Automatically Modeling Conversations as Processes of Interrelated Speech Intentions

Elena Epure

▶ **To cite this version:**

Elena Epure. Automatically Modeling Conversations as Processes of Interrelated Speech Intentions. Artificial Intelligence [cs.AI]. Paris 1 - Panthéon-Sorbonne, 2018. English. NNT : . tel-02021609v1

HAL Id: tel-02021609

https://paris1.hal.science/tel-02021609v1

Submitted on 19 Feb 2019 (v1), last revised 21 Oct 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse
pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PARIS 1 PANTHÉON-SORBONNE
Spécialité: Informatique

Présentée par:

ELENA VIORICA EPURE

Titre:

Automatically Modeling Conversations as Processes of Interrelated Speech Intentions

Thèse soutenue publiquement le 14 décembre 2018 devant le jury composé de :

Camille SALINESI	Université Paris 1 Panthéon-Sorbonne	Directeur
Rébecca DENECKÈRE	Université Paris 1 Panthéon-Sorbonne	Co-directeur
David NACCACHE	École Normale Supérieure Ulm	Rapporteur
Alain WEGMANN	École Polytechnique Fédérale de Lausanne	Rapporteur
Christophe CERISARA	Centre National de la Recherche Scientifique	Examinateur
Frank HOPFGARTNER	University of Sheffield	Examinateur

TRADUCTION FRANÇAISE DU RÉSUMÉ

La prolifération des données numériques a permis aux communautés de scientifiques et de praticiens de créer de nouvelles technologies basées sur les données pour mieux connaître les utilisateurs finaux et en particulier leur comportement. L'objectif est alors de fournir de meilleurs services et un meilleur support aux personnes dans leur expérience numérique. La majorité de ces technologies créées pour analyser le comportement humain utilisent très souvent des données de logs générées passivement au cours de l'interaction homme-machine. Une particularité de ces traces comportementales est qu'elles sont enregistrées et stockées selon une structure clairement définie. En revanche, les traces générées de manière proactive sont très peu structurées et représentent la grande majorité des données numériques existantes. De plus, les données non structurées se trouvent principalement sous forme de texte. À ce jour, malgré la prédominance des données textuelles et la pertinence des connaissances comportementales dans de nombreux domaines, les textes numériques sont encore insuffisamment étudiés en tant que traces du comportement humain pour révéler automatiquement des connaissances détaillées sur le comportement.

L'objectif de recherche de cette thèse est de proposer une méthode indépendante du corpus pour exploiter automatiquement les communications asynchrones en tant que traces de comportement générées de manière proactive afin de découvrir des modèles de processus de conversations, axés sur des intentions de discours et des relations, toutes deux exhaustives et détaillées.

Plusieurs contributions originales sont faites. Il y est menée la seule revue systématique existante à ce jour sur la modélisation automatique des conversations asynchrones avec des actes de langage. Une taxonomie des intentions de discours est dérivée de la linguistique pour modéliser la communication asynchrone. Comparée à toutes les taxonomies des travaux connexes, celle proposée est indépendante du corpus, à la fois plus détaillée et exhaustive dans le contexte donné, et son application par des non-experts est prouvée au travers d'expériences approfondies. Une méthode automatique, indépendante du corpus, pour annoter les énoncées de communication asynchrone avec la taxonomie des intentions de discours proposée, est conçue sur la base d'un apprentissage automatique supervisé. Pour cela, deux corpus "ground-truth" validés sont créés et trois groupes de caractéristiques (discours, contenu et conversation) sont conçus pour être utilisés par les classificateurs. En particulier, certaines des caractéristiques du discours sont nouvelles et définies en considérant des moyens linguistiques pour exprimer des intentions de discours, sans s'appuyer sur le contenu explicite du corpus, le domaine ou les spécificités des types de communication asynchrones. Une méthode automatique basée sur la fouille de processus est conçue pour générer des modèles de processus d'intentions de discours interdépendantes à partir de tours de parole, annotés avec plusieurs labels par phrase. Comme la fouille de processus repose sur des logs d'événements structurés et bien définis, un algorithme est proposé pour produire de tels logs d'événements à partir de conversations. Par ailleurs, d'autres solutions pour transformer les conversations annotées avec plusieurs labels par phrase en logs d'événements, ainsi que

l'impact des différentes décisions sur les modèles comportementaux en sortie sont analysées afin d'alimenter de futures recherches.

Des expériences et des validations qualitatives à la fois en médecine et en analyse conversationnelle montrent que la solution proposée donne des résultats fiables et pertinents. Cependant, des limitations sont également identifiées, elles devront être abordées dans de futurs travaux.

Abstract

The proliferation of digital data has enabled scientific and practitioner communities to create new data-driven technologies to learn about user behaviors in order to deliver better services and support to people in their digital experience. The majority of these technologies extensively derive value from data logs passively generated during the human-computer interaction. A particularity of these behavioral traces is that they are structured. However, the pro-actively generated text across Internet is highly unstructured and represents the overwhelming majority of behavioral traces. To date, despite its prevalence and the relevance of behavioral knowledge to many domains, such as recommender systems, cyber-security and social network analysis, the digital text is still insufficiently tackled as traces of human behavior to automatically reveal extensive insights into behavior.

The main objective of this thesis is to propose a corpus-independent method to automatically exploit the asynchronous communication as pro-actively generated behavior traces in order to discover process models of conversations, centered on comprehensive speech intentions and relations. The solution is built in three iterations, following a design science approach.

Multiple original contributions are made. The only systematic study to date on the automatic modeling of asynchronous communication with speech intentions is conducted. A speech intention taxonomy is derived from linguistics to model the asynchronous communication and, compared to all taxonomies from the related works, it is corpus-independent, comprehensive—as in both finer-grained and exhaustive in the given context, and its application by non-experts is proven feasible through extensive experiments. A corpus-independent, automatic method to annotate utterances of asynchronous communication with the proposed speech intention taxonomy is designed based on supervised machine learning. For this, validated ground-truth corpora are created and groups of features—discourse, content and conversation-related, are engineered to be used by the classifiers. In particular, some of the discourse features are novel and defined by considering linguistic means to express speech intentions, without relying on the corpus explicit content, domain or on specificities of the asynchronous communication types. Then, an automatic method based on process mining is designed to generate process models of interrelated speech intentions from conversation turns, annotated with multiple speech intentions per sentence. As process mining relies on well-defined structured event logs, an algorithm to produce such logs from conversations is proposed. Additionally, an extensive design rationale on how conversations annotated with multiple labels per sentence could be transformed in event logs and what is the impact of different decisions on the output behavioral models is released to support future research. Experiments and qualitative validations in medicine and conversation analysis show that the proposed solution reveals reliable and relevant results, but also limitations are identified, to be addressed in future works.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Camille Salinesi and Rébecca Deneckère, for their kindness, friendliness and constant support and scientific guidance throughout my PhD. You challenged me to become a better researcher and teacher. You encouraged me and believed in me from the beginning, even when I decided to venture into several different research directions at the same time. I learned a lot from you, beyond the academic knowledge you have skillfully shared with me over these years.

I want to thank Charlotte Hug, whom I was lucky to have as a third supervisor in my first year of the thesis. We started this adventure together and the advice you gave me, and everything I learned from you during this first year and during our collaboration when I was still a Master's exchange student, followed me all these years. You taught me courage, determination and dedication.

On this occasion, I would also like to show my greatest appreciation to my Utrecht University professor, Sjaak Brinkkemper, who was my mentor during my master's degree and recommended me in the first place for a research exchange abroad. Spending eight months in the research lab where I would later start my PhD was a revealing experience and a milestone in my decision to pursue a research career.

I would like to offer my special thanks to Prof. David Naccache from École Normale Supérieure and Prof. Alain Wegmain from École Polytechnique Fédérale de Lausanne for kindly accepting to be the referees of my thesis. Your valuable recommendations significantly helped me to highlight the contributions of my thesis and to improve the manuscript in general.

I would like to thank Dario Compagno from Institut de Recherche Médias, Cultures, Communication et Numérique, Sorbonne Nouvelle University for a fascinating multi-disciplinary collaboration of more than 2 years. You taught me a lot about linguistics and semiotics, about beautiful writing and about how to maintain a positive attitude despite challenges.

Many thanks to Jon Espen Ingvaldsen from Norwegian University of Science and Technology for introducing me to the field of news recommendations, for facilitating the validation of our proposal with Adresseavisen, the largest newspaper of the third largest media group in Norway, and for later helping me collaborate with Distributed Artificial Laboratory, Technical University of Berlin for a large-scale evaluation. This latter step would not have been possible without the implication of Benny Kille from TU Berlin, who had very extensive experience in recommender systems and access to data from several German newspapers. Thank you so much, Benny for showing me that successful collaborations are possible without having met before and without being in the same place. It was a real pleasure to have finally met you in person at RecSys2017, where we presented the fruitful results of our work.

I would like to thank Slavko Žitnik and Prof. Marko Bajec for their initiative to collaborate on my thesis topic and for warmly welcoming me to the University of Ljubljana for two research exchanges, during the summers of 2016 and 2017. Marko, I still remember your challenging

questions at the beginning of my PhD and this certainly helped me to better address my topic. Slavko, I highly appreciate your help with implementing a part of my experiment, your wise advice, your kindness and reliability, and your being the best guide to Slovenia.

I am thankful to Patricia Martin-Rodilla from Institute of Heritage Sciences, Spanish National Research Council for introducing me to multi-disciplinary research during her mission in our lab. I knew very little about archaeology before meeting you and I was very happy to work together on something that could help archeologists to structure their methodological knowledge. Thank you also for all the advice you gave me at the beginning of my PhD.

I would like to thank Dr. Nhân Pham-Thi from Pasteur Institute and Prof. Neil Maiden from City University of London. The exchanges I had with you about the importance of medical stories, especially those of patients, have greatly helped me to understand the field and to later establish the relevance of my work to health care.

I am very grateful to all the professors in Centre de Recherche en Informatique, Paris 1 who have given me the opportunity to enhance my academic profile. In particular, I warmly thank Carine Souveyet, Bénédicte Le Grand and my supervisors for entrusting me with giving lectures, supervising research projects and with reviewing articles for the RCIS2016 and RE2017 international conferences. Also, thank you for sponsoring my participation in two summer schools: Deep Learn 2017 (Bilbao) and Web Intelligence 2016 (Saint-Étienne).

I was lucky to be in a lab where there were many female computer scientists and grateful to see the progress that many organizations are making with increasing diversity. I would like to thank Google for inviting me on several occasions to Women Techmakers events in Paris and for financially supporting me to attend the ACM-W womENcourage conference in Linz in 2016. The exchanges I had there with fellow female researchers were very inspiring.

My life as a PhD student would have not been the same without all my lab mates: Asmaa, Elena, Angela, Luisa, Ali, Fabrice, Danny, David, Afef, Raouia, Amina, Danillo, Lamia, Floriane, Sabrine, Houssem... We shared so many moments together, which I will definitely never forget. I also want to thank Stéphane and Astrid: you have always been so kind helping me with the organization of my classes and research missions abroad. Stéphane, I always enjoyed talking about music with you.

I was also very fortunate to be surrounded by many friends here, during this important and sometimes challenging phase of my life. I wholeheartedly thank Irina, Kim, Romana, Séverine, Cécile, Arielle and Iryna for having encouraged me all this time and for having pushed me to have a life beyond the PhD.

Finally, I dedicate this work to my strongest pillars of support and love: my partner Marc-Antoine, my sister Cristina and my parents, Rodica and Cristi. You have always helped me find my north.

TABLE OF CONTENTS

Page

List of Tables **xiii**

List of Figures **xvii**

Traduction française de l'introduction **xix**

Contexte de Recherche xix
Problématique xx
Cadre de Recherche, Objectif et Questions xxii
Périmètre des données en entrée xxii
Périmètre des models en sortie xxiv
Objectif de Recherche et Questions xxvi
Contributions et Publications xxvi
Organisation de la Thèse xxxi

1 Introduction **1**

1.1 Research Context 1
1.2 Problem Statement 2
1.3 Research Scope, Objective and Questions 3
1.3.1 Input Scope 4
1.3.2 Output Scope 5
1.3.3 Research Objective and Questions 7
1.4 Research Approach 9
1.5 Contributions and Publications 13
1.6 Overview 16

2 Literature Review **19**

2.1 Research Method 20
2.1.1 Research Questions, Foci and Goals 20
2.1.2 Data Collection 21
2.1.3 Inclusion/Exclusion Criteria and Other Aspects of Data Evaluation 23

2.1.4 Analysis and Interpretation . . . 24
2.2 A Conceptual Framework for Analysis and Interpretation . . . 25
2.2.1 Input-centered Facets . . . 25
2.2.2 Output-centered Facets . . . 27
2.2.3 Practice-centered Facets . . . 27
2.2.4 Method-centered Facets . . . 28
2.2.5 Outcome-centered Facets . . . 33
2.3 Results . . . 34
2.3.1 Asynchronous Communication Corpora . . . 35
2.3.2 Applications and Beneficiaries . . . 36
2.3.3 Speech Intention Taxonomies for Asynchronous Communication . . . 38
2.3.4 Manual Annotation of Asynchronous Communication . . . 44
2.3.5 Automatic Annotation of Asynchronous Communication . . . 47
2.3.6 Considering Relations among Speech Intentions . . . 51
2.4 Conclusions . . . 53
2.4.1 Research Questions Revisited . . . 53
2.4.2 Study Limitations . . . 56
2.4.3 Summary of Contributions . . . 57
2.4.4 Directions for Further Research . . . 58

3 Modeling Public Tweets with Speech Intentions 61
3.1 A Speech Intention Taxonomy for Public Tweets . . . 63
3.1.1 Experiments . . . 65
3.1.2 Results . . . 68
3.2 Automatically Annotating Tweets with Speech Intentions . . . 70
3.2.1 Ground-truth Corpus . . . 70
3.2.2 Discourse Features of Speech Intentions . . . 71
3.2.3 Experiments . . . 75
3.2.4 Results . . . 81
3.3 Conclusion . . . 83
3.3.1 Relevance of Designed Approach to Medicine . . . 83
3.3.2 Study Limitations . . . 85
3.3.3 Directions for Further Research . . . 86

4 A Structural Taxonomy of Speech Intentions 87
4.1 Taxonomy Design . . . 88
4.1.1 Design Rationale . . . 89
4.1.2 Deriving a Structural Speech Intentions Taxonomy . . . 90
4.2 Experiments . . . 94

4.3 Results . . . 97
4.4 Conclusion . . . 102
4.4.1 Study Limitations . . . 103
4.4.2 Directions for Further Research . . . 103

5 Modeling Conversations with Speech Intentions 105
5.1 Medicine as Application Domain . . . 107
5.2 Ground-truth Corpora . . . 109
5.3 Feature Engineering . . . 113
5.4 Experiments . . . 115
5.5 Results . . . 117
5.6 Conclusion . . . 124
5.6.1 Study Limitations . . . 125
5.6.2 Directions for Further Research . . . 126

6 Modeling Conversations as Processes over Speech Intentions 129
6.1 Introduction to Process Mining . . . 131
6.1.1 Heuristic Miner Algorithm . . . 133
6.1.2 Fuzzy Miner Algorithm . . . 135
6.1.3 Discussion . . . 137
6.2 Process Mining to Model Asynchronous Conversations . . . 138
6.2.1 Generating Event Logs from Annotated Asynchronous Conversations . . . 138
6.2.2 Alternative Heuristics to Map Utterances on Events . . . 142
6.3 Evaluation Methods . . . 144
6.4 Results . . . 146
6.4.1 Scenario-Based Evaluation Results in Medicine . . . 146
6.4.2 Observational Evaluation Results in Conversation Analysis . . . 149
6.5 Conclusions . . . 153
6.5.1 Study Limitations . . . 154
6.5.2 Directions for Further Research . . . 155

7 Conclusion and Perspectives 157

A Introduction to Design Science 163

B Running Example to Illustrate the Contribution 165

C Research Process for the Systematic Literature Review 173

D Literature Search Results 175

E Speech Act Verbs **179**

F Experimental Resources for Manual Annotation **183**

G Recommending Personalized News in Short User Sessions **189**
G.1 Introduction . . . 190
G.2 Defining Reading Behavior in News . . . 191
G.3 Experiments . . . 193
G.3.1 Data Description . . . 193
G.3.2 Data Transformation . . . 194
G.3.3 News Reading Dynamics . . . 194
G.3.4 Recommendation Policy Comparison . . . 196
G.4 Results and Discussion . . . 197
G.4.1 News Reading Dynamics . . . 197
G.4.2 Recommendation Policy Comparison . . . 198
G.4.3 Reflections on News Recommendation . . . 201
G.5 Literature Review . . . 201
G.5.1 News Reading Interests . . . 202
G.5.2 News Recency, Popularity, and Variety . . . 202
G.5.3 Session-based Recommendation . . . 203
G.5.4 Process Models for Recommendation . . . 204
G.5.5 News Reading Behavior Analysis . . . 204
G.6 Conclusion and Future Works . . . 205

H Automatic Process Discovery from Textual Methodologies **207**
H.1 Introduction . . . 208
H.2 Methodology . . . 209
H.2.1 Problem Investigation . . . 209
H.2.2 Design of the Technique . . . 210
H.2.3 Validation of the Technique . . . 210
H.3 Illustration . . . 210
H.4 Solution . . . 215
H.4.1 *TextCleaner* . . . 216
H.4.2 *TextProcessMiner* . . . 217
H.4.3 Discussion . . . 220
H.5 Validation and Preliminary Evaluation . . . 222
H.5.1 Validation and Evaluation Setup . . . 222
H.5.2 Results . . . 223
H.6 Related Works . . . 225

H.7 Conclusion and Future Works 227

Bibliography **229**

List of Tables

Table **Page**

1.1 Summary of the solution-driven problem investigation in the current work. 10

2.1 The research method followed for conducting the systematic literature review [42, 164]. 20
2.2 Inclusion and exclusion criteria for assessing relevance and quality. 24
2.3 Classification scheme with multiple facets and their associated categories. 26
2.4 Interpretation of Kappa scores—at least moderate results are preferred. 34
2.5 Confusion matrix example for predicting a certain class. 34
2.6 Overview of the number of works per type of asynchronous communication and year. 35
2.7 Forum-related taxonomies—an alignment of classes proposed or used across different works (cited in the first row). 42
2.8 Statistical measures used in the related works for validating ground-truth corpora: the first column shows the type, the second column shows the percentage of works that uses it from the total number of works reporting validation. The sum is not 100% because several works use multiple statistical measures. 46
2.9 Percentage of related works using each feature group—only the supervised and semi-supervised machine learning solutions are considered. 49
2.10 Supervised and semi-supervised algorithms used in the related works. 50

3.1 Identified speech intentions for the Twitter public communication. 64
3.2 κ scores for speech act types and intentions. At least substantial results are underlined. 69
3.3 Confusion matrix showing the agreements (the values on the diagonal) and the disagreements between two annotators in using the proposed speech intentions. . . . 69
3.4 Proposed discourse features for discovering speech intentions. 72
3.5 Overall results of the classification experiment using various feature sets and only the single-labeled tweets. The measures are macro-weighted. The best results are in bold. 81
3.6 The results of the classification experiment using various feature sets for each class and only the single-labeled tweets. The support of each class is presented in the first column, under the label. The measures are macro-weighted. Best results for each speech intention and for each type of classifier are in bold. Best results overall for each speech intention are underlined. 82

4.1 Fleiss' Kappa scores (κ) obtained for each speech act type and overall, computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The second to last row contains the median (Med) of all κ scores per speech act type across datasets. All p-values are 0 except for Other—groups 3, 4, 5 and 6. The last row presents how often each speech act type was used in the experiment, computed by considering all labels provided by each annotator. 98

4.2 Fleiss' Kappa scores (κ) obtained for *assertive* speech intentions computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The last row contains the median (Med) of all κ scores per speech intention. All p-values < 0.05, except for *disagree*—groups 1, 5, 7, 8 and 10, and *agree*—group 8. 99

4.3 Fleiss' Kappa scores (κ) obtained for *expressive* speech intentions computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The last row contains the median (Med) of all κ scores per speech intention. All p-values < 0.05, except for *rejoice*—group 6, *wish*—group 6, *apologize*—group 1, *thank*—group 1, and *greet*—group 5. 100

4.4 Fleiss' Kappa scores (κ) obtained for *directive* and *commissive* speech intentions computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The last row contains the median of all κ scores per speech intention. All p-values < 0.05, except for *engage*—groups 5 and 8, *accept*—groups 1, 2, 5, 6, 7 and 9, and *refuse*—groups 1, 2, 3, 4, 8 and 10. 100

4.5 Confusion matrix for the proposed speech intentions. For the sake of brevity, only the first three letters of each class are used to denote the class. The pairs without the empty choice, for all groups of annotators, were taken in consideration for computing the matrix. The last row presents how often each class was used in the experiment, computed by considering all labels provided by each annotator. 101

5.1 Statistics of the *Reddit* corpus and of the two external corpora, *Bhatia* and *SWBD*. *Bhatia* is released as threads of conversations and each turn is labeled. *SWBD* is composed of synchronous conversations and does not contain threads; each utterance or part of it is labeled. *Reddit* is labeled at the utterance level (grammatical sentences). 110

5.2 Description of the speech intentions. The frequency of each class is presented under the name—if missing, the class is not used in the automatic experiments. *Assert* includes also *sustain*, *agree* stands for *agree* and *disagree*, *engage* includes also *accept* and *refuse*. 111

5.3 Mapping of the taxonomy classes on *Bhatia* and *SWBD* classes. Not all classes could be aligned. The number of instances for each class from the external corpora is in parentheses. 113

5.4 Proposed features (3 main groups): 1–*Content*, 2–*Discourse*, 3–*Conversation*; when "highly correlated" features are mentioned, these are selected with the χ^2 statistical test. 114

5.5 Precision (*P*), Recall (*R*) and F1-score (*F1*) —macro values, in percentage—and Cohen's Kappa score (κ) obtained by Logistic Regression, Linear SVM and Random Forest in classifying intentions and speech act types on the *Reddit* corpus. For each cell, the first line contains the score for intention classification and the second line for speech act type classification. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best F1 and κ scores of each classifier in predicting the intentions and speech act types are in bold. Best overall F1 and κ scores are underlined. 118

5.6 Per-intention binary F1-scores (in percentage) and κ scores obtained by the Logistic Regression classifier in predicting speech intentions on the *Reddit* corpus, in an one-versus-all setup. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best scores for each speech intention are in bold. 120

5.7 Hamming and macro F1-scores obtained by Logistic Regression in predicting intentions on the *Reddit* corpus, in a multi-label setup. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best Hamming and F1-scores are in bold. 121

5.8 Per-intention binary F1-score (in percentage) and κ score obtained by the Logistic Regression classifier in predicting the selected speech intentions on the *Reddit* corpus augmented with instances from *Bhatia* and *SWBD*. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best scores for each intention are in bold. 123

5.9 Macro F1-scores (in percentage) obtained by the Logistic Regression classifier in predicting intentions on *Bhatia* and *SWBD* corpora. From these corpora, only the classes that could be aligned with the current taxonomy were selected. For each cell, the first line contains the score for speech intention classification and the second line the score for speech act type classification. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best scores are in bold. 123

6.1 Evaluation methods used in the current work, selected from [93]. 144

B.1 Sentences annotated with speech intentions. 165

G.1 News categories and their associated codes 193

H.1 Example of a methodology fragment before and after cleaning. 211
H.2 Treebank for the first sentence of the fragment. 212
H.3 Treebank for the second sentence of the fragment. 213
H.4 Treebank for the third sentence of the fragment. 213
H.5 Activities extracted from the fragment. 214
H.6 Symbols in the process model representation. 214
H.7 Process instance discovered from the fragment. 215
H.8 Rules to mine activities relationships. 220

List of Figures

Figure **Page**

3.1 Medical terms and jargon, lexicon proposed by Ridpath et al. [170]. 66

4.1 The proposed structure for the five main types of speech acts. 90
4.2 The speech intention taxonomy with speech act types in the first-level nodes, speech intentions in the leaves and oppositional traits in the intermediary nodes [41]. 92

6.1 The overall approach to model conversations as processes over speech intentions. The current chapter is focused on applying process mining, emphasized in the bottom part of the figure. Existing process mining techniques are reused. The challenge then is how to transform conversations annotated at utterance-level in well-formed logs of representative verbal behavior in order to discover relevant processes for conversation analysis. 139
6.2 Within-turn processes over speech intentions, mined from a *Reddit* corpus with Disco. 147
6.3 Process model mined from all corpus turns with Searle's speech act types as activities. 150
6.4 Process model mined from all corpus turns with speech intentions as activities. 152

A.1 Design Science Research Cycles. 163

B.1 Example of a Reddit conversation. 167
B.2 The conversation tree obtained from the annotated conversation. 168
B.3 Another thread of the conversation introduced as a running example. 169
B.4 The resulting event log as a csv file. 170
B.5 The process model obtained from the event log. 170
B.6 Another view of the process model obtained from the event log. 171
B.7 Global overview of a conversation corpus based on the process mining models obtained at each level in the conversation trees. 171

E.1 Semantic tree model for commissives [212]. 179
E.2 Semantic tree model for expressives [212]. 179
E.3 Semantic tree model for assertives [212]. 180
E.4 Semantic tree model for directives [212]. 181

G.1 Histogram of events per session (log scale) 194
G.2 Probabilities per month for transitions from General News—code 7, to all the other categories in 2015. 198
G.3 The cumulative distribution function (CDF) of the hourly response rates for the compared recommenders. The inclusion of long and medium-term interests in the recommendation policy improves the results over using the short-term interests only. 199
G.4 The pair-wise comparison of the three recommendation policies: T_1 v. B, T_c v. B and T_1 v. T_c. The strategies with long and short-term interests yield consistently the best results. 199
G.5 Analysis of the news variety ensured by T_1 and T_c. The two histograms show the distribution across the spectrum. T_1 has a peak of very high variety, which T_c lacks. 200

H.1 Proposed solution for automatically discovering process models from text: an overview. 216
H.2 Fragment of the discovered log. 224
H.3 Fragment of the discovered process instance model. 226

Contexte de Recherche

L'utilisation d'Internet dans la vie personnelle et professionnelle quotidienne est omniprésente de nos jours [143]. En conséquence, chaque individu génère des données très diverses dans des volumes sans précédent. Les utilisateurs créent et partagent du contenu de manière *proactive* lors de publications sur les réseaux sociaux [11], d'écriture de commentaires sur les plateformes en ligne [191], de rédaction de rapports de travail [177] ou de l'envoi de code dans des dépôts en ligne [6]. De plus, il existe aussi une génération *passive* de données. En effet, les interactions homme-machine sont fréquemment tracées par les logiciels [1, 3, 6]. L'objectif est alors de mieux connaître les utilisateurs finaux et in fine, fournir de meilleurs services et un meilleur support aux personnes dans leur expérience numérique.

La prolifération des données numériques a permis aux communautés de scientifiques et de praticiens de créer de nouvelles technologies basées sur les données pour mieux connaître les utilisateurs finaux et en particulier leur *comportement* [1, 3, 6, 11, 191].

Définition 1.1 Comportement: Le comportement est défini comme "la manière dont les humains agissent et interagissent"[1].

De nombreux domaines et applications tournent autour de la modélisation automatique du comportement. Sur les marchés financiers, le comportement des investisseurs est analysé en tant que modèle de trading afin d'identifier la fraude ou le trading spéculatif [47]. Dans la lutte contre le terrorisme, la découverte de comportements à partir de données en ligne facilite les enquêtes et renforce la prévention [38]. Dans les systèmes de recommandation pour le commerce en ligne, les films, la musique, les nouvelles, etc., le comportement des utilisateurs est identifié par rapport aux éléments consommés et aux similarités avec les autres utilisateurs [3]. Sur les réseaux sociaux, le comportement collectif est analysé en profilant les structures sociales afin de mieux comprendre les relations humaines [104]. La communauté scientifique dédiée à la détection d'anomalies propose des méthodes pour découvrir des valeurs comportementales aberrantes à partir de données, appliquées aux domaines de la finance, de la santé et de la cybersécurité [32]. Aussi, le comportement est modélisé en tant que processus et est découvert automatiquement à partir de données, pour la gestion opérationnelle et l'évolution des systèmes

[1] https://www.nature.com/subjects/human-behaviour

d'information [1].

La majorité de ces technologies créées pour analyser le comportement humain utilisent très souvent des données de logs générées passivement au cours de l'interaction homme-machine [1, 3]. Une particularité de ces *traces comportementales* est qu'elles sont enregistrées et stockées selon une structure clairement définie. Ce type de traces est appelé *structuré* ou *semi-structuré*. Celles semi-struturées sont souvent constituées de logs de navigation Web. La structure existante des données joue un rôle dans l'extraction des connaissances comportementales et dans la création des technologies afférentes. En revanche, les traces générées de manière proactive sont *très peu structurées* et représentent la grande majorité des données numériques existantes[2]. De plus, les données non structurées se trouvent principalement sous forme de *texte*.

Problématique

À ce jour, malgré la prédominance des données textuelles et la pertinence des connaissances comportementales dans de nombreux domaines, les textes numériques sont encore insuffisamment étudiés en tant que *traces du comportement humain* pour révéler automatiquement des *connaissances détaillées sur le comportement* [222]. Les technologies d'analyse de texte sont orientées sur des tâches essentielles du traitement automatique de la langue, telles que la désambiguïsation et l'analyse du sens des mots, ainsi que sur des tâches d'extraction d'informations spécifiques, telles que la reconnaissance d'entités nommées [60, 134]. Dans une certaine mesure, les aspects comportementaux issus des textes sont étudiés au moyen de classification et clustering [60, 134]. L'analyse des sentiments, également connue sous le nom d'exploration d'opinion, est une application très fréquente de ces technologies d'analyse de texte. Elle propose des techniques pour analyser le comportement linguistique, permettant d'identifier et de quantifier les états affectifs [152].

Cependant, l'association entre le comportement et le texte est un sujet beaucoup plus complexe [222], qui va au-delà de la découverte d'opinions positives ou négatives. Parmi les modes de représentation et les moyens de communication, le texte est le plus fréquemment utilisé, notamment sous forme numérique. La communication, en tant que production du langage, est l'un des principaux moyens d'agir et d'interagir. Ce sujet a beaucoup été étudié en sciences humaines et sociales [9, 12, 95, 184, 197]. En particulier, le 20ème siècle a vu une croissance de l'intérêt porté à l'étude de la communication en tant que forme du comportement humain et interactions, dans les domaines de la philosophie et de la linguistique.

Dans ses cours "How to do things with words", John Austin [12] a introduit ce nouveau point de vue sur l'utilisation du langage en linguistique. Plus précisément, il a avancé l'idée que le langage n'est pas seulement utilisé pour communiquer des faits vrais ou faux, détachés de l'action et de la compréhension sociales; mais aussi que les gens réalisent des actions par le discours. La

[2]Il est communément partagé que 80% des données numériques sont non structurées [18].

reconnaissance mutuelle des intentions de ces actions constitue la base de l'interaction humaine à travers le langage. Les *actes de langage* ont été inventés pour conceptualiser et différencier les *intentions du discours* réalisées par la production et l'interprétation des *énoncés* écrits ou parlés [197].

Définition 1.2 Énoncé: Un énoncé a plusieurs définitions[3]. Dans le discours, un énoncé est l'unité de la parole délimitée par le silence ou les pauses du locuteur. Dans les dialogues, un énoncé peut être un tour de parole. En linguistique, un énoncé est l'unité de parole ou de texte qui est étudiée.

Définition 1.3 Intention du discours: Une intention du discours est utilisée pour désigner l'intention linguistique d'un locuteur de produire une expression telle que les promesses, les demandes, les salutations, les affirmations, etc. Elle est également connue comme la fonction illocutoire d'une expression [184].

Définition 1.4 Acte de langage: Un acte de langage est un acte accompli par un locuteur lorsqu'il prononce un énoncé. Cela inclut la prononciation ou l'écriture de mots et l'intention du discours associée.
Par exemple, l'énoncé d'un enseignant "Rendez les devoirs d'ici vendredi" est un acte de langage, et l'intention du discours associée peut être incluse dans la classe des requêtes. Ainsi, le langage parlé ou écrit peut être analysé en tant que comportement humain composé d'actes de langage produits à travers l'énonciation de phrases.

Par ailleurs, l'extraction des connaissances du comportement humain à partir d'un texte est rendue très difficile par la nature même des textes [60]. Outre le manque de structure, le texte peut apparaître dans des styles très variés, dépendant :

- de la fonction du discours : un article de blog est écrit différemment qu'un message sur les réseaux sociaux;
- de la technologie utilisée : Twitter, par exemple, a conduit à une forme de discours unique en imposant un maximum de 140 caractères par message;
- des caractéristiques du discours de chaque individu.

De plus, les technologies de base pour le traitement automatique du texte, par exemple pour effectuer automatiquement des analyses morphologiques ou sémantiques [134], sont encore imparfaites. Par conséquent, la conception de solutions plus intelligentes d'analyse de texte est souvent limitée par l'efficacité du logiciel sous-jacent.

Bien que le défi puisse être technologique en ce qui concerne le traitement des textes en entrée, les limitations pour la découverte des connaissances comportementales en sortie proviennent d'autres raisons:

[3] `https://glossary.sil.org/term/utterance`

- Dans le cas de la modélisation du comportement basée sur la théorie linguistique et centrée sur des actes de langage, les solutions sont souvent en pratique, conçues pour des problématiques spécifiques. Par exemple, les actes de langage sont utilisés pour améliorer les outils de gestion d'emails / tâches [28, 40], ou afin de faciliter la recherche d'informations et la gestion des discussions dans les forums [10, 118, 163].

- Il existe également d'autres approches pour la découverte du comportement à partir de textes [163, 213, 234]. Ces approches se basent sur des représentations générales d'actes de langage alignées sur les modèles théoriques existants de la linguistique [184]. Dans ce cas, le niveau de détail des connaissances comportementales découvertes est souvent faible, comme cela sera détaillé au Chapitre 2.

- Inversement, des définitions très détaillées du comportement linguistique au travers des actes de langage ont également été utilisées dans des approches automatiques [110, 196]. Cependant, leur niveau de représentation très détaillé présente un défi pour leur interprétation et leur réutilisation par des non-experts en analyse de discours.

Cadre de Recherche, Objectif et Questions

Avant de présenter l'objectif et les questions de recherche, une discussion plus approfondie sur la communication numérique et sa portée dans la présente thèse est présentée. Ensuite, d'autres aspects théoriques pertinents du comportement par le langage et la conceptualisation du comportement au travers des intentions de discours, sont introduits.

Périmètre des données en entrée

La vie quotidienne est de plus en plus liée au digital. En conséquence, les pratiques de communication ont évolué. La communication se fait par des canaux et formats modernes dictés par le nouvel ordre numérique [218]. Le texte peut être considéré, d'une part, comme *un produit de la communication*, transmettant le contenu de la communication. Il peut aussi être considéré en tant que *structure verbale* ou formulation représentant physiquement la communication [95]. T. Hillesund affirme que ces deux visions conceptuelles du texte ne sont pas contradictoires [95]. Il propose une définition unifiée: "le texte est une représentation visuelle de l'information verbale".

En général, la communication peut être caractérisée à travers plusieurs dimensions: modalité, support, registre, nombre de participants à la conversation et immédiateté [100]. La modalité de communication distingue les échanges oraux et écrits. Le support de communication fait référence aux moyens par lesquels la communication a lieu. Cela inclut en particulier les supports analogiques, tels que les conversations téléphoniques, ainsi que les supports numériques, correspondant aux communications par ordinateur. Le registre fait référence à la formalité de la langue utilisée; il peut être informel ou formel. L'immédiateté fait référence à l'aspect *synchrone*

de la communication, par exemple dans les messageries instantanées, ou à l'aspect *asynchrone*, comme dans les emails et les publications sur les réseaux sociaux. Par ailleurs, un autre moyen pour caractériser les types de communication est l'objectif de communication porté communément par les interlocuteurs lors d'une conversation [100]. La recherche d'informations, la fourniture d'instructions ou le partage d'une histoire sont des exemples d'objectifs de communication.

Dans le cadre de cette thèse, seules *les communications textuelles sur support numérique* sont couvertes. Une fois qu'un message a été énoncé, les données textuelles stockées deviennent *une trace* du message. Aucune règle sur le nombre de participants, le registre de la communication ou l'objectif de communication n'a été imposée. Cependant, on se concentrera sur *la communication asynchrone* en entrée.

Définition 1.5 Communication asynchrone: La communication asynchrone comprend les conversations qui peuvent avoir lieu sur de plus longues périodes et auxquelles les individus peuvent participer à tout moment en répondant aux messages existants. Etant donné que ces conversations ne sont pas en temps réel et que les interventions sont peu structurées, des fils de discussion sont souvent utilisés pour structurer ces conversations et aider les utilisateurs à les suivre.

Définition 1.6 Fil de conversation: un fil de conversation se compose d'une liste de messages connexes, qui apparaissent dans la conversation de la manière suivante: le message d'origine, une réponse au message d'origine, une réponse à la réponse précédente, etc. Si une conversation est imaginée sous la forme d'un arbre dont la racine est le message d'origine et les noeuds sont les autres messages, où chaque relation parent-enfant implique que le message dans le noeud enfant est une réponse au message dans le noeud parent, alors un fil de conversation contient tous les messages correspondant à un chemin entre le noeud racine et un noeud feuille.

Des recherches sont toujours en cours pour comprendre la nature de la communication asynchrone. Par exemple, comme rapporté par Goldstein et al. [79], des analyses empiriques révèlent que ce genre est hybride et présente des caractéristiques à la fois de la parole et de l'écriture. En effet, la communication asynchrone numérique ressemble beaucoup aux correspondances par lettres, mais, à en juger par le niveau d'interaction, elle est beaucoup plus proche des discours synchrones, qui sont conversationnels. Les dialogues synchrones convergent vers des modes de communication relativement standard, mais la diversité des échanges asynchrones est beaucoup plus grande.

Comparée aux conversations synchrones, la communication asynchrone est composée de messages plus longs, contenant des éléments conversationnel et non conversationnel, tels que des documents joints dans des courriers électroniques ou des extraits de code dans des messages de forum [194]. En outre, la communication asynchrone a tendance à être plus souvent un texte expositoire [163] et à avoir des "modèles ou conventions stéréotypés" spécifiques à chaque sous-catégorie, tels que les emails ou les tweets [138, 234]. Les communications synchrones ont une structure linéaire et ont généralement des objectifs délimités clairs (par exemple, des messages

échangés pour négocier une décision). Au contraire, dans une communication asynchrone, les objectifs varient et peuvent être combinés dans un même message [100] (par exemple, le même email peut être utilisé pour planifier des tâches et fournir des informations). En outre, la structure des conversations asynchrones est généralement un graphe, où les noeuds sont des tours de parole et les arêtes sont des liens entre ces tours consécutifs dans les fils de conversation. Les participants interviennent dans la discussion dans un ordre peu structuré et peuvent répondre à des messages plus anciens que les messages plus récents [199].

Par conséquent, comparée à la communication synchrone, la communication asynchrone est très variée et implique des défis et des particularités qui ont un impact sur la conception d'une solution. De plus, si une solution appropriée pour la communication asynchrone est trouvée, elle sera alors transférable à la communication synchrone. En effet, de nombreuses caractéristiques liées à la communication synchrone sont intégrées dans celles de la communication asynchrone ou en sont des simplifications. Par conséquent, la démarche de cette thèse commence par étudier des conversations asynchrones à base de texte numérique, puis montrera en détail comment ces conversations représentent des traces de comportement.

Périmètre des models en sortie

Prolongeant le travail d'Austin sur les actes de langage [12], John Searle [184] a proposé une taxonomie de ces derniers, qui est devenue la base de la théorie des actes de langage. Cinq types principaux d'actes de langage expliquent le sens et l'effet de la plupart des énoncés: *assertif*, *commissif*, *directif*, *expressif* et *déclaratif*. Un acte de langage *assertif* est utilisé pour énoncer des informations ou des croyances sur le monde (par exemple, "Il suit également le cours d'apprentissage automatique"); un acte de langage *commissif* est utilisé lorsque le locuteur s'engage à réaliser une action future (par exemple, "Je suivrai le cours d'apprentissage automatique"); un acte de langage *directif* est une tentative visant à amener l'interlocuteur à agir (par exemple, "Souscrivez au cours d'apprentissage automatique, s'il vous plaît!"); un acte de langage *expressif* est utilisé pour énoncer des états psychologiques (par exemple, "Ce cours est vraiment bien expliqué et amusant!"); et un acte de langage *déclaratif* influe un changement immédiat de la situation du monde (par exemple, "Je déclare ce cours en ligne terminé" ou un juge déclarant une personne coupable).

L'analyse conversationnelle [183], domaine de recherche de plus en plus ouvert aux corpus de conversations numériques [207], utilise souvent des actes de langage pour analyser les conversations. Dans une telle approche, *les paires adjacentes* sont étudiées.

Définition 1.7 Paire adjacente: Une paire adjacente est une paire d'actes de langage correspondant à deux tours de conversation effectués par des participants différents et considérés comme étant fonctionnellement liés, tels que invitation-acceptation, salutation-salutation, question-réponse.

Ce transfert de concepts d'une communauté théorique—l'étude des actes du langage, à une

autre—l'étude des conversations, a donné lieu à des positions contradictoires sur la manière dont les actes de langage contribuent à construire des échanges verbaux plus longs [187]. Des questions complexes ont été soulevées: existe-t-il dans les conversations des modèles récurrents qui pourraient caractériser formellement les interactions verbales? Existe-t-il une "grammaire" de la conversation orientant les choix des participants à une conversation, et structurant leur comportement langagier en échanges ciblés, fondés sur des règles [41]?

Certains chercheurs, comme Emanuel Schegloff [183], affirment que les conversations ont leur propre structure intrinsèque, soumise à des règles constitutives. Les actes de langage pourraient ainsi être utilisés pour décrire une telle structure et, en conséquence, une séquence d'actes de langage pourrait être utilisée pour prédire les suivants. D'autres chercheurs, tel que Searle [186], ne croient pas que les conversations suivent des règles intrinsèques et prétendent que les actes de langage ne peuvent pas être utilisés pour étudier les fils globaux des énoncés dans les conversations. Cependant, Andreas Jucker [187] soutient que si les conversations sont effectivement des communications non structurées, ne permettant pas d'anticiper les tours de parole sur la base des précédents, ils peuvent toujours être modélisés en tant que *processus* avec des principes d'organisation locaux.

Définition 1.8 Processus: Un processus est une série d'actions orientées vers un objectif donné. Dans une conversation, le processus peut être vu comme une série de tours de parole, d'énoncés ou d'actes de langage produits pour atteindre un certain objectif, qui est généralement local et non associé à un objectif global de la conversation.

En d'autres termes, le langage, en tant que comportement, pourrait être résumé au travers d'actes de langage. Ensuite, des normes de conversation peuvent être définies en plus des actes de langage, en fonction de l'interprétation locale des échanges verbaux. Conceptuellement, identifier ces normes présente des difficultés, mais des recherches empiriques peuvent s'avérer utiles pour mieux comprendre comment les actes de langage sont reconnus dans le contexte des conversations et comment ils forment des fils de conversation complexes.

En résumé, un texte peut être associé au comportement langagier de différentes manières. Le texte peut contenir un comportement décrit. Par exemple, un chercheur, écrivant sur la démarche scientifique suivie, rend compte des activités réalisées, de leur ordre d'exécution et des conditions possibles. Dans ce cas, le texte est interprété comme un produit de la communication. Par ailleurs, un texte peut révéler un comportement à travers le langage, comme présenté dans cette sous-section. *Les processus des intentions de discours interdépendantes* peuvent modéliser ce type de comportement et éventuellement guider l'échange verbal local.

Définition 1.9 Processus d'intentions de discours interdépendantes: Un processus d'intentions de discours interdépendantes est un processus dont les intentions de discours sont des blocs constitutifs ou des actions. Le processus montre comment les intentions de discours qui le composent sont liées entre elles pour la construction de tours de parole ou d'échanges verbaux.

Les lecteurs interprètent les tours de parole, perçoivent le sens de la communication, et éventuellement réagissent. Pour les travaux menés dans le cadre de cette thèse, c'est cette dernière association entre le texte et le comportement qui est adoptée. Par ailleurs, le point de vue retenu pour l'interprétation du texte et l'extraction de connaissances comportementales au travers d'actes de langage est celui des *lecteurs du texte*.

Objectif de Recherche et Questions

Cette thèse aborde l'objectif de recherche suivant:

Objectif de la recherche: *Proposer une méthode indépendante du corpus pour exploiter automatiquement les communications asynchrones en tant que traces de comportement générées de manière proactive afin de découvrir des modèles de processus de conversations, axés sur des intentions de discours et des relations, toutes deux exhaustives et détaillées*.

Définition 1.10 Indépendant du corpus: la propriété d'être indépendant du corpus implique que la méthode proposée puisse être appliquée à n'importe quel type de corpus conversationnel (i. e. synchrone, asynchrone, emails, discussions de forum, etc.). Implicitement, une méthode indépendante du corpus est également indépendante du domaine d'applications car le corpus peut appartenir à n'importe quel domaine (par exemple, médecine, informatique, etc.). Comme la méthode peut être utilisée dans n'importe quel domaine d'application et potentiellement par des personnes d'horizons variés, le résultat de la méthode en sortie doit également être indépendant du domaine et donc compréhensible pour des non experts en linguistique.

Par son niveau de détail, une représentation peut avoir un granularité fine ou macro. Pour évaluer l'exhaustivité des intentions de discours, la théorie des actes de langage, qui constitue une étape importante en linguistique, est prise comme référence: si les cinq types d'actes de langage sont présents, alors les intentions de discours sont exhaustives; si les intentions de discours correspondent aux classes détaillées des cinq types d'actes de langage, alors elles sont considérées comme ayant une granularité fine.

Afin d'atteindre l'objectif de recherche, trois questions de recherche doivent être abordées:

RQ1: *Comment formaliser des conversations avec des intentions de discours et des relations de processus, toutes deux exhaustives, détaillées et indépendantes du corpus*?

RQ2: *Comment découvrir automatiquement les intentions de discours proposées à partir de conversations asynchrones, indépendamment du domaine et des caractéristiques du corpus?*

RQ3: *Comment découvrir automatiquement des processus d'intentions de discours interdépendantes issus de conversations asynchrones, indépendamment des caractéristiques du domaine et du corpus*?

Contributions et Publications

Plusieurs contributions originales sont faites:

1. Une taxonomie des intentions de discours dérivée de la linguistique—à partir des types d'actes de langage de Searle [184] et des verbes d'actes de langage de Vanderveken [212], pour modéliser la communication asynchrone. Comparée à toutes les taxonomies des travaux connexes, celle proposée est indépendante du corpus, à la fois plus détaillée et exhaustive dans le contexte donné, et son application par des non-experts est prouvée au travers d'expériences approfondies.

   Dans une première version, une taxonomie composée de 6 intentions de discours est proposée pour couvrir la communication publique sur Twitter—des classes d'une granularité plus fine appartenant aux types d'acte langage assertif et directif sont définies. Comparée aux travaux connexes sur la modélisation de tweets publics avec des actes de langage, cette représentation est à la fois indépendante du corpus et plus détaillée [51] (Chapitre 3). Ensuite, en collaboration avec un linguiste [41], la taxonomie précédente est étendue à 18 classes d'intentions de discours afin de pouvoir être appliquée à tout type de communication. De plus, des critères d'opposition des intentions de discours sont ajoutées à la taxonomie pour structurer les classes et permettre une annotation manuelle plus facile (Chapitre 4).

   Les expériences d'annotation manuelle avec des experts et des non-experts montrent une cohérence dans la perception de la plupart des intentions de discours dans les deux itérations de création de la solution. De plus, l'utilisation rare de la classe "*other*" par les annotateurs confirme de manière empirique que la taxonomie est exhaustive pour modéliser la communication dans deux types de corpus: les tweets [51] et les conversations de forum [41].

2. Une méthode automatique, indépendante du corpus, pour annoter les énoncées de communication asynchrone avec la taxonomie des intentions de discours proposée, conçue sur la base d'un apprentissage automatique supervisé [50, 51, 57]. Pour cela, deux corpus "ground-truth" validés sont créés (un avec des tweets[4] et un avec des conversations de forum[5]) et trois groupes de caractéristiques (discours, contenu et conversation) sont conçus pour être utilisés par les classificateurs. En particulier, certaines des caractéristiques du discours sont nouvelles et définies en considérant des moyens linguistiques pour exprimer des intentions de discours, sans s'appuyer sur le contenu explicite du corpus, le domaine ou les spécificités des types de communication asynchrones.

   Les expériences sur le corpus Twitter [51] (Chapitre 3) montrent que les intentions de discours peuvent être découvertes de manière satisfaisante en utilisant les classificateurs SVM Linéaire ou Régression Logistique avec uniquement des caractéristiques de discours (F1-mesure compris entre 0,73 et 0,80). Par rapport aux autres travaux connexes, la méthode proposée annote chaque tweet avec des intentions de discours multiples et plus

[4] `http://tinyurl.com/hk9t83y`
[5] `http://tinyurl.com/yc8mjtb2`

détaillées, en ne s'appuyant que sur des caractéristiques de discours indépendantes du domaine ou du corpus.

Les expériences sur le corpus Reddit [50] (Chapitre 5) montrent que les mêmes classificateurs, SVM Linéaire et Régression Logistique, conduisent à de meilleurs résultats, dans une configuration "one versus all", avec à la fois des caractéristiques de contenu et de discours (F1-mesure entre 0,53 et 1). En outre, afin d'améliorer l'identification des intentions de discours avec des F1-mesures faibles, deux stratégies sont proposées et évaluées: utiliser une classification multi-label pour les paires d'intentions de discours qui co-occurrent fréquemment, ou compléter le corpus d'apprentissage par des exemples d'intentions de discours les moins représentées, même à partir de corpus externes hétérogènes. Bien que les caractéristiques du contenu soient nécessaires pour l'annotation précise des phrases, l'analyse détaillée réalisée montre que ce sont des n-grammes génériques, indépendants du corpus ou du domaine et liés aux intentions de discours, qui sont les plus importants pour la prédiction. De plus, cette approche montre une robustesse satisfaisante pour découvrir les intentions de discours d'autres corpus, y compris la communication synchrone. Comparée aux autres solutions existantes d'annotation des conversations de forum utilisant un apprentissage automatique supervisé, cette approche est générale et la plus exhaustive et détaillée. Cette solution utilise des classes détaillées pour multi-labeler les phrases. De plus, des expériences approfondies ont été menées pour sélectionner la meilleure configuration pour la classification multi-label, explorer des stratégies permettant d'améliorer les performances de cette classification et évaluer la validité externe de l'approche.

La pertinence et l'impact des résultats obtenus des deux séries d'expériences sur la médecine sont également abordés, en montrant diverses directions pour intégrer cette approche dans des technologies et des études concernant la recherche et la diffusion d'informations de santé, ainsi que la médecine narrative.

3. Une méthode automatique basée sur la fouille de processus [1] conçue pour générer des modèles de processus d'intentions de discours interdépendantes à partir de tours de parole, annotés avec plusieurs labels par phrase [50, 58] (Chapitre 6). Comme la fouille de processus repose sur des logs d'événements structurés et bien définis, un algorithme est proposé pour produire de tels logs d'événements à partir de conversations. Par ailleurs, d'autres solutions pour transformer les conversations annotées avec plusieurs labels par phrase en logs d'événements, ainsi que l'impact des différentes décisions sur les modèles comportementaux en sortie sont analysées afin d'alimenter de futures recherches. De plus, la technique Fuzzy Miner [85], implémentée dans l'outil Disco, est utilisée pour exploiter les logs produits. En effet, elle convient bien aux processus non structurés et est hautement interactive et visuelle.

   Comparés aux travaux existants modélisant les relations entre les intentions de discours à

partir de conversations annotées sous forme de diagrammes de transition, les modèles de processus découverts à l'aide de la fouille de processus sont plus généralsetplus complets, capturant également d'autres types de relations—boucles significatives de longueur 2, séquences indirectes et concourentes. Ces processus peuvent être explorés de manière interactive à différents niveaux de détail avec les outils de fouille de processus. Parmis les rares travaux existants sur la modélisation automatique des relations entre les intentions de discours, seulement un utilisait la fouille de processus [214]. Ce dernier repose sur un scénario plus simplifié et dépendant du domaine: chaque tour de parole est annoté avec un seul acte de langage associé à des discussions de cours en ligne. La solution développée dans cette thèse va au-delà et propose un scénario complémentaire et général. Contrairement à la précédente méthode existante, celle développée est applicable aux conversations annotées avec plusieurs libellés par phrases provenant de tout type de domaine.

L'évaluation qualitative de l'approche proposée débute par une analyse de la pertinence des modèles de processus obtenus pour la médecine. Puis dans un second temps, les résultats d'une étude observationnelle préliminaire menée avec un chercheur en linguistique sont présentés afin d'évaluer la pertinence pour l'analyse conversationnelle.

Les résultats montrent que l'approche proposée conduit à la découverte de modèles intéressants de construction de tours de parole dans une conversation et à la formulation de nouvelles hypothèses concernant les actes de langage et les conversations.

4. L'étude menée en Chapitre 2 est la seule revue systématique à ce jour sur la modélisation automatique des conversations asynchrones avec des actes de langage.

   Pour cela, un cadre d'analyse et de comparaison de la littérature existante sur le sujet est développé. Il comprend plusieurs aspects : taxonomies d'intention de discours, relations entre les intentions de discours, domaines d'application et bénéficiaires potentiels, création et validation manuelle de corpus. Ce cadre intègre aussi des approches automatiques pour modéliser des conversations avec des intentions de discours et avec des relations entre les intentions de discours, incluant divers groupes de caractéristiques utiles à la classification des labels.

Cette thèse contient les publications et les manuscrits suivants :

- Epure, E.V., Compagno, D., Salinesi, C., Deneckere, R., Bajec., M., Zitnik, S. (2018). Process Models of Interrelated Speech Intentions from Online Health-related Conversations. *Artificial Intelligence in Medicine*. doi = https://doi.org/10.1016/j.artmed.2018.06.00

- Compagno, D., Epure, E.V., Deneckere-Lebas, R., Salinesi, C. (2018). Exploring digital conversation corpora with process mining. *Corpus Pragmatics, 2*(2) 193-215.

- Epure E.V., Deneckere R., Salinesi C. (2017). Analyzing Perceived Intentions of Public Health-Related Communication on Twitter. In ten Teije A., Popow C., Holmes J., Sacchi L.

(Eds), *Proceedings of 16th Conference on Artificial Intelligence in Medicine* (pp. 182-192). Vienna, Austria. Springer.

- Epure, E.V., Zitnik, S., Compagno, D., Deneckere, R., Salinesi, C. (2017). Automatic analysis of online conversations as processes, presented at *Journées Analyse de Données Textuelles en Conjonction avec EDA 2017*, Lyon, France.

- Epure, E.V., Deneckere, R., Salinesi, C. (2018): *Automatically modeling asynchronous conversations with speech intentions: A systematic study*. Manuscript in preparation.

- Epure, E.V., Salinesi, C., Deneckere, R., Bajec., M., Zitnik, S. (2018). *A linguistic approach to reveal tacit requirements engineering knowledge from online discussions*. Manuscript in preparation.

Pour des raisons de cohérence thématique, cette thèse ne comprend pas les articles publiés suivants—travaux réalisés en parallèle du sujet traité dans cette thèse :

- Epure E.V., Kille B., Ingvaldsen J.E., Deneckere R., Salinesi C., Albayrak S. (2017) Recommending Personalised News in Short User Sessions. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)* (pp. 121-129). Como, Italy. ACM.

- Epure E.V., Kille B., Ingvaldsen J.E., Deneckere R., Salinesi C., Albayrak S. (2017) Modeling the Dynamics of Online News Reading Interests. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)* (pp. 363-364). Bratislava, Slovakia. ACM.

- Epure E.V., Ingvaldsen J.E., Deneckere R., Salinesi C. (2016). Process mining for recommender strategies support in news media. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-12). Grenoble, France. IEEE.

- Epure, E. V., Deneckere, R., Salinesi, C., Kille, B., Ingvaldsen, J.E. (2017). Devising News Recommendation Strategies with Process Mining Support. Presented at *Interdisciplinary Workshop on Recommender Systems*. Paris, France.

- Epure E.V., Martin-Rodilla P., Hug C., Deneckere R., Salinesi C. (2015). Automatic process model discovery from textual methodologies. In *2014 IEEE Ninth International Conference on Research Challenges in Information Science (RCIS)* (pp. 19-30). Athens, Greece. IEEE.

- Gonzalez, C., Martin-Rodilla, P., Epure, E.V. (2016). Formalization and Reuse of Methodological Knowledge on Archaeology across European Organizations. Presented at *the 44th Computer Applications and Quantitative Methods in Archaeology Conference*. Oslo, Norway.

Le lecteur le souhaitant, peut consulter une sélection représentative de ces publications dans les Annexes G et H. Ces articles se concentrent sur la découverte automatique du comportement en

tant que modèles de processus à partir de données, mais dans des contextes différents : à partir de logs Web pour créer des algorithmes de recommandation de nouvelles en ligne [54, 56] et aussi à partir de textes décrivant des comportements afin de favoriser le partage de connaissances dans le domaine des sciences humaines [57].

Enfin, une illustration simplifiée de la solution créée est présentée à l'Annexe B.

Organisation de la Thèse

Le chapitre 2 présente les résultats des deux questions de recherche concernant la littérature existante. En effectuant une étude systématique de la littérature, les travaux associés qui modélisent automatiquement les conversations asynchrones avec des actes de langage sont passés en revue. En outre, une discussion sur la mesure dans laquelle ces travaux répondent aux objectifs et aux propriétés de la solution définis dans la phase d'investigation de la problématique est présentée.

Le Chapitre 3 présente la contribution suivante de la thèse. Une méthode indépendante du domaine et du corpus pour modéliser automatiquement les tweets avec des intentions de discours exhaustives et détaillées est conçue, répondant ainsi aux première et deuxième questions de recherche (**RQ1** et **RQ2**). Des expériences y sont menées pour prouver la validité interne. La validité externe y est aussi discutée.

Le Chapitre 4 traite des limites de la solution présentée au Chapitre 3 en ce qui concerne le caractère exhaustif et détaillé des intentions de discours proposées (**RQ1**). Une taxonomie étendue de 18 intentions de discours qui respecte tous les critères établis dans la phase d'investigation de la problématique est proposée et validée.

Le Chapitre 5 est étroitement lié au précédent, formant la deuxième itération du cycle de conception. Des classificateurs d'apprentissage automatique supervisé, ainsi que diverses configurations de classification et ensembles de caractéristiques sont évalués pour l'annotation automatique des phrases issues de conversations de forum avec la taxonomie proposée des intentions de discours. Ainsi, la deuxième question de recherche est ré-examinée (**RQ2**). Les validités internes et externes sont déterminées par des expériences.

Le Chapitre 6 présente la troisième itération du cycle de conception, permettant d'atteindre la conception d'une solution satisfaisante et complète (**RQ3**). La fouille de processus est choisie pour découvrir des processus d'intentions de discours interdépendantes issus de conversations asynchrones annotées. Notamment, il y est proposé une méthode pour transformer les conversations annotées en logs bien définis qui sont requis en entrée des techniques de fouille de processus. La

pertinence de l'approche globale est illustrée dans le domain de la médecine, puis validée dans une étude observationnelle pour la linguistique / l'analyse conversationnelle.

Le Chapitre 7 présente une synthèse des contributions répondant aux questions de recherche proposées, et identifie les limites, ainsi que les orientations pour de futurs travaux.

CHAPTER 1

INTRODUCTION

1.1 Research Context

The steep rise of Internet use in everyday personal and professional life is highly evident nowadays [143]. As a result, each individual generates larger and more diverse amounts of data than ever before. When posting on social media [11], commenting news online [191], writing work reports [177] or uploading code to repositories [6], people create and share content *pro-actively*. Besides this, a *passive* data generation also exists. Human-computer interactions are frequently traced by the underlying software [1, 3, 6]. The objective is to learn about end-users and through this deliver better services and support to people in their digital experience.

The proliferation of digital data has enabled scientific and practitioner communities to create new data-driven technologies to learn about end-users, in particular about their *behavior* [1, 3, 6, 11, 191].

Definition 1.1 Behavior: Behavior is defined as "the way humans act and interact"[1].

Many areas and applications revolve around automatically modeling behavior. In capital markets, investor behavior is mined as trading patterns to identify fraud or speculative trading [47]. In counter-terrorism, behavior discovery from data supports investigations and enforces prevention [38]. In recommender systems for e-commerce, movies, music, news and so forth, user behavior is expressed with regard to the consumed items and to the similarity to other users [3]. In social networks, the collective behavior is analyzed by outlining social structures in order to get insights into human relationships [104]. The anomaly detection community proposes methods to discover behavioral outliers from data with applications in finance, health-care and cyber-security [32]. Moreover, behavior is modeled as processes and automatically discovered from data for business management and information system evolution [1].

[1] https://www.nature.com/subjects/human-behaviour

The majority of these technologies created for human behavior analysis extensively derive value from data logs passively generated during the human-computer interaction [1, 3]. A particularity of these *behavioral traces* is that they are logged and stored according to a clearly defined structure. This type of trace is called *structured* or *semi-structured*—the latter being a frequent terminology for web logs. The existing structure plays a role in the extraction of the behavioral knowledge and the creation of subsequent technologies. In contrast, pro-actively generated traces are highly *unstructured* and represent the overwhelming majority of the existing digital data[2]. Furthermore, the unstructured data is mainly found as *text*.

1.2 Problem Statement

To date, despite the digital text's prevalence and the relevance of behavioral knowledge to many domains, the *digital text* is still insufficiently studied as *traces of human behavior* to automatically reveal *extensive insights into behavior* [222]. Text analytics technologies focus on core natural language processing tasks such as word sense disambiguation and parsing, and on specific information extraction tasks, such as entity recognition [60, 134]. To some extent, behavioral aspects have been investigated through text categorization and clustering [60, 134]. A very extensive application is sentiment analysis, also known as opinion mining, which proposes techniques to analyze verbal behavior to identify and quantify affective states [152].

However, the association between behavior and text is a much more complex topic [222], extending beyond the discovery of positive and negative opinions. Text is one of the most frequent representations and means of communication, especially digitally. Communication, as a production of language, is one of the main ways in which people act and interact. This topic has been thoroughly studied in humanities and social sciences [9, 12, 95, 184, 197]. In particular, the 20th century has seen increasing interest into studying communication as a form of human behavior and interactions, in the philosophy and linguistics domains.

In his lectures "How to do things with words", John Austin [12] pioneered this new view on the use of language in linguistics. Specifically, he put forward the idea that language is not only used for communicating true or false facts, detached from social understanding and action; but also that, through speech, people realize actions and that the mutual recognition of the intentions of these actions is the basis of human interaction through language. *Speech acts* have been coined to conceptualize and differentiate the *speech intentions* realized by the production and interpretation of written or spoken *utterances* [197].

Definition 1.2 Utterance: An utterance has multiple definitions[3]. In speech, an utterance is the unit of speech bounded by speaker's silence or pauses. In dialogues, an utterance can be a turn. In linguistics, an utterance is the unit of speech or of written text being studied.

[2] A widely accepted estimation is that 80% of the digital data is unstructured [18].
[3] `https://glossary.sil.org/term/utterance`

Definition 1.3 Speech intention: A speech intention is used for denoting the linguistic intention of a speaker for producing an utterance such as *promise*, *request*, *greet*, *assert* etc. It is also known as the illocutionary force of an utterance [184].

Definition 1.4 Speech act: A speech act is an act performed by a speaker when making an utterance. It includes the uttering or writing of words and the speech intention associated with the uttered or written words.
For instance, a teacher's utterance "Hand in the homework by Friday." is a speech act and its associated speech intention can be included in the class *request*. Thus, spoken or written language can be analyzed as human behavior encompassing acts brought about through the wording.

What makes the extraction of human behavioral knowledge from text very challenging is the nature of the *input* [60]. Apart from lacking structure, text can appear in very diverse styles depending on the discourse function—a blog article is written differently than a social media post; on the enabling technology—Twitter, for instance, has led to a unique form of discourse by enforcing a maximum of 140 characters per message; and, ultimately, on the characteristics of each individual person's discourse. Moreover, the basic technologies for processing text, for instance, for conducting automatic morphological or semantic analyses [134], are still imperfect. Consequently, deriving more intelligent text analytics solutions is often restricted by the effectiveness and efficiency of the underlying text processing software.

While the challenge may be technological regarding the input processing, the limitations of the *output*, the discovered behavioral knowledge, come from other sources. When conceptualizations of behavior from linguistic theoretical frameworks centered on speech acts are adopted, the practical solutions are frequently designed for specific applications or problems. For example, speech acts are addressed in the specific context of emails to enhance email tracking tools [28, 40] or in the context of questions and answers forums, to support the information seeking and discussion management [10, 118, 163]. There are also other approaches for behavior discovery from text [163, 213, 234], adopting general representations of speech acts, exactly aligned with the existing theoretical frameworks from linguistics [184]. In these cases, the level of detail of the discovered behavioral knowledge is often low, as will be shown in detail in Chapter 2. Conversely, very detailed definitions of behavior through speech acts have also been used in automatic approaches [110, 196], but their very specific level of representation presents a challenge in their interpretation and re-use by non-experts in discourse analysis.

1.3 Research Scope, Objective and Questions

Before presenting the research objective and questions, a more thorough discussion on digital communication and its scope in the current thesis is presented in Subsection 1.3.1. Then, other relevant theoretical aspects of behavior through language and the behavior's conceptualization through speech intentions are introduced in Subsection 1.3.2.

1.3.1 Input Scope

Everyday life is increasingly connected with digital life. As a result, the communication practices have been impacted. Communication takes place through modern channels and formats dictated by the new digital order [218]. Text is considered, on the one hand, as a *product of communication*, which conveys the communication's content, and on the other hand, as a *verbal structure*, the wording that physically represents the communication [95]. Hillesund claims that these two conceptual views on text are not contradictory and proposes a unified definition: "text is a visual representation of verbal information" [95].

In general, communication can be characterized through multiple dimensions: modality, medium, register, number of conversation participants and immediacy [100]. The modality of communication distinguishes between spoken and textual exchanges. The medium of communication refers to the means over which communication takes place; specifically, it includes analog media, such as phone conversations, as well as digital media, which is any type of computer-mediated communication. The register refers to the formality of the language that is used; it could be informal or formal. Immediacy relates communication to the the traits of being *synchronous*, such as in chats and instant messaging, or *asynchronous*, such as in emails and social media posts. Furthermore, another way to delimit communication types is the communicative goal, which represents a common objective followed by interlocutors during conversation [100]. Examples of communicative goals are searching for information, providing instructions or sharing a story.

In the current work, only *text-based communication* over *digital media* is covered. Once a message has been uttered, the stored textual data becomes a *trace* of the message. Exclusion criteria based on the number of participants, the communication register and the communicative goal are not imposed. However, *asynchronous communication* is the input that will be focused on.

Definition 1.5 Asynchronous communication: Asynchronous communication includes conversations which could take place over longer periods of time and in which individuals can participate at any time by replying to any existing message. As the conversations are not in real-time and the interventions are loosely structured, threads are often used to structure these conversations and to help individuals to follow them.

Definition 1.6 Conversation thread: A conversation thread consists of a list of related messages that appear in the conversation in the following manner: the original message, a reply to the original message, a reply to the previous reply and so on. If a conversation is imagined as a tree with the root being the original message and the nodes being the other messages, with each parent-child relation marking the fact that the message in the child node is a reply to the message in the parent node, then a conversation thread contains all the messages corresponding to a path from a root node to a leaf node.

There is still ongoing research to understand the nature of asynchronous communication. For example, as reported by Goldstein et al. [79], empirical analyses reveal that this genre is hybrid, embodying characteristics of both speech and writing. Specifically, it is very similar to letters, but

when judging on the level of interaction, it is much closer to synchronous discourses, which are conversational. Synchronous dialogues converge to relatively standard ways of communication. However, there is much more diversity in asynchronous exchanges.

Compared to synchronous conversations, asynchronous communication is composed of longer messages, containing both conversational and non-conversational content such as attached documents in emails or snippets of code in forum posts [194]. Also, the asynchronous content tends to be more often expository text [163] and to have associated "stereotypical templates or conventions" per specific subtype, such as emails or tweets [138, 234] . Synchronous communications have a flat structure and tend to have clear local goals (e.g. messages exchanged to negotiate a decision). On the contrary, in asynchronous communication, the goals vary and can be mixed within same messages [100] (e.g. the same email could be used to plan tasks and deliver information). Also, the structure of asynchronous conversations is usually a graph, where the nodes are the turns while the edges are the links between consecutive turns in conversational threads. Participants intervene in discussion in a loosely structured order and could reply to much further messages than the immediate ones [199].

Consequently, asynchronous communication compared to synchronous communication comes with a lot of variety and implicitly with challenges and particularities that impact the design of a solution. Moreover, if a suitable solution for asynchronous communication is found, then it will be transferable to synchronous communication, as many of the synchronous-related characteristics are embedded in or simplifications of the former type. Therefore, this investigation starts with *asynchronous text-based conversations* and will be shown in detail how these conversations represent traces of behavior.

1.3.2 Output Scope

Extending the work of Austin on speech acts [12], John Searle [184] proposed a speech act taxonomy, which became the basis for the Speech Act Theory. Five main types of speech acts account for the meaning and effect of most existing utterances: *assertive*, *commissive*, *directive*, *expressive* and *declarative*. An *assertive* speech act is used to state information or beliefs about the world (e.g. "He follows the machine learning course too"); a *commissive* speech act is used to commit the speaker to a future action (e.g. "I will follow the machine learning course"); a *directive* speech act is an attempt to get the interlocutor to do something (e.g. "Subscribe to the machine learning course too, please!"); an *expressive* speech act is used to state psychological states (e.g. "This course is really well explained and fun!"); and a *declarative* speech act affects an immediate change on the world's state of affairs (e.g. "I hereby declare this online course finished." or a judge declaring a person guilty).

Conversation Analysis [183], a domain of research increasingly open to digital corpora [207], often uses speech acts to analyze conversations. In such an approach, *adjacency pairs* are studied.

Definition 1.7 Adjacency pair: An adjacency pair is a pair of speech acts corresponding to

two conversation turns by different conversation participants that are considered functionally related, such as invitation-acceptance, greeting-greeting, question-answer.

This transfer of concepts from one theoretical community—the study of speech acts, to the other—the study of conversations, has produced conflicting stances about how speech acts contribute to building longer verbal exchanges [187]. Complex questions have been raised: are there recurring patterns in conversations which could formally characterize verbal interactions? Does a "grammar" of conversations exist directing the choices of conversation participants and structuring their speech behavior into purposeful rule-based exchanges [41]?

Some scholars, such as Emanuel Schegloff [183], claim that conversations have a proper intrinsic structure, which is subject to constitutive rules. Speech acts could thus be used to describe such a structure and, as a consequence, a sequence of speech acts may be used to predict the following ones. Other scholars, such as Searle [186], do not believe that conversations follow intrinsic rules and claim that speech acts cannot be used to study the global threads of utterances in conversations. However, Andreas Jucker [187] argues that if conversations are indeed unstructured communication, not allowing turns to be anticipated based on the previous ones, they can still be modeled as *processes* with local organization principles.

Definition 1.8 Process: A process is a series of actions directed to some goal. In a conversation, the process could be seen as a series of turns, utterances or speech acts produced to achieve a certain goal, which is usually localized and not associated with a global conversation goal.

In other words, language as a form of behavior could be abstracted through speech acts and conversational norms can be defined on top of speech acts, depending on the local interpretation of the verbal exchanges. Conceptually identifying these norms is challenging, but empirical research may prove useful to better understand how speech acts are recognized within conversational contexts and develop complex threads in conversations.

To sum up, text can be related to behavior in different ways. Text can contain reported behavior. For instance, a researcher writing about the followed scientific approach reports the undertaken activities, the order in which they were performed and the possible conditions. In this case, text is interpreted as a product of communication. Further, text reveals behavior through language as discussed in this subsection. *Processes of interrelated speech intentions* could abstract this type of behavior and possibly guide the local verbal exchange.

Definition 1.9 Process of interrelated speech intentions: A process of interrelated speech intentions is a process which has speech intentions as building blocks or actions and exposes how the composing speech intentions relate to each other for building turns or verbal exchanges.

Readers interpret the turns, perceive the communication meaning and eventually react. In the current work, this latter association of text and behavior is adopted. Furthermore, the viewpoint on text interpretation and behavioral knowledge extraction with speech acts is taken from the *perspective of the text readers*.

1.3.3 Research Objective and Questions

This thesis addresses the following research objective:

Research Objective: *Propose a corpus-independent method to automatically exploit the asynchronous communication as pro-actively generated behavior traces in order to discover process models of conversations, centered on comprehensive speech intentions and relations.*

Definition 1.10 Corpus-independent: The property of being corpus-independent entails that the proposed method could be applied to any type of conversational corpus (e.g. synchronous, asynchronous, emails, forum discussions etc.). Implicitly, a corpus-independent method is also domain-independent because the corpus could belong to any domain (e.g. medicine, software engineering etc.). As the method could be used in any application domain, potentially by people with varied backgrounds, the output of the method must also be domain-independent and thus comprehensible by non-experts in linguistics.

Definition 1.11 Comprehensive: Comprehensiveness in this work refers to two desired properties of speech intentions and process relations. On the one hand, it refers to exhaustiveness. On the other hand, it refers to the level of detail, making a representation fine-grained or coarse-grained. For instance, to assess the comprehensiveness of speech intentions, the speech act theory, which is a milestone in linguistics, is taken as a reference: if all five types of speech acts are present, then the speech intentions are exhaustive; if the speech intentions are detailed classes of the five speech act types, then the speech intentions are considered fine-grained.

In order to achieve the research objective, three research questions must be addressed:

RQ1: *How to formalize conversations with comprehensive and corpus-independent speech intentions and process relations?*

The output of the proposed method is knowledge that represents how humans behave through written language. Based on linguistic theory of language, the adopted definition of behavior is processes of interrelated speech intentions. Moreover, speech intentions could be expressed through Searle's five speech act types [184]: assertive, expressive, directive, commissive and declarative. Nevertheless, this representation still needs to be completed in order to specify the process interpretation of conversations. Additionally, the final representation should comply with two established properties of the envisioned solution: the target behavioral knowledge must be comprehensive, but corpus-independent. Consequently, a general diversification of the speech intentions is essential, resulting in a fine-grained *taxonomy of speech intentions*. A trade-off between the level of detail and the interpretation and the use of the taxonomy by non-experts is also desired. Let's notice that no constraint on the knowledge representation related to the type of communication is imposed at this step. Ultimately, as already mentioned, a corpus-independent taxonomy of speech intentions implies that it should be applicable to any domain and any type of corpus. The same properties, comprehensiveness and corpus-independence, must be ensured for the process formalization too.

RQ2: *How to automatically discover the proposed speech intentions from asynchronous conversations independently of the domain and corpus characteristics?*

Once a valid representation of behavioral knowledge is designed, the next step is to automatize the knowledge discovery. At this point in the design process, the focus on asynchronous conversations is relevant. As previously discussed, asynchronous communication entails many particularities that render more challenging the technology creation: it is very diverse compared to chats (emails, tweets, other types of social media and forum posts); it appears in complex structures such as graphs; it contains both conversational and non-conversational elements; it has associated communicative goals co-occurring within same conversation turn compared to a single local communicative goal spanning multiple turns in chats.

To annotate asynchronous conversations with speech intentions means to associate speech intentions to individual discourse units. However, multiple decisions have to be made and motivated: What should the discourse unit be: a complete turn, a sentence or a part of a sentence? Should the association be one-to-one, meaning that each discourse unit has related an unique speech intention, or should more speech intentions be allowed per unit? Could domain and corpus-independent characteristics of discourse signaling specific speech intentions be identified and effectively used? Which text mining and machine learning approaches should be used for annotation? And eventually, how to enable the validation of the proposed solution?

The adopted approach is to first annotate asynchronous conversations with multiple speech intentions per utterance[4], followed by transforming the annotated conversations in processes—covered in the third research question. Another approach would have also been possible, namely to concurrently discover speech intentions and their relations. In the literature, the direct discovery of processes from conversations has been realized with unsupervised machine learning approaches, using for instance Hidden Markov Models [107, 109, 154, 171]. However, as it will be exposed in more details in Chapter 2, multiple challenges exist. In particular, humans are required to interpret the obtained clusters in order to associate speech intentions, the algorithmic complexity is high and the evaluation of such models is much more complex. Additionally, an advantage of not jointly pursuing the discovery of speech intentions and processes is the flexibility to use the results separately. Only speech intention discovery could be very useful for deriving descriptive statistics about behavior through language, for analyzing communication in different communities or on various platforms and for supporting the development of intelligent technologies such as chat bots and search engines allowing to search by both keywords and speech intentions [10, 14, 149].

RQ3: *How to automatically discover processes of interrelated speech intentions from asynchronous conversations independently of the domain and corpus characteristics?*

The discussion of **RQ2** stated that the discovery of processes from asynchronous conver-

[4]An utterance in the current work is a sentence, apart from the case of tweets when an utterance is a tweet.

sations is realized after annotating conversations with speech intentions. By answering **RQ1**, a corpus-independent and comprehensive process representation as a process meta-model is selected. Further, the objective is to propose a method to automatically transform the annotated asynchronous conversations in this representation.

If each turn in conversation has associated multiple speech intentions, the input of the designed method could be a graph with each node representing a set of speech intentions. Simplifications of the input are possible by considering conversations as sequences or trees and by associating one single speech intention with each turn. Conversely, in a more complex representation of the input, a conversation could be a forest of graphs when multiple unrelated discussions subsequently take place within the same conversation. Additionally, each turn could be represented as sequences of interrelated speech intentions if each sentence of the turn, annotated with multiple interrelated speech intentions, is mapped on a sequence element.

A process can be derived in different ways. A process could describe individual behavior as in a single conversation turn such as an email or a forum message. Also, a process could reflect behavior as an individual human interaction by being mined from a single conversation. By contrast, processes could reveal *collective* behavior by aggregating interactions from multiple conversations or multiple turns, thus exposing human communication strategies.

The diversity of input representations and output interpretations should be carefully considered when designing a solution. In addition, the usability and relevance of the solution to reveal insights into behavior in a wide range of domains should be illustrated.

1.4 Research Approach

To answer the proposed research questions multiple research methods are followed. These specific methods are introduced in the chapter reporting their use. However, the overall research work for this thesis has been conducted under the principles of *design science*, which is introduced in Appendix A.

Problem investigation. The first phase of design science consists in understanding the application domain problems and formulating the solution requirements / acceptance criteria. Medicine and linguistics will be introduced as potential application domains. However, the current work is in fact driven by a technological shortcoming that if solved could be an opportunity for a larger class of practical problems, thus not only for a specific domain. This particular case is identified by Wieringa [223] as *solution-driven investigation*. This type of problem investigation starts with identifying the goals and current technologies. Then, it continues with defining the new solution requirements as desired properties. The problem investigation phase is summarized in Table 1.1.

Table 1.1: Summary of the solution-driven problem investigation in the current work.

Solution goals	**Solution properties**	**Related technologies**
Input: text (asynchronous conversations)	Automatic	Text mining
	Corpus-independent	Machine learning
Output: behavioral knowledge (interrelated speech intentions)	Effective (correct, complete comprehensive, relevant)	Natural language processing

The current problem statement outlines multiple issues. Existing technologies for automatic human behavior analysis have limitations with regard to the *input*–the unstructured textual data as behavioral traces is underexploited, and to the *output*–the discovered behavioral knowledge is limited. The output limitations emerge from multiple causes:

- The behavioral knowledge is either uncomprehending, focusing on certain aspects of behavior such as affective states;
- The behavioral knowledge targets specific domains, corpora and use cases;
- When more general conceptual frameworks are adopted, such as based on speech act theory, these are either too high-level to allow for detailed analyses or too detailed, making it challenging to be re-used in empirical studies especially by non-experts.

Consequently, the *goals* of the envisioned solution refer to both input and output. The input is text produced in any type of asynchronous communication. Grounded in widespread theoretical interpretation of language, the output conceptualizes this type of behavior through speech intentions and processes of interrelated speech intentions.

The most related current *technologies* for these goals are text mining, natural language processing and machine learning. Their relevance is motivated by the nature of the input data and the research objectives, as detailed in the presentation of the research questions. Techniques and approaches belonging to these domains are considered for creating the solution.

Finally, the desired *properties* of the solution are restated. The method must be automatic and corpus-independent. Then, the solution must be effective. The effectiveness is defined in relation to the obtained behavioral knowledge. Specifically, the behavioral knowledge representation must be based on comprehensive speech intentions and process relations and must be relevant for at least one application domain. Also, the results must be correct and complete.

This phase corresponds to the relevance cycle in the design science paradigm. Activities from the rigor cycle were also applied when defining the link between text and behavior and for discovering the related technologies. Two *knowledge questions* were answered: "What is the addressed problem?" and "What are the goals and properties of an envisioned solution?".

Solution design, validation and evaluation. The design of a solution consists in outlining the means to achieve the goals, while ensuring the properties formulated in the problem

investigation [223]. The validation of a design is linked to *knowledge questions* and its objective is to identify if the proposed artifacts satisfy the solution goals and properties defined in the problem investigation phase (internal validation). Additionally, an external validation that investigates if the solution used in different contexts satisfies the goals and properties defined in the problem investigation phase also takes place. Then, the validated designs are implemented and the obtained artifacts are thoroughly evaluated. In addition, the limitations of the proposed artifacts are discussed. Four types of threats are addressed regarding:

- The *internal validity*—it enforces the confidence in the causal relationship between the input and output variables, while mitigating extraneous factors accounting for the result.
- The *external validity*—it investigates the extent to which the results can be generalized.
- The *construct validity*—it ensures that the measurements used are correctly designed for what they are claimed to measure and that the concepts modeling the treatments and the outcomes correctly reflect the cause and effect constructs.
- The *conclusion validity*—it reflects the statistical confidence in the correctness of the results and of the reached conclusions.

In this phase, the proposed research questions (**RQ1**, **RQ2** and **RQ3**), which are *design questions*, are answered. As a preparation step preceding the design, one knowledge question is addressed through a systematic literature study presented in Chapter 2: "How do the related works automatically model asynchronous conversations with speech intentions?" (**RQ_L1**). Additionally, the internal validation of the related works is conducted. Thus, a second knowledge question related to the existing related solutions complements the previous one: "How well do the related works meet the goals and properties of the solution, identified in the problem investigation phase?" (**RQ_L2**).

A full specification of the solution design is rarely an one-step activity. The evaluation results and the solution limitations can trigger changes in the solution design leading to new iterations. In the current work, *three design iterations* were conducted:

1. In *the first iteration*, a solution design, which addresses the first two research questions (**RQ1** and **RQ2**), was created. Specifically, comprehensive and corpus-independent speech intentions were derived from the theoretical background and corpora analysis to formalize tweets. The proposed 6 classes of speech intentions were then validated in a manual annotation experiment, assessing if two human annotators consistently associated the proposed speech intentions with the tweets. Then, a valid dataset was derived from the tweet corpus used in the manual annotation experiment. Finally, a domain and corpus-independent method relying on supervised machine learning was created to automatically annotate individual tweets with multiple speech intentions. To this end, characteristics of

speech intentions were defined manually which together with several types of classifiers were examined for efficacy.

2. In *the second iteration*, an improved solution design was created (**RQ1** and **RQ2**), considering the results obtained in the previous iteration. Though promising, the previous design was incomplete because the speech intentions proposed to conceptualize tweets as a form of communication did not cover all speech act types[5]. Therefore, the proposed speech intentions were extended from 6 classes to 18 classes by analyzing a richer type of corpus with regard to the range of conveyed speech intentions: forum conversations. Additionally, oppositional traits between the speech intentions were defined to be used in the manual classification. Following a similar method as in the first iteration, the speech intentions were validated in a manual annotation experiment, this time on forum conversations and with two experts and ten external annotators. The annotation was per sentence and multi-label. Then, a ground-truth corpus was created and used in the training and evaluation of the domain and corpus-independent automatic method to discover the extended speech intentions. More features were defined and more thorough experiments compared to the previous iteration were deployed to improve the effectiveness of speech intention discovery and to assess the robustness of the classifiers on external, heterogeneous corpora.

3. In *the third iteration*, the previous design was completed in order to answer the third research question (**RQ3**). The validation and evaluation of the previous artifacts, corresponding to the first two research questions, reached satisfactory results in the second iteration. However, the process perspective on conversations was not yet tackled. Consequently, the solution design was extended to fully achieve the established goals and ensure the defined properties. The approach taken was to use process mining to discover conversational processes. Process mining [1] proposes a suite of techniques to discover and model human behavior from structured event logs generated during the interaction with information systems. Although the annotation of conversations with speech intentions resulted in a more structured output, the input required by process mining is conceptually different. Consequently, to answer this research question, a mapping method to transform the annotated conversations in well-formed structured logs, which could allow process mining tools to discover processes of interrelated speech intentions, was created.

To summarize, this phase is presented in detail as follows: the literature study in Chapter 2 (**RQ_L1**, **RQ_L2**); the first iteration in Chapter 3 (**RQ1**, **RQ2**); the second iteration in Chapter 4 (**RQ1**) and Chapters 5 (**RQ2**); the third iteration in Chapter 6 (**RQ3**).

[5] This was because the proposed speech intentions were derived by considering existing theoretical knowledge, but also by empirically analyzing the corpus. The tweets collected through the official Twitter API (`https://developer.twitter.com/en/docs`) represented public communication, which proved more limited with regard to the speech intentions it conveyed. For instance, people seldom made promises publicly; thus, commissive speech acts were very rare among the collected tweets.

1.5 Contributions and Publications

Multiple original contributions are made:

1. A speech intention taxonomy derived from linguistics—from Searle's speech act types [184] and Vanderveken' speech act verbs [212], to model asynchronous communication. Compared to all taxonomies from the related works, the proposed one is corpus-independent, comprehensive—as in both finer-grained and exhaustive in the given context, and its application by non-experts is proven feasible through extensive experiments.

   In the first form, the taxonomy consisting of 6 speech intentions is proposed to cover the Twitter public communication—finer-grained classes belonging to the assertive and directive speech act types are defined and, compared to the related works on modeling public tweets with speech acts, this representation is both corpus-independent and more detailed [51] (Chapter 3). Then, in a collaboration with a linguist [41], the previous taxonomy is extended to 18 classes of speech intentions in order to be applicable to any type of communication. Additionally, oppositional traits of speech intentions are added to the taxonomy to structure the classes and to enable an easier manual annotation (Chapter 4).

   The manual annotation experiments with both experts and non-experts show consistency in the perception of most speech intentions[6] in both iterations. Moreover, the rare use of the class *other* by annotators empirically supports the completeness of the taxonomy to formalize communication in two types of corpora: tweets [51] and forum conversations [41].

2. A corpus-independent, automatic method to annotate utterances of asynchronous communication with the proposed speech intentions taxonomy designed based on supervised machine learning [50, 51, 57]. For this, validated ground-truth corpora are created (one with tweets[7] and one with forum conversations[8]) and groups of features—discourse, content and conversation-related, are engineered to be used by the classifiers. In particular, some of the discourse features are novel and defined by considering linguistic means to express speech intentions, without relying on the explicit content or domain of the corpus or on specificities of the asynchronous communication types.

   The experiments on the Twitter corpus [51] (Chapter 3) show that the speech intentions can be satisfactorily discovered by using Linear SVM or Logistic Regression with discourse features only (F1-scores between 0.73 and 0.80). Compared to the related works, the proposed method annotates each tweet with multiple, finer-grained speech intentions by relying only on discourse characteristics, which are independent of the domain or corpora.

[6] The aligned perception may be interpreted as a measure of correctness too, when comparing the reasoning of the expert annotator with the reasoning of a non-expert annotator.

[7] `http://tinyurl.com/hk9t83y`

[8] `http://tinyurl.com/yc8mjtb2`

The experiments on the Reddit corpus [50] (Chapter 5) show that the same classifiers, Linear SVM and Logistic Regression, lead to the best results, in an one-versus-all setup, with both content and discourse features (F1-score between 0.53 and 1). Additionally, in order to improve the identification of speech intentions with lower F1-scores, two strategies are proposed and assessed: to use a multi-label classification for commonly co-occurring pairs of speech intentions or to augment the training corpus with instances of the least represented speech intentions even from external, heterogeneous corpora. Although content features appear now necessary for the accurate annotation of sentences, the detailed analysis shows that generic n-grams, independent of the corpus or domain and related to speech intentions are the most important in prediction. Additionally, the approach shows satisfactory robustness in discovering speech intentions from other corpora, including synchronous communication. Compared to the existing solutions that annotate forum conversations with supervised machine learning, this approach is general and the most comprehensive—uses detailed classes to multi-label sentences, and extensive experimentation is conducted to select the best setup for multi-label classification, to explore strategies to improve the classification performance and to assess the external validity of the approach.

The relevance and impact of the results obtained from both sets of experiments to medicine are discussed too, by showing various directions to integrate the approach in technologies and studies regarding health-information seeking and dissemination, persuasion and health compliance-gaining and narrative medicine.

3. An automatic method based on process mining [1] designed to generate process models of interrelated speech intentions from conversational turns, annotated with multiple labels per sentence [50, 58] (Chapter 6). As process mining relies on well-defined structured event logs, an algorithm to produce such event logs from conversations is proposed. Additionally, an extensive design rationale on how conversations annotated with multiple labels per sentence could be transformed in event logs and what is the impact of different decisions on the output behavioral models is released to support future research. Further, the Fuzzy Miner technique [85], implemented in the Disco tool, is selected to exploit the produced logs because it is suitable for unstructured processes and is highly interactive and visual.

   Compared to the existing works which discover relations among speech intentions from annotated conversations as transition diagrams, the process models mined with process mining are general and more comprehensive, capturing other types of relations too—significant length-2 loops, indirect sequences and concurrencies, and could be interactively explored at different levels of detail. Additionally, the only other work, among the few existing ones, to put process mining to test for conversation analysis [214] relies on a more simplified and domain-dependent scenario—each turn is annotated with a single speech act related to discussions in online courses. In contrast, the current solution proposes a

complementary and general scenario, motivating why the previous one is not applicable to conversations annotated with multiple labels per sentences.

The qualitative evaluation of the proposed approach starts with a discussion about the relevance of the obtained process models to medicine. Then, it presents the results of a preliminary observational study conducted with a linguistic researcher to assess the relevance to conversation analysis. The results show that the proposed approach leads to the discovery of insightful models of turn building in conversations and to the formulation of new hypotheses regarding speech acts and conversations.

4. The first systematic study to date on the automatic modeling of asynchronous conversations with speech acts is conducted (Chapter 2). For this, a framework to analyze and compare the related literature is developed, consisting of multiple facets regarding the speech intentions taxonomies, the relations among speech intentions, the application domains and potential beneficiaries, the manual corpus creation and validation, the automatic approaches for conversation modeling with speech intentions and with relations among speech intentions, including the diverse groups of features relevant to the class discovery.

This thesis contains the following peer-reviewed publications and manuscripts in preparation:

- Epure, E.V., Compagno, D., Salinesi, C., Deneckere, R., Bajec., M., Zitnik, S. (2018). Process Models of Interrelated Speech Intentions from Online Health-related Conversations. *Artificial Intelligence in Medicine*. doi = https://doi.org/10.1016/j.artmed.2018.06.00

- Compagno, D., Epure, E.V., Deneckere-Lebas, R., Salinesi, C. (2018). Exploring digital conversation corpora with process mining. *Corpus Pragmatics, 2*(2) 193-215.

- Epure E.V., Deneckere R., Salinesi C. (2017). Analyzing Perceived Intentions of Public Health-Related Communication on Twitter. In ten Teije A., Popow C., Holmes J., Sacchi L. (Eds), *Proceedings of 16th Conference on Artificial Intelligence in Medicine* (pp. 182-192). Vienna, Austria. Springer.

- Epure, E.V., Zitnik, S., Compagno, D., Deneckere, R., Salinesi, C. (2017). Automatic analysis of online conversations as processes, presented at *Journées Analyse de Données Textuelles en Conjonction avec EDA 2017*, Lyon, France.

- Epure, E.V., Deneckere, R., Salinesi, C. (2018): *Automatically modeling asynchronous conversations with speech intentions: A systematic study*. Manuscript in preparation.

- Epure, E.V., Salinesi, C., Deneckere, R., Bajec., M., Zitnik, S. (2018). *A linguistic approach to reveal tacit knowledge for requirements engineering from online discussions*. Manuscript in preparation.

For thematic coherence reasons, this thesis does not comprise the following articles that were also published—works carried out in parallel with the subject treated in this thesis:

- Epure E.V., Kille B., Ingvaldsen J.E., Deneckere R., Salinesi C., Albayrak S. (2017) Recommending Personalised News in Short User Sessions. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)* (pp. 121-129). Como, Italy. ACM.

- Epure E.V., Kille B., Ingvaldsen J.E., Deneckere R., Salinesi C., Albayrak S. (2017) Modeling the Dynamics of Online News Reading Interests. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)* (pp. 363-364). Bratislava, Slovakia. ACM.

- Epure E.V., Ingvaldsen J.E., Deneckere R., Salinesi C. (2016). Process mining for recommender strategies support in news media. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-12). Grenoble, France. IEEE.

- Epure, E. V., Deneckere, R., Salinesi, C., Kille, B., Ingvaldsen, J.E. (2017). Devising News Recommendation Strategies with Process Mining Support. Presented at *Interdisciplinary Workshop on Recommender Systems*. Paris, France.

- Epure E.V., Martin-Rodilla P., Hug C., Deneckere R., Salinesi C. (2015). Automatic process model discovery from textual methodologies. In *2014 IEEE Ninth International Conference on Research Challenges in Information Science (RCIS)* (pp. 19-30). Athens, Greece. IEEE.

- Gonzalez, C., Martin-Rodilla, P., Epure, E.V. (2016). Formalization and Reuse of Methodological Knowledge on Archaeology across European Organizations. Presented at *the 44th Computer Applications and Quantitative Methods in Archaeology Conference*. Oslo, Norway.

The interested reader may consult a representative selection of these publications in Appendices G and H, which focuses on the automatic discovery of behavior from data as process models, but in different contexts: from web logs to create news recommendation algorithms [54, 56] and from text reporting behavior to support domain knowledge sharing in humanities [57].

Finally, a simplified illustration of the created solution is presented in Appendix B.

1.6 Overview

Chapter 2 presents the results of the two research questions regarding the existing literature (**RQ_L1** and **RQ_L2**). By conducting a systematic literature study, the related works that automatically model asynchronous conversations with speech acts are reviewed. Additionally, a discussion is taken up on the extent to which these works satisfy the solution goals and properties defined in the problem investigation phase.

Chapter 3 presents the next contribution of this thesis. Specifically, a domain and corpus-independent method to automatically model tweets with comprehensive speech intentions is designed, addressing thus the first and the second research questions (**RQ1** and **RQ2**). Experiments are conducted to prove the internal validity and the external validity is discussed as well.

Chapter 4 addresses the limitations of the previous solution presented in Chapter 3 with respect to the comprehensiveness of the proposed speech intentions (**RQ1**). An extended taxonomy of 18 speech intentions that complies with all the criteria established in the problem investigation phase is proposed and validated.

Chapter 5 is closely linked to the previous one, forming the second iteration of the design cycle. Supervised machine learning classifiers and various classification setups and sets of features are explored for the automatic annotation of sentences from forum conversations with the proposed speech intentions taxonomy. Thus, the second research question is revisited (**RQ2**). The internal and external validities are determined through experiments.

Chapter 6 reports the third iteration of the design cycle, which completes the design of a satisfactory and complete solution (**RQ3**). Process mining is chosen to reveal processes of interrelated speech intentions from annotated asynchronous conversations. Precisely, a method to transform the annotated conversations in well-defined logs, required as input by process mining techniques, is proposed. The relevance of the overall approach is discussed for medicine and validated for linguistics / conversation analysis in an observational study.

Chapter 7 presents an overview of the contributions, as answers to the proposed research questions, and identifies limitations and final directions for future work.

CHAPTER 2

LITERATURE REVIEW

Epure, E.V., Deneckere, R., Salinesi, C. (2018): *Automatically modeling asynchronous conversations with speech intentions: A systematic study*. Manuscript in preparation.

Contributions: E.E.V. designed and conducted the research and wrote the manuscript. S.C and D.R. provided feedback on the research design and on the manuscript.

This chapter presents a systematic literature study that:

- discusses how the related works model asynchronous conversations with speech intentions.
- compares the proposed solutions from several perspectives: types of asynchronous communication, envisioned applications and beneficiaries, speech intention taxonomies and relations among speech intentions, manual and automatic annotation methods.

In the late '90s-early 2000s, the focus in the related works was put on modeling synchronous communication, in particular spoken, such as phone conversations [110, 196] and recorded multi-party meetings [140, 192]. Subsequently, with the increased adoption of chat rooms and instant messaging, researchers began to focus on modeling text-based synchronous conversations too [103, 112, 228]. The study of speech intentions in the context of asynchronous communication arose much later. Although the first work in this direction was published by Cohen et al. [40] in 2004 using emails as study corpus (see Table 2.6 in Section 2.3), it was not until 2010 that asynchronous conversations received much greater attention than the text-based synchronous type. The increased popularity of asynchronous communication modeling over the past 10 years is most likely associated with the increased use of the Internet, new online media such as social media and portable devices.

As Petersen et al. [159] emphasize, once the number of publications in a research area reflects a rapid increase, an analysis of existing works becomes important for a better understanding of that area and for identifying research gaps. Although there are several detailed literature studies on the automatic modeling of synchronous communication [123, 217], there is none centered on the automatic modeling of the asynchronous type to date.

Therefore, the main purpose of this chapter is to systematically review the existing literature that reports empirical solutions to automatically model asynchronous communication with speech intentions. Multiple sub-objectives are identified:

1. Variables and concepts relevant to this area of research that could be used as dimensions for comparing the related literature are identified.
2. Methods and research techniques are described to gain methodological knowledge.
3. Different perspectives on the studied research area are presented.
4. By analyzing what has been done, new lines of inquiries and improvements are suggested. Also, attempts that have led to unsatisfactory results can be discussed.

2.1 Research Method

The research method followed for conducting this systematic literature review is summarized in Table 2.1 [42, 164]. Each step is briefly defined in Appendix C and the implementation of each step in the current work is detailed in the following subsections.

Table 2.1: The research method followed for conducting the systematic literature review [42, 164].

Step 1	**Step 2**	**Step 3**	**Step 4**	**Step 5**
Problem formulation	Data collection	Data evaluation	Analysis & interpretation	Public presentation

2.1.1 Research Questions, Foci and Goals

Two research questions guide the current literature review[1]:

RQ_L1 *How do the related works automatically model asynchronous conversations with speech intentions?*

RQ_L2 *How well do the related works meet the goals and properties of the solution, identified in the problem investigation phase?*

As mentioned earlier in the introduction to the current chapter, there is no systematic literature study centered on the automatic modeling of asynchronous conversations to date. The closest ones identified focus only on synchronous communication [123, 217].

[1]These questions are coded with the prefix **RQ_L** to indicate that they are in addition to the three main research questions and are addressed only in this chapter through the study of literature.

The second question must be interpreted in the context of the present thesis. As specified in Chapter 1, the solution's goal is to model asynchronous conversations as processes of interrelated speech intentions. Several properties of the solution have been highlighted. The solution should be automatic, domain-independent and effective by leading to comprehensive and relevant knowledge. Automatically discovered knowledge should be accurate and complete. Thus, effectiveness is also related to the algorithmic performance of the proposed technology. The foci of the current work are research methods, theories and, to some extent, outcomes and practices and applications.

The first research question explores two aspects of the existing works: 1) the method and 2) the speech acts and, if any, their relations. Therefore, methodological and theory reviews are needed to answer this research question. In practice, through methodological review, the key components, variables, analysis methods and rationales that inform the adopted designs are identified in the first instance. Then, these findings are interpreted to summarize the advantages and disadvantages of different methodological approaches and to compare them for different types of asynchronous communication or historically. The theory review is necessary to establish the existing conceptualizations of speech intentions and the relations among them.

Methodological and theory reviews are also necessary for the second research question. The interpretation of the adopted speech acts can reveal if they are corpus-independent and comprehensive. The analysis of research methods is also necessary to validate whether the proposed solution is general. For instance, domain or corpus-specific characteristics could be integrated into the designed solutions. Then, since the second research question also investigates the relevance of the work, the applications envisioned by the authors should be illustrated or at least be discussed. Hence, a more shallow review of the practices and applications is conducted. Finally, reviewing the outcomes of the existing works is necessary to discuss the algorithmic effectiveness of the proposed solutions. However, a comparison of the outcomes is not possible because the output and types of asynchronous communication differ significantly in the studies.

The current goals refer to all aspects outlined by Cooper [42]: integrating and criticizing existing works and identifying central issues. The integration of review data illustrates the field of research in a concise view by modeling its map. Integration is applied to review research methods, theories and applications. Criticism is needed to identify research gaps and limitations of existing works. Critical analysis and identification of central issues can motivate future investigation. All foci of the current work are concerned with these later goals.

2.1.2 Data Collection

The digital libraries used for data collection are: ACM, IEEE Xplore, SpringerLink, Science Direct, Microsoft Research, AAAI, ACL Anthology, JSTOR, ArXiv and DBLP. This list was compiled by selecting from Wikipedia's list of academic databases and search engines[2] those related to

[2] https://en.wikipedia.org/wiki/List_of_academic_databases_and_search_engines

computer science. In addition, the top ten pages of results returned by Google Scholar and Semantic Scholar search engines were also analyzed.

The use of the term "intention" for searches proved to be very broad, as intentions have been studied in other contexts too, besides communication. For example, identifying behavioral intentions is relevant to recommendation algorithms [190] or modeling processes in information systems [53]. Querying by "speech intention" or "intention in speech" did not improve the outcomes either because intentions are also extensively studied as types of engagements in future behavior. Discovering people's commitments is an objective of many communities, including marketing and e-commerce [30, 97].

As linguistic theories refer to speech intentions as speech acts, an assumption was made. Any existing solid work that addresses the subject of speech intentions should be familiar with the theoretical background. Hence, the concept of *speech act* must be mentioned in the text. Consequently, the first term selected for the search query was "speech act". Initial search by "speech act" has shown that often speech acts have sometimes been used interchangeably with dialogue acts. Dialogue acts are defined as specialized speech acts to interpret utterances in dialogues, especially in the synchronous and spoken ones [217]. Although asynchronous communication is targeted, sometimes dialogue act taxonomies are also used in this context [149]. Hence, "dialogue act" was also added to the search query, as an alternative term to "speech act".

Further, the goal was to select only empirical papers that propose automatic solutions to annotate conversations with speech intentions. For this reason, "automatic discovery" and "automatic classification" and "automatic annotation" were included in the search query. Moreover, because these solutions are often based on supervised or unsupervised approaches, "machine learning" was added as a substitute. To further cover cases where the intention discovery was automatic but based on other approaches such as rule or knowledge-centered, the terms "text mining" and "natural language processing" were added. A second assumption was made here. No matter what modeling approach was proposed, text processing was a mandatory step. Thus, through these two terms, alternative solutions to machine learning would appear in search results. The final search query is further summarized:

("speech act" OR "dialogue act") AND ("automatic classification" OR "automatic discovery" OR "automatic annotation" OR "machine learning" OR "text mining" OR "natural language processing").

The search query was sometimes subdivided into multiple sub-queries when the results returned for some digital libraries were very scarce and a query processing issue was suspected. The complete list of queries and the number of articles returned by each is presented in Appendix D.

The adopted search strategy is exhaustive with selective citation, which means that the search of the related publications is thorough but constrained by certain conditions. These conditions can be considered as initial inclusion or exclusion criteria, such as the publication to be in English, and are detailed in the next section. However, a specific condition is presented

here as it played a role in the search process, being used as a filter when the interface allowed it. Only the publications between 2000 and 2017 were taken into account. The rationale of this decision is explained in the introduction of this chapter. As mentioned there, the solutions that model asynchronous communication have been published since 2004 (see Table 2.6 in Section 2.3). Nonetheless, the four years preceding this date and the references of the first publication on the topic [40] were also investigated for caution. One last remark about the data collection is that the relevant references of the retrieved publications were also checked.

2.1.3 Inclusion/Exclusion Criteria and Other Aspects of Data Evaluation

Only works reported in English were considered. However, this condition did not constrain the text on which speech acts have been researched. Thus, publications reporting automatic discovery of speech intentions from conversations in other languages were included.

A large body of publications returned by the defined search query was within the artificial intelligence domain, specifically to formalize the agent-based communication and language [117, 219]. These works were excluded because they are not empirical studies, but theoretical proposals. Theoretical models of agent behavior in general or in speech are defined in a top-down manner and after implemented.

Further, many retrieved publications reported the exploitation of recorded real-life conversations, most frequently analyzed with dialogue acts. The nature of communication allows for the use of a larger spectrum of features, apart from the text-based ones, in modeling conversations with speech intentions. These features include speech characteristics such as prosody [8] and even physical indicators collected with external devices, for instance, for projecting the person's emotions [20]. These automatic solutions were excluded because they did not focus on asynchronous communication and they also employed text-unrelated features.

Then, another category of published works was observed, which relied on data already tagged with speech acts. Their main goal was not the modeling of communication, but the use of already annotated conversations for varied applications. For instance, speech acts were used to discover major life events in social media [125], to reveal effective teaching strategies [208], to select unresolved threads in forums [61], to model communications skills and personality traits [147], and to generate relevant dialogues utterances with neural networks [236]. While these articles are out of scope for the current literature review, a selection of them was kept in order to provide recommendations on use cases and application domains that could benefit from modeling the communication with speech intentions.

Quality criteria were defined too. Works published only in ranked journals and conferences were included. For the conferences, CORE or Qualis databases were checked and a rank of at least B, respectively B2, was the limit for inclusion. For the journals, the impact factor had to be at least 1.5. Any article published in conferences, journals, workshops that appeared as influential—cited at least 10 times—was included. Book chapters were included only if published

Table 2.2: Inclusion and exclusion criteria for assessing relevance and quality.

Relevance Criteria	**Quality Criteria**
R1: Publication year must be at least 2000.	**Q1**: Include only published works.
R2: Publication must be reported in English.	**Q2**: If a conference article, the conference is ranked with at least B in CORE or B2 in Qualis.
R3: The modeling of communication with speech intentions must be empirical and bottom-up—exclude papers formalizing agent languages.	**Q3**: If a journal article, the journal impact factor is at least 1.5.
R4: The exploited corpus must be composed of asynchronous conversations—exclude works using recorded conversations or synchronous chats.	**Q4**: Include influential articles—cited at least 10 times. Any venue (journal, conference, workshop) is accepted and the rank is not relevant.
R5: Exclude papers using already modeled conversations with speech acts for deriving knowledge or designing new intelligent technologies.	**Q5**: Book chapters are included if published by well-known academic publishers.

by well-known academic publishers such as Springer.

The relevance and quality criteria are summarized in Table 2.2. The process to apply the defined criteria is further described. First, the title, abstract and keywords of the publication were analyzed and a scanning of the article content took place in order to gain a preliminary understanding. If no clear decision regarding its inclusion or exclusion could be made, a more comprehensive understanding of the articles was necessary. Thus, the introduction and conclusion were read and the section describing the corpus was located in order to grasp if the tackled communication was asynchronous. Based on this step, the final decision was made.

2.1.4 Analysis and Interpretation

A conceptual framework to extract data from the relevant publications and to enable an in-depth analysis of methodologies, theories on speech intentions, outcomes and applications is presented in Section 2.2. The proposed conceptual framework is a *classification scheme* containing multiple facets. Initially, this scheme had much fewer facets, regarding types of asynchronous communication and used automatic approaches (e.g. supervised, unsupervised). However, through in-depth analysis of the relevant literature, the conceptual framework evolved iteratively.

As for the means of analysis, narrative summaries are mainly used in the current work. The discussions are presented by themes mapped on types of asynchronous communication. Critical comparison of the findings are reported through narrative explanations associated with graphical tables and models. The result analysis and interpretation are exposed in Section 2.3.

2.2 A Conceptual Framework for Analysis and Interpretation

The classification scheme is summarized in Table 2.3. It contains 13 facets, each having associated multiple categories—apart from the last one.

The focus represented in the first column associates a facet to a specific part of the proposed solution: its input, output, method, envisioned applications and the outcomes of the experimental evaluation. Their brief introduction is presented below:

- The *input* facet is relevant to conduct a thematic analysis by considering specific types of asynchronous communication: emails, tweets, forums and social media posts.
- The *output* facets support a theoretical review with regard to the conceptualizations of speech intentions and their relations, if present.
- The *practice* facet is used in the applications-related review.
- The *outcome* facet is useful to analyze, compare and integrate the results of the proposed automatic, speech intention-centered approaches to model asynchronous communication.
- The remaining facets, all focusing on the *method*, help extracting and analyzing information in order to conduct the methodological review.

To answer the first research question (**RQ_L1**) on how the related works automatically model asynchronous conversation with speech intentions and processes of interrelated speech intentions, the output and method facets need to be considered. The second research question (**RQ_L2**) assesses whether the related works comply with the goals and properties of the expected solution for satisfying the thesis main objective. The goals are by default satisfied as only publications reporting automatic modeling of asynchronous communication with speech intentions are included in the review. However, to discuss compliance with the expected properties, multiple facets need to be examined: the practice facet to analyze the relevance, the output and feature group facets to analyze the corpus- and domain-independence and the outcome facet to discuss the algorithmic effectiveness. In the next subsections, a detailed description of each facet is provided.

2.2.1 Input-centered Facets

Three types of asynchronous communication are identified from the selected literature: emails, forum and social media and tweets.

The large adoption of email as a means of communication in everyday life has resulted in accepted conventions for email production and interpretation, recognized by Mildinhall and Noyes [138] as "stereotypical templates". Conversational in nature, an email is composed of sequences of utterances and the email exchange could develop in threads among multiple participants [79].

Emails are often similar to forum and some social media posts. Additionally, email, forum and social media conversations could reassemble in form [10, 14]. However, their terminology

Table 2.3: Classification scheme with multiple facets and their associated categories.

Focus	**Facet**	**Categories**
Input	Asynchronous communication	email (1), Twitter (2), forums and other social media (3)
Output	Speech intention taxonomy	general (1), domain or corpus-specific (2), coarse-grained (3), fine-grained (4)
Output	Relations among intentions	present and automatically discovered (1), present as context features in solution (2), absent (3)
Practice	Applications & beneficiaries	discussed (1), not discussed (2)
Method	Annotation unit	turn (1), sentence (2), part of sentence (3)
Method	Annotation strategy	single-label (1), multi-label (2)
Method	Ground-truth corpus	present and validated (1), present and not validated (2), absent (3)
Method	Text preprocessing	stop-words exclusion (1), stop-words inclusion (2), lemmatization (3), stemming (4), placeholders (5), segmentation (6), none specified (7)
Method	Sampling (unbalanced corpus)	present (1), absent (2)
Method	Automatic modeling approach	supervised (1), unsupervised (2), semi-supervised (3)
Method	Feature groups	bag-of-words (1), n-grams (2), text similarity (3), time, date, numbers, urls (4), morphological & syntactic (5), contextual (6), verb-related (7), pronoun-related (8), sentiment-related (9), taxonomy-related expressions (10), punctuation (11), unit length (12), position in conversation (13), author identity (14), communication type-specific (15), domain-specific lexicons (16), general-language lexicons (17), other features (18)
Outcome	Modeling results	not applicable—numerical values

is different. In forums and social media, the initial turn is usually called the starting *post* and any other post replying to an already existing one is called *comment* [10]. Moreover, the communicative goals could be rather different between emails, forums and social media.

Compared to these types of asynchronous communication, tweets have much more different characteristics. A maximum of 140 characters allowed per tweet has led to a very condensed style of writing with frequent linguistic noise and non-grammatical sentences [234].

2.2.2 Output-centered Facets

The output-centered facets have two goals:

1. to tell if the speech intention taxonomy used in the specific research work is corpus-independent and comprehensive;

2. to identify if relations among speech intentions are present in the specific research work and how they are represented.

A speech act taxonomy is considered *corpus-independent* if it can be applied to any type of corpus. For instance, the types proposed by Searle [184] are corpus-independent. However, when composed of speech acts customized to a certain domain, corpus or application, a speech act taxonomy is considered *corpus-dependent*.

Then, in order to adjudicate the comprehensiveness of a speech act taxonomy, the classes associated to the speech act theory—assertive, expressive, directive, commissive, declarative—are taken as reference and they all should be present. However, if only these wide speech act classes are used—or a subset of them, the speech act taxonomy is considered *coarse-grained*. Otherwise, if at least one speech act is mapped on multiple intentions, the taxonomy becomes *fine-grained*.

Theoretical and empirical research, in particular from linguistics [181, 183, 187], has also put forward the *process* perspective on conversations. The discourse could be studied, not only through standalone speech intentions, but also by investigating the way these intentions interrelate (e.g. adjacency pairs) for building conversations. During the iterative definition of this conceptual framework, it was noticed that some related computer science works also adopt this perspective, being central to the design of their proposed solution. Two approaches were identified: either previous speech acts are used as features in the method to predict the next classes, or relations among speech acts are automatically discovered from corpora—usually as common sequences; hence the categories of the "Relations among intentions" facet were formulated.

2.2.3 Practice-centered Facets

This facet was defined to generally summarize if the studied literature motivates and discusses applications of automatic modeling of asynchronous communication with speech intentions or identifies potential beneficiaries. This is necessary in order to prove the relevance of the proposed

contribution. Although individual thematics were identified during the study of the selected publications, they were limited to certain types of asynchronous communication—for example, email-specific applications. Consequently, specific facet categories referring to these applications and groups of beneficiaries were not introduced. Nonetheless, these thematics are discussed using narrative summaries.

2.2.4 Method-centered Facets

In general, the methods to automatically model asynchronous communication with speech intentions are composed of several steps:

1. An existing speech intention taxonomy is adopted or a new one is defined.
2. Data of the targeted asynchronous communication is collected and a ground-truth corpus is created by manually annotating this data, or a sub-selection of it, with speech intentions[3].
3. An automatic modeling technique is built by first deciding the approach type: supervised, unsupervised, semi-supervised. Also, sampling methods can be explored to mitigate the effects of corpus unbalance (unequally represented classes) on algorithmic performance .
4. Relevant features of speech act classes are defined and further used in algorithms. Feature extraction may be preceded by text preprocessing, aimed at normalizing the input.
5. Algorithms are trained and evaluated and parameter tuning also takes place.

The choice of annotation unit has an impact on both manual and automatic annotations. When trying to model discourse with speech acts, conversational *turns*, turn *sentences* or *parts of sentences* can be individually analyzed. If parts of sentence or sentences are the chosen units in manual annotation, this does not imply that automatic algorithms are created to identify speech intentions for these units. Instead, automatic algorithms may be trained per turn. So, the labels associated to each turn sentence or parts of sentence are aggregated to represent the labels of that specific turn. This is possible due to the inclusion relationship that characterizes the three types of annotation units.

Then, whatever the unit of annotation, a second decision for either manual or automatic annotation is if multiple intentions are allowed per unit or just one. The former case is called a *multi-label* annotation, the latter case a *single-label* annotation. A manually single-labeled corpus could be treated as a multi-labeled one in the automatic approach if the units of annotation in the manual annotation are parts of sentence or sentences and are aggregated into sentences or turns for building the technique. A multi-labeled corpus could be treated as a single-label one in

[3]These first two steps could be skipped if an already annotated dataset is used. Moreover, in unsupervised learning approaches, the creation of the ground-truth corpus is not necessary; a speech intentions taxonomy could be used to enable the naming of the clusters, but ad-hoc classes of speech acts could be proposed too, in a bottom-up manner, by manually analyzing the clusters.

the automatic approach when algorithms are created for each specific speech intention. In this case, for each target intention, a new corpus is formed by specifying for each unit from the initial multi-labeled corpus if it has associated the target speech intention (positive instance) or not (negative instance). Even if this case is implemented, the overall approach may still automatically annotate each unit with multiple speech intentions, by showing as final result the set of classes for which their corresponding algorithms yielded true. In machine learning, inherent multi-label prediction is also possible. Then, the same algorithm can identify all possible intentions per unit, compared to having a classifier for each speech intention type, as in the previous case.

In conclusion, the "Annotation unit" and "Annotation strategy" facets should be discussed separately for manual and automatic annotations, although very often they match.

In the beginning of this section, it was mentioned that a ground-truth corpus may be needed or not, depending on the automatic modeling approach. The preparation of a ground-truth corpus requires humans annotating units. A natural question that arises regarding this task is how to ensure the reliability of the annotation and thus of the corpus. Compared to other types of corpora, for instance for entity recognition, annotating a corpus with speech intentions highly relies on human subjective judgments. While the annotation could follow some instructions or, at least, refer to some clear class definitions, how is it indeed ensured that the obtained corpus is reliable and also what it is claimed to represent is valid across multiple human perceptions?

The most common answer to this is to manually label the same corpus by multiple annotators and to use Kappa statistic [27] (discussed in subsection 2.2.5) to measure the reliability or agreement. In this way, a ground-truth corpus is created and validated. In the selected publications, it was observed that ground-truth corpora were sometimes created, but their validation was not discussed or reported. Hence, three categories were defined for the "Ground-truth corpus" facet to cover all possible cases.

Then, when the speech intentions chosen by two or more annotators for the same unit are compared, one could observe an agreement or a disagreement. By default, the ground-truth corpus contains all the units for which agreements are observed. However, how should disagreements be treated? In some studied works, three strategies were mentioned, while in others, nothing related to this aspect was discussed. The three strategies are:

1. to exclude the units with annotating disagreements from the ground-truth corpus;
2. to have the expert annotators—usually the originators of the speech intention classes—discuss the disagreements and make a decision;
3. for the case when at least three human annotators label each unit, a ground-truth corpus can be formed with all the units for which a majority decision among annotators is reached (e.g. 3 out of 5 annotators choose a certain class).

The "Text preprocessing" facet is defined to capture varied steps followed to transform each unit of the dataset in a canonical form before being used by the technique:

- *Segmentation* is used to identify individual units from larger portions of text. For instance, turns can be segmented into grammatical sentences that represent individual utterances or sentences could be segmented in parts of sentences, such that each individual part conveys a single speech intention. *Segmentation* can be also used to identify turns in conversations, when posts are not delimited by default, such as in Wikipedia Talk pages [64].

- Lemmatization and stemming are linguistic procedures aimed at transforming words into a common base form. In language, words would appear in different inflectional forms (e.g. "see, sees, seeing") or derivational forms (e.g. "large", "enlarge"). *Stemming* derives the common base form by chopping the end of words under the assumption that it would be sufficient to obtain the root of the word. In practice, Porter's stemming [134], the most famous stemming algorithm, highly effective in many circumstances, consists of more complex rules of word chopping grouped in five sequential phases. *Lemmatization* exploits complex knowledge of the language by using a vocabulary and by analyzing the morphological form of the word beforehand. Then, the common form of the word, called lemma, is discovered.

- *Placeholders* are keywords used to replace specific text such as urls, nouns or code snippets. In order to simplify the input and allow the algorithm to exploit the most interesting input knowledge, placeholders are chosen to encode and normalize otherwise too diverse or domain-specific information that is not explicative in terms of speech intentions.

- *Stop-words* are the most common words in a language such as prepositions or auxiliary verbs. They are also known as function words and they have no or little lexical meaning, compared to their opposite group—content words.

In a dataset, the class distribution might be non-uniform and while some classes are highly represented, some others may rarely appear. This aspect could negatively impact the performance of automatic classifiers. Therefore, sampling procedures can be used to adjust the class distribution. The most common sampling strategies are undersampling—when units of the majority classes are removed, and oversampling—when units of the minority classes are duplicated. More complex sampling technique are also available such as SMOTE [37]. The "Sampling" facet informs if sampling techniques are used or not in the related works and, when present, the sampling techniques can be discussed.

Three types of machine learning solutions were observed in the related works studied:

1. In *supervised* learning, the technique learns the most effective function to map the input—conversational units, on the output—speech intentions, from a given set of examples—the ground-truth corpus. Common supervised learning algorithms are Naive Bayes, Decision Trees, Logistic Regression, Support Vector Machine (SVM) and Neural Networks [17].

2. *Unsupervised* learning aims at learning a function capable of unveiling a hidden structure in the input data. This time, the input data is unlabeled. Sometimes, labeled data can

be used with unsupervised learning. However, the fact that the data is labeled does not influence the learning of the function, but it enables a more thorough evaluation in terms of prediction accuracy. Examples of unsupervised learning approaches are clustering (e.g. K-means), neural networks (e.g. Deep Neural Networks) and statistical models (e.g. Hidden Markov Models) [17].

3. *Semi-supervised* learning is a type of supervised learning that learns the mapping function by exploiting both labeled and unlabeled data. Usually, the labeled dataset is much smaller compared to the unlabeled one. One advantage of this approach is that it requires significantly less human effort for manually preparing a ground-truth corpus.

Whatever the selected learning strategy, features extracted from the input are necessary. These features capture properties of the input that could describe the phenomenon aimed to be observed, in this case, speech intentions from text. Several clusters of features were iteratively defined, by analyzing the related works. Each cluster is further described:

1. *Bag-of-words* (shortened BoW): is one of the most frequent and simplest representations of a textual input. The text is transformed in a dictionary of unique words and a value measuring the presence of those words in the text. This measure could simply mark if a specific word appears in text or not or how many times a word appears in the text (counts or frequencies with regard to the total number of words). Sometimes, a more complex measure—term-frequency, inverse document frequency (TF-IDF)—is also used in order to account for the bias that frequent words with minimum information content may cause. In order to avoid sparse representations that could lead to higher computational complexity, only the most frequent words may be selected to be part of the BoW dictionary.

2. *N-grams*: improves the BoW representation by taking in consideration the order and context of words, apart from their measure of presence in text. N-grams are contiguous sequences of words: unigram—for single words, bigram—for two consecutive words, trigram—for three consecutive words and so on. For measuring the presence of a n-gram in text and for handling sparsity, the previously discussed strategies for BoW apply here too.

3. *Text similarity* is used to compare the similarity of two texts and quantitatively represent the degree of resemblance. The similarity could be measured character-wise or word-wise. A commonly employed measure is the cosine similarity, when each text is represented as a vector (BoW, N-grams) and the cosine of the angle between these vectors is computed.

4. *Time, data, numbers, urls* features: are used either to mark the presence of times, dates, numbers, links in text or to report the time between two turns in conversation.

5. *Morphological and syntactic* features: incorporate information related to the morphology and syntax of the text. Specifically, there are features related to part-of-speech (POS) and

dependency trees. POS tags could reveal if a word is a noun, verb, pronoun and so on. Dependency trees could show syntactic relations between the words of a sentence, such as that two nouns are united through the preposition "of" or a noun is the agent of a verb.

6. *Contextual* features: exploit the context of an unit by informing what are the speech intentions for the previous and/or following units. Contextual features could be related to the true unit labels, as specified in the ground-truth corpus. However, a more realistic approach, in particular when dealing with new unlabeled data, is to use estimated speech intentions for the neighboring units.

7. *Verb-related* features: represent various cues of speech intentions focused on verbs. A first category here is the use of existing corpora of speech act verbs, manually defined in theoretical linguistic works [79, 163]. Speech act verbs are verbs used to achieve various speech intentions and are explicitly outlined in discourse (e.g. the verb "advise" in "I advise you to check again.") Then, features regarding the presence of verbs in general, of modal verbs, of tenses, of moods and the positions of first verbs are also highly relevant.

8. *Pronoun-related*: similar to verbs, pronouns appear as relevant cues for speech intentions. The presence of certain pronouns, in particular the first and second persons, and their association with certain verb tenses and moods are used as speech act predictors.

9. *Sentiment-related*: are features that show the presence of sentiments in utterances and turns, highly relevant for identifying expressive speech acts. Sentiment analysis algorithms can compute sentiment scores, revealing if negative or positive opinions are expressed. A wider range of sentiments can be computationally discovered too [10], or implicitly through the identification of emoticons.

10. *Taxonomy-related expressions*: are manually collected expressions and words observed to be frequently used to convey the speech intentions defined in the taxonomy. These features are similar to verb corpus features mentioned in the verb-related group. However, taxonomy-related expressions appear to be identified by relying mainly on empirical corpus analyses and less on theoretical linguistic works.

11. *Punctuation*: contains features about the presence of interrogative and exclamation marks and of other symbols. Their position in the start of an utterance may be also relevant.

12. *Unit length*: is usually computed as counts or frequencies of words or characters. The lengths of the previous or next units are sometimes considered too.

13. *Position in conversation*: refers either to the position of a turn in a dialogue or the position of a sentence in a turn. The position of the previous turn is also included sometimes.

14. *Author identity*: can be described through multiple characteristics—the role of the author (e.g. student versus instructor [10, 118]), the authority score and post counts (frequent in question and answer forums), thread-related characteristics (the authors is the initiator of the thread), position of the author and of the previous author in thread and the same author has two consecutive posts. Speech intention classes can also be used to represent writer identity either by creating his/her profile as a distribution of classes or by specifying the previous speech intention that the writer conveyed.

15. *Communication type-specific*: are features capturing characteristics of each type of asynchronous communication, such as the presence of attachments in emails, of post vote scores in forums or of special characters in Twitter (e.g. "#").

16. *Domain-specific lexicons*: contain terms belonging to specific domains (e.g. medicine).

17. *General-language lexicons*: contain words and expressions used with a certain role in language, such as interjections, vulgar words or online-specific abbreviations.

18. *Other features*: groups features that cannot be included in any other feature category.

2.2.5 Outcome-centered Facets

The "Modeling results" facet is designed to enable the analysis, comparison and integration of the results obtained in manual and automatic annotations of asynchronous communication with speech intentions. The results are numerical and can represent different types of measures.

A basic approach to quantitatively evaluate the manual annotation of corpora includes two types of annotators, expert and non-expert, and counting the number of times two annotators, one of each type, agree. If the type of annotators is not relevant, the problem can be formalized as having N annotators that need to classify units using M exclusive labels. A more comprehensive way to model the agreement is then *Kappa statistic*, which originates from the domain of content analysis and corrects for the expected chance agreement [27]. Kappa values vary between 0—for no agreement, and 1—for full agreement. A scale of agreement is also proposed in order to unify the interpretation across different works (see Table 2.4) [27].

In the Kappa test, there is no distinction between different types of annotators, experts and non-experts being equally treated[4]. However, if expert opinion is important, Carletta [27] proposes to compute the Kappa scores by considering pairwise results between experts and non-experts, instead across all annotators.

The measures used for the assessment of the automatic annotation of corpora depend on the type of machine learning algorithms used: supervised versus unsupervised. Often, in the

[4]Carletta [27] argues that this is a correct way of approaching the problem as, in reality, most of the judgments in discourse analysis are subjective. Also, what is important is to see if the coding instructions convey what the originators aimed to achieve, by encouraging their use by non-experts too.

evaluation of supervised machine learning techniques, the reported measures are accuracy, precision, recall and F1-score [17]. The accuracy shows how many times a correct classification or prediction is made. The precision shows how many instances classified as belonging to a class belong indeed to that class. The recall shows how many from the instances belonging to a class were correctly discovered by the algorithm. The F1-score is the harmonic mean between precision and recall. The confusion matrix can be also used to visualize the performance of a technique; it is a table with a particular layout as in Table 2.5. The accuracy, precision and recall can be then derived from the values of the confusion matrix. For instance, precision is the ratio between the count of true positives and the total count of instances predicted as positive.

Table 2.4: Interpretation of Kappa scores—at least moderate results are preferred.

Judgment	**Score value**
Poor	0.00
Slight	0.01-0.20
Fair	0.21-0.40
Moderate	0.41-0.60
Substantial	0.61-0.80
Almost perfect	0.81-0.99

Table 2.5: Confusion matrix example for predicting a certain class.

	Predicted: positive	**Predicted: negative**
Actual: positive	true positive count (TP)	false negative count (FN)
Actual: negative	false positive count (FP)	true negative count (TN)

In unsupervised learning, if a ground-truth corpus is available, the same measures as in supervised learning could be used. Otherwise, other options exist too, depending on the context. For the current context, generating permutations from conversation turns and predicting the permutation closest to the true ordering of turns or qualitatively visualizing clusters have been proposed as alternatives for scenarios with unlabeled dialogues [154, 171].

2.3 Results

The created conceptual framework for analysis and interpretation was applied to the selected body of research on modeling asynchronous communication with speech intentions. The results are further presented and thematically discussed by considering each type of asynchronous communication. The integration of results is reported in graphical tables and figures and through narrative summaries; also, when considered appropriate, critical comparisons are made.

2.3.1 Asynchronous Communication Corpora

The datasets used in the selected literature are projected per type of asynchronous communication and per year in Table 2.6.

Table 2.6: Overview of the number of works per type of asynchronous communication and year.

year	**'04**	**'05**	**'06**	**'07**	**'08**	**'09**	**'10**	**'11**	**'12**	**'13**	**'14**	**'15**	**'16**	**'17**	%
emails	1	1	1		1	2				2			1		33%
posts				1		1	2	2	3	2		1	1	1	50%
tweets							1	1	3		1		2	1	27%

Forum and social media posts are most common, followed by emails and tweets[5]. In the beginning, emails were mostly studied; then, starting with 2010, the focus shifted on tweets and forum conversations. Further, a discussion with regard to the dimensions of communication (communicative goals, number of participants) is conducted by taking in consideration each asynchronous communication type.

Email. The *email* conversations used in the related work are either task-oriented (negotiation tasks in [28, 29, 40] and exchanges between work collaborators in [178]), personal [79] or a mix between these types—multiple publications [100, 106, 138] report using emails extracted from the Enron corpus[6], which can be personal, social or work-related.

Twitter. *Tweet* corpora can target a specific context, such as customer service in [150], rhetorical question in [165] or politics in [91]. The other related works on tweets create more general corpora by covering conversations on varied topics [171, 213, 234, 235]. Also, the topics are chosen to take into account diverse entities (e.g. artists), pieces of news and long-standing subjects (e.g. traveling, cooking).

Online Forum. The most frequently found *forum* corpora are extracted from online course forums and contain discussions among students and instructors [10, 13, 118, 166, 189]. Apart from these, there are also corpora with conversations from question and answer forums in the veterinary medicine [163], travel and restaurants [106] and computers [14, 194] domains. Ferschke et al. [64] release a corpus extracted from Wikipedia Talk pages[7].

Regarding the *number of participants* in conversations, all the related works contain corpora with many-to-many conversations with one exception: Oraby et al. [150] analyze customer service interactions which are one-to-one.

[5] The total percentage exceeds 100% as some works use corpora of multiple types of asynchronous communication.

[6] Enron corpus was publicly released by the US Federal Energy Regulation Commission in 2003 and includes email exchanges between 150 employees of the Enron corporation that went bankrupt in 2001 because of a large-scale fraud.

[7] The Talk pages are used to coordinate the editing process with the goal to improve the Wikipedia content [64].

2.3.2 Applications and Beneficiaries

The majority of the publications in the selected literature (84%) discuss applications and beneficiaries. Further, these are presented in three parts corresponding to the types of asynchronous communication: email, tweet and forum interactions.

Email. A first category of applications in the context of emails is to create new or to enhance existing email tracking tools with capabilities to monitor discussions. The goal of discussion monitoring is to manage shared tasks and track the status of ongoing activities [28, 40, 178]. By identifying speech act classes in emails, requests towards recipients could be automatically extracted and thus recipients can be assisted in creating to-do lists. The emails replying to sender requests could be tracked in order to identify if questions have been answered or commitments to actions have been accepted or refused. These automatic capabilities added to email managing tools can help users in scheduling, delegating and prioritizing from large volumes of emails. Consequently, one positive claimed effect is the alleviation of information overload by decreasing the effort spent for email management [79, 178].

A second category of applications is understanding email as a conversational genre by automatically uncovering structures and interactions through speech acts and adjacency pairs [79, 106, 107, 109, 199]. Multiple communities in humanities and social science, including linguistics, conversation analysis and psychology, are considered to benefit from such automatic modeling of email conversations [79]. The uncovered email structures and user interactions could further represent the starting point for more focused types of analyses. For instance, as email is a mandatory tool in organizations, email conversations annotated with speech acts can be analyzed in order to outline organizational behavior, such as leadership roles, engagement in meetings or team interactions [28, 79, 107, 138]. Moreover, marketing analyses can be conducted in order to enable email personalization for different targets [138]. Additionally, email conversations as sequences of speech acts can be used as knowledge in the creation of automatic dialogue agents [107, 199] or in conversation summarization tools [106, 199].

Twitter. Similar to the latter email applications, automatic speech act modeling of Twitter conversations is argued to help with revealing behavioral knowledge through the large-scale analysis of this type of communication. Hemphill et al. [91] propose an approach to analyse how politicians use social media, what they intend to achieve through their tweets and how their online, public posts are associated to other political activities. By identifying rhetorical questions, Ranganath et al. [165] aim at investigating persuasive tactics surrounding these types of questions, replies given by the community to these persuasive tactics and traits of people using them. Oraby et al. [150] work towards creating automatic customer service agents in Twitter. For this, several aspects are investigated from the annotated Twitter conversations: differences in speech behavior between customers and agents, automatic prediction of a conversation outcome

in terms of customer satisfaction, frustration and problem resolution, and identifying issues in live settings to allow the agents to appropriately adapt to a given problematic situation.

General applications of the automatic modeling of tweets with speech intentions target information seeking by allowing users to efficiently search for tweets that convey a specific intention such as questions or statements [165, 234]. Moreover, topics emerging in Twitter could be also studied in terms of speech act distributions, revealing the verbal behavior of the community towards those topics [213, 234]. Finally, within a Twitter conversation, speech act use could mark topic shifts or behavioral adaptation of users to specific communities [234].

Online Forum. The automatic modeling of forum discussions with speech intentions targets the improvement of information extraction, user assistance and community management.

Forum users could search posts by their conveyed intention [118, 158, 194]. Users may be interested in finding solutions to specific issues. In this case, speech act annotation can improve the results of information retrieval by weighting posts or threads containing accepted solutions higher than the others which are not relevant (e.g. off-topic conversations) or are not considered suitable answers (e.g. contain negative feedback) [14, 118, 158]. In order to assist users to navigate through the forum content, summarization based on speech acts may be another option. Qadir and Riloff [163] propose to summarize the questions being asked in forums over time, while Bhatia et al. [14] suggest to summarize a thread by showing only the posts containing questions and solutions. Similarly, in Wikipedia, by summarizing the discussion on article editing and publicly exposing it, readers could know what kind of issues are behind the article [64]. Also, in domain-specific discussions, when someone wants to get acquainted with the beliefs shared by the community, assertive utterances may be automatically highlighted [163].

Another direction to assist the users is to identify unanswered questions, in particular, in the context of online learning, where thousands of students need to be managed by a very small course staff [10, 166, 189]. Instructors can become aware of the existing problems communicated in the forum and of specific students who might need assistance.

Moreover, automatically identifying unanswered questions is also very relevant in technical forums. For instance, by presenting threads or posts containing unanswered questions first, experienced users can more easily identify them and offer help [14].

Interaction analyses are another point of interest enabled by the automatic modeling of forums with speech acts. By identifying student interaction patterns, insights are gained into strategies leading to resolved discussions, into student performance reflected by the forum participation and into the effectiveness of the course support [10, 13, 166, 189]. The interactions described as speech act exchanges could also reveal processes of collaborations in a group and help with identifying successful collaboration strategies [64]. Interaction analysis could also enable the identification of user roles in forum conversations [14, 189].

2.3.3 Speech Intention Taxonomies for Asynchronous Communication

Speech intention taxonomies from the studied literature are chronologically presented, grouped by the type of asynchronous communication. However, before starting, more details about dialogue acts are provided, as they are referred by several related works.

Dialogue act taxonomies cover a large range of conversational functions, specialized for the interpretation of utterance roles in dialogues. A widespread dialogue act taxonomy is Dialog Act Markup in Several Layers (shortly DAMSL) initially created by Core and Allen [7] and later modified by Stolke et al. [196] (version known as SWBD-DAMSL) in order to be more suitable for its application to telephone conversations. The modifications were aimed to linguistically discover interesting classes, but also to ensure an easier manual annotation. SWBD-DAMSL taxonomy is composed of 42 mutually exclusive classes. The major groups are statement and opinion, question, back-channel, turn exit and abandoned utterance, and answer and agreement. Back-channel utterances are short utterances through which one of the speakers suggests that the other should continue talking such as "Um."; they are also called continuer utterances in the conversation analysis literature. Turn exits and abandoned utterances appear like unfinished utterances and speakers breaking off in conversation. However, their actual role is to transfer the turn of speaking to the interlocutor. As it could be noticed, these two last groups of dialogue acts are very specific to spoken and, more general, to synchronous conversations.

Email. Cohen et al. [40] propose a taxonomy of verbs representing classes of speech intentions. These classes reflect negotiation-related speech intentions, but also some other conversational intentions such as *greetings* and *reminders*. Consequently, this taxonomy contains directive speech acts (*propose* and *request*) and commissive speech acts (*commit* and *refuse*). Apart from these, some classes specific to the communicative goals of negotiation or "common non-linguistic uses of emails" are also used, such as *deliver*—when an artifact is delivered after a commitment or information is provided, *remind*—when a coming deadline is communicated, and *amend*—when a previous proposal, subject to modifications, is communicated. This taxonomy is later used in two other research works [28, 29].

Goldstein and Sabin [79] propose a taxonomy of email acts, which is a mix of pragmatic and email-specific classes. The email-specific classes are *self emails*, *non-personal bulk* such as advertising or spam, and *transmissive emails* when forwarding documents. The pragmatic classes consist in both dialogue act-related classes and speech acts. The classes linked to dialogue acts are *responses*—when answers to questions are provided, and *responses with forward function*—similar to the previous category, but it also contains further questions. Then, the authors include the speech act classes mapped on Searle's taxonomy with one difference: directives are split in *requests for action* and *requests for information*. Also, a class *Other* is introduced to collect greeting-related and other introductory utterances.

Mildinhall and Noyes [138] adopt an already existing, general-purpose taxonomy: the verbal

response modes created by William B. Stiles [195]. This taxonomy has been extensively used in psychotherapy for studying therapist interventions. Similar to Searle's taxonomy, the proposed categories are extensive and exhaustive; thus, any spoken or written utterance can be associated with one of the classes. An alignment between Searle's and Stile's classes is possible, although not all verbal response modes have an equivalent speech act [124]. *Disclosure* is associated with both expressive and commissive, *edification* is mapped on assertive, directive contains *advisement* and *question*. Apart from these, the *confirmation* class subsumes agreements, disagreements, acknowledgements—serving to acknowledge a previous utterance (but not for agreeing, disagreeing) or for salutations, *interpretation* is a type of judgment on the addressee or an explanation, and *reflection* aims to put the addressee's experience into the speaker's words—to repeat or to clarify.

Jeong et al. [106] adapt dialogue acts for application to email and forum corpora. As the authors emphasize, some classes designed for the analysis of spoken corpora that materialize interactions specific to colloquial style (e.g. backchannel, floor grabbing) are not applicable to asynchronous conversations. Therefore, these groups of dialogue acts are not considered. Most of the used classes capture the diversity of questions including *rhetorical*, *open-ended* and *yes-no question*. Apart from these directives, requests and suggestions grouped under the *Action motivator* class are also found. Then, four classes characterize response types: *accept*, *reject*, *uncertain response* and *acknowledgement*. *Polite mechanism* class consists of expressive speech acts such as showing sympathy, thanking, apologizing, welcoming. Finally, there is a class aligned to assertive and is called *statement*. This taxonomy is later employed by Tavafi et al. [199].

Aiming at an abstract taxonomy that enables the analysis of adjacency pairs, Hu et al. [100] define categories inspired from dialogue acts too. Thus, two types of directive are proposed: *request for information* and *request for action*, *inform* that is a type of assertive, *commit*, *conventional* that partially overlaps with expressive, but contains also introductions, and *performative* that appears as declarative. However, in the automatic solution, only the directives, inform and conventional speech acts are used. These classes and the released corpus are also used in [149].

Twitter. While the taxonomies employed with email corpora have many classes in common, the taxonomies for analyzing tweets vary much more across the studied works.

The first solution on interpreting intentions behind tweets is proposed by Ritter et al. [171]. The classes are not defined beforehand as unsupervised machine learning is used. The speech intention classes are proposed by manually reasoning on the characteristics of the discovered clusters. The classes are similar to dialogue acts, in the sense that they represent to a less extent pragmatic intentions; instead, they aim to capture the role of different tweets: *question* to followers or to individual users, *reference broadcast*, *reaction*, *comment*, *answer* and *response*. Reactions and comments are semantically equivalent. Nonetheless, their associated tweets have a clearly differentiated form, shown by the discovery of two separate clusters, one for reactions

and one for comments. The same remarks apply to answers and responses. The corpus released by Ritter et al. [171] is later used by Paul [154].

Contrary to the previous work, Zhang et al. [234] start with a manual pragmatic analysis of Twitter communication for defining the classes. Searle's taxonomy is used as a point of departure. However, several adaptations are made. Assertives are entirely covered by the *statement* class. Directives are split in two classes—*question* and *suggestion*. As for expressives, it is claimed that they are mapped on the class *comment*. However, from the example presented in the paper, it appears that the class *comment* could also have other roles apart from revealing feelings. Finally, declarative and commissive are grouped into a generic category named *miscellaneous*.

Although the work of Hemphill et al. [91] is focused on the analysis of political tweets, the proposed taxonomy is rather abstract. However, the specific context and goal of this work appear to influence the definition of the classes. The assertive category contains, on the one hand, the *narrating* class used for tweets reporting stories or descriptions and, on the other hand, the *positioning* class used for tweets where a Congress member positions himself or herself in relation to other politicians or political issues. Then, compared to all the previous works, a novel class within the directive is defined: *directing to information*, which groups tweets that re-direct readers to external resources through urls. Apart from this, the taxonomy contains three other classes: r*equesting action*, *giving thanks* and *other*.

Vosoughi and Roy [213] start from Searle's taxonomy for defining their own. *Assertive* and *expressive* are maintained as in the original work. The directive class is composed of multiple speech intentions: *recommendations*, *questions* and *requests*. Moreover, similar to Zhang el al. [234], a *miscellaneous* class captures instances of commissive and declarative.

Ranganath et al. [165] focus only on identifying *rhetorical* and *non-rhetorical* questions from tweets. Thus, only these specific classes belonging to directive speech acts are used in the solution.

The most elaborated taxonomy used in the context of tweets is proposed by Oraby et al. [150], which is also the most recent work modeling tweets with speech intentions. The taxonomy is created starting from dialogue acts and is represented at two different levels. At the highest, more general level, the *greeting*, *statement*, *request*, *question*, *answer* and *social act* classes are found. At the more specific level, *greetings* can be of type *openings* or *closings*; *statements* are a highly mixed category containing assertives, positive and negative expressives and commissives; *requests* could be *requests for information* or *for action*; specific types of questions are proposed: *yes-no questions*, *open question*, *questions starting with 5w1h*[8]; answers could be *positive*, *negative* or *acknowledgements*; social acts include *thanking* and *apologizing*.

Online Forum. Many forum-related taxonomies are created to be applied to problem-solution conversations, either in the context of educational or frequently asked question (FAQ) forums.

Ravi and Kim [166] propose several classes in this direction: *questions* represent inquiries

[8]Who, What, Where, When, Why, How

about problems or about previous answers; *suggestions and advices* are responses to questions; the *inform* class includes utterances with information and commands[9], *elaboration* refers to utterances providing details to previous questions or answers. Finally, two classes conceptualize reactions to proposed solutions: either positive through *acknowledging, complimenting or supporting* a solution or negative through *correcting, objecting or complaining* about it.

Similarly, Kim et al. [118] propose a taxonomy related to asking questions and receiving answers. This work is later used by Tavafi et al. [199] and Jo et al. [107]. There are various types of *questions* in the taxonomy: questions that start a thread, additional or follow-up questions to the existing ones, confirmation questions emphasizing mistakes in the previous questions without providing solutions or confirming details, and correction questions showing corrections to existing questions. The *answers* are also diversified into simple answers that represent solutions to questions, answers that are elaboration on previous answers, confirmation answers that outline errors without providing solutions, objection answers and correction answers. The *resolution* class represents answers confirmed to work. Compared to the previous taxonomy [166], a new class *reproduction* is introduced to either confirm the presence of a certain problem or that the an already accepted solution works for other forum users too.

Other taxonomies related to problem-solutions conversations are proposed in the literature [10, 13, 14, 189]. The alignment of speech intention classes across all these works is presented in Table 2.7. The taxonomies created by Arguello and Shaffer [10] and Bayat et al. [13] largely overlap and slightly differ from the others. Both have the following common classes: *questions*, *answers* and *positive* and *negative acknowledgements*. One main difference is that, although both have a class called *Issue*, in [10] this class contains utterances conveying dissatisfaction with the logistics of the course, being directed towards the course staff, while *Issue* in [13] refers only to misunderstandings and unclear concepts for solving a problem. Also *Issue* in [10] can have associated an answer of type *Issue resolution*, representing a solution to the raised issue and not to an asked question regarding the course content. In [13], a class called *Reference* related to answers is used too when only hints or suggestions are offered, but not complete solutions.

Compared to the taxonomies presented so far which use ad-hoc classes designed for specific contexts, Qadir and Riloff [163] adopt an abstract taxonomy by explicitly using Searle's speech acts, excepts for declaratives.

Then, Song and Diederich [194] extend Searle's by splitting directives in *questions* and *actions*. Moreover, assertive is also divided in different speech intentions: statements of *facts*, *self situations*, *preferences* and *opinions*. Compared to Searle's taxonomy, expressives are not explicitly covered, but implicitly, through preferences and opinions. Also, new speech act classes are introduced: *prohibitive*—specifying interdictions, *permissive*—specifying permissions, and *condition*—used to express conditional situations [194].

[9]Being in the context of an online course, the *inform* class is a mix between assertive and directive and the corresponding utterances appear to be mainly written by instructors or course staff [166].

Table 2.7: Forum-related taxonomies—an alignment of classes proposed or used across different works (cited in the first row).

[166]	[107, 118, 154, 199]	[14, 158]	[189]	[10]	[13]
Question	Question-Question, Reproduction—confirms same issue	Question, Repeat question	Problem presenting	Question, Issue	Question, Issue
Answer, Suggestion, Advice	Answer-Answer	Solution	Solving	Answer, Issue resolution	Answer, Reference
Acknowledgement, Support, Compliment	Resolution, Question-confirmation, Answer-Confirmation, Reproduction—answer works	Positive feedback	Solution appreciation	Positive acknowledgement	Positive acknowledgement
Correction, Objection, Complaint	Question-Correction, Answer-Correction, Answer-Objection, Question-confirmation, Answer-Confirmation	Negative feedback	Solution objecting	Negative acknowledgement	Negative acknowledgement
Elaboration of a question or answer	Question-Add, Answer-Add	Further details	Problem understanding, Solution understanding		
Information, Command, Announcement	Other	Junk		Other	

Ferschke et al. [64] design a taxonomy in two levels with intentions related to the collaborative work process to create and improve Wikipedia. The classes are dependent on the communicative goal and often an individual class subsumes multiple speech acts. The *article criticism* class differentiates between multiple types of problems; *explicit performative* includes suggestions, recommendations, requests, but also commitments to actions; *information content* is used for utterances providing information, requesting information or suggesting changes to existing content; *interpersonal* is related to positive and negative attitudes and to acceptances and rejections regarding a previous request.

Comparison Analysis. Apart from Cohen et al. [40] and Goldstein and Sabin [79] who introduce ad-hoc classes, personalized for the communicative goal and corpus type, the taxonomies used in the automatic speech act annotation of emails aim to be general. Also, a part of these classes can be semantically aligned and to a great extent cover the main Searle's speech act types. Both dialogue act and speech act classes are the foundation of these taxonomies.

With regard to the particularities of email taxonomies, [106] is the only work that expands questions into five independent classes. Moreover, except from the taxonomy proposed by Goldstein and Sabin [79], no other contains declarative speech acts. Then, the taxonomies in [100] and [40] are significantly less comprehensive compared to the others. However, Cohen et al. [40] annotate entire emails, not individual utterances, fact that increases the likelihood that at least one of the proposed classes can be associated with the content. Moreover, a class *Other* is introduced by these works for when none of the other proposed classes is applicable [40, 100].

Several tendencies can be observed regarding the email taxonomies. First, directives are often differentiated between questions or requests for information and request for action that includes advices and requests. Second, most of the works [40, 79, 106, 138] propose classes of speech intentions, dependent on previous utterances, such as agreement, accept, refuse, acknowledgement and response. Consequently, the adjacency pairs appear central to the study of emails. As stated by most of these works, the automatic annotation of emails with speech acts is aimed at the creation of email tracking systems, which clearly motivates the need to investigate the development of conversations, as sequences of speech acts. Third, expressives are rarely represented as standalone classes; they are partially covered by greetings or combined with other speech act classes under more ample categories (e.g. disclosure, conventional, polite mechanism).

In the pragmatic study of tweets by the computer science community, the directive and assertive-related classes are the most common. Furthermore, a similarity shared with the taxonomies for email corpora is that expressives rarely appear as a standalone class. They are partially covered through thanks, greetings or are part of comments. Moreover, the directive speech acts continue to be studied as multiple sub-classes, specific to questions, requests, suggestions and, the newly proposed, directing-to-information [91].

Commissive speech acts are never used for tweet corpora. The Twitter-specific taxonomies are

empirically defined by manually analyzing the corpora or the clusters obtained from unsupervised learning. Most often, datasets are collected through Twitter streaming API leading to mainly public tweets composing the corpora. Public commitments are rather rare, promises being more frequently uttered in personal settings—unless there is a particular context, such as a political discourse. Furthermore, commitments, in the form of accepts and refusals, are expressed in conversations as a result of directives; nonetheless, such exchanges are more often private, hence not collected through the stream. Then, in Twitter, there is less focus on adjacency pairs in taxonomies, apart from the work of Oraby et al. [150], which is inspired from dialogue acts, and two other solutions [171, 234], which superficially cover this aspect through the class comment. The taxonomy of Oraby et al. [150] is employed on private conversations for customer support, which also explains the existence of more diversified speech act classes and adjacency pairs.

The majority of taxonomies for forum conversations are highly dependent on the communicative goal. Moreover, as presented in Table 2.7, these taxonomies specific to problem-solution discussions overlap. Only one work uses dialogue act classes [106] and two taxonomies pragmatics-based classes [163, 194]. Compared to the taxonomies for tweets and emails, expressives appear more frequently here and can be positive or negative. Moreover, commitments and specialized classes of answers, relevant to the study of adjacency pairs, are present in all the studied works.

When creating new taxonomies or adopting existing ones, the originators appear to prioritize different requirements. The majority of the proposed taxonomies are dependent on the conversation dimensions and communicative goals. Also, a focus on adjacency pairs guides the efforts of many works in the context of emails and forums, but it is not a priority for Twitter. There are some works that explicitly favor general schema and others claiming that general taxonomies fall short [150]. Tavafi et al. [199] argue that domain-independent taxonomies, applicable to different types of corpora, could enable the analysis of conversations spanning various media, such as conversations starting in live meetings and continuing by email. However, the corpora used in their solution—originally released by [106] and [118], is tagged with both general and specialized classes. Then, Hu et al. [100] claim that their adopted classes of speech acts are suitable for any type of communication across the presented dimensions; while this is true, the proposed classes are not exhaustive. By relying on the original work of Searle [184], the taxonomy proposed by Qadir and Riloff [163] tries to be as domain-independent and exhaustive as possible. Nonetheless, compared to taxonomies emerging from dialogue acts, their classes are coarse and do not enable a detailed analysis of dialogues. In conclusion, exhaustiveness and high granularity appear as rare traits of general, pragmatics-based taxonomies in the studied computer science works.

2.3.4 Manual Annotation of Asynchronous Communication

The manual annotation with speech act classes is necessary for two situations:

1. for the validation of a newly proposed speech intention taxonomy;

2. for the creation of a ground-truth corpus, for training and evaluating supervised and semi-supervised machine learning algorithms.

In the former case, the originators of the taxonomy want to demonstrate that the taxonomy is used in the same way by multiple humans, showing thus that the interpretation of the classes is consistent. Apart from the use in supervised learning, the ground-truth corpus may be also useful in unsupervised approaches to evaluate the results; nonetheless, other evaluation methods not relying on the existence of an annotated dataset also exist, as already mentioned.

Annotation unit. The unit of annotation in the majority of the studied works, 81% of the unique corpora, is the turn. A turn can be a post, an email or a tweet and may be composed of multiple sentences. There are two corpora annotated per sentence [106, 163] and two other corpora for which the annotation is applied to parts of sentences [100, 194]. These four corpora include email and forum conversations, while tweets are always annotated per turn.

Annotation strategy. With regard to the single or multi-label choice in manual annotation, several tendencies are observed. Twitter corpora is almost always a single-label corpus, with two exceptions [91, 150]. These two works claim that, although tweets are rather short messages, they often realize multiple speech intentions. Moreover, allowing multiple labels per tweet is claimed to be more informative in the analysis [150]. Forum corpora is predominantly multi-label. In three situations, posts [14, 189] and parts of sentence [194] have associated single classes. By contrary, the annotation strategies for email corpora appear more mixed with single and multi-label annotations equally preferred. All of the works opting for multi-label annotation apply it per turn [28, 29, 40, 178], while the single-label annotation is applied per sentence [106], per part of sentence [100] and per turn [79]. An argument in favor of multi-label annotation is that posts or emails could simultaneously contain multiple speech act classes such as requests and answers to previous questions [154].

In general, from the studied literature it appears that, when the unit of annotation is a turn, multiple speech acts are associated with it. However, in the case of spoken and synchronous conversations, turns are often segmented in individual units (e.g. sentences, parts of a sentence), such that each unit conveys a single dialogue act. Automatic segmentation is a challenging task and, sometimes, solutions for simultaneous automatic segmentation and speech acts tagging have been also explored [237]. The community focusing on the speech act analysis of asynchronous conversations seem to have mixed opinions about segmentation. In the case of tweets, segmentation is considered to cumber the annotation for several reasons: these messages are very short; the speech acts could overlap, hence the same content could convey multiple speech intentions; it is a significant increase in effort for the human annotators [150].

Aligned with the last remark, Sappelli et al. [178] also state that a previous work [124]

showed that human annotation led to more reliable results when applied per turn than per part of sentence. However, in the works presented so far, segmenting by grammatical sentences or parts of sentence is sometimes applied. Moreover, segmentation is suggested as future work to improve the annotation results in terms of annotator's agreement[10] by several works [40, 63]. Another way to handle the existence of multiple speech acts per turn without segmentation, while in practice allowing for a single label is to propose classes that incorporate multiple speech acts. Such an example is *responses with forward function* proposed in [79]. These mixed speech act classes are also found in works with multi-label corpora—the class *propose* implies both a *direct* and a *commit* to the proposed actions in [40].

Ground-truth corpus. The creation of a ground-truth corpus is necessary in supervised learning. Several of the studied works based on supervised learning do not report whether the used corpora are validated [100, 166, 194, 234]. Ranganath et al. [165] created the ground-truth corpus by collecting tweets which contained the hashtags "#rhetoricalquestion" or "#dontanswerthat" for positive instances and which contained question marks for negative instances. Implicitly, it was assumed that the hashtag was always correctly used, which may be a threat to validity. In the same time, it was an approximate and sufficiently reliable way to collect many instances with minimum effort.

When the developed strategies for the identification of speech intentions are unsupervised and the measures proposed to analyze the performance of the algorithms do not rely on true labels, the existence of a ground-truth corpus is not required (e.g. [154, 171]).

Most of the unique corpora used in the related works are nonetheless validated (76%). The most common approach is to have annotated a selected dataset by 2 human coders and then compute the Cohen's Kappa score to evaluate the inter-rater agreement (see Table 2.8) [40, 79, 91, 118, 138, 163]. The Cohen's Kappa score is computed for each class and, when reported as an overall score, it appears averaged over all scores obtained for each class. Ferschke et al. [64] claim that a better alternative to averaging is to use the pooled Kappa score.

Table 2.8: Statistical measures used in the related works for validating ground-truth corpora: the first column shows the type, the second column shows the percentage of works that uses it from the total number of works reporting validation. The sum is not 100% because several works use multiple statistical measures.

Type of statistical measure	**Percentage**
Cohen's Kappa	60%
Fleiss' Kappa	13%
Other statistics	27%
Not explicit	27%

[10] As reminder, annotator agreement is a statistical measure (e.g. Kappa statistics) which quantifies the extent to which multiple annotators similarly apply some classes on a given corpus [27].

Some datasets are annotated by more than 2 human annotators. In this situation, either Fleiss' Kappa is used [10, 150], which is a type of Kappa statistics suitable for any number of coders, or Cohen's Kappa is applied instead for each pair of coders [178]. Another strategy is reported too for the case when expert annotations are important. Then, the classes selected by the majority of annotators are compared with the expert annotations using Cohen's Kappa [10]. Using the majority vote to decide on final labels in the ground-truth corpus is a strategy used in several other works too (at least 2/3 votes or 3/5 votes) [14, 150, 213].

Among other types of measures used for validation, there are: a method based on Jaccard distance for assessing the validity of the annotations of independent annotators with regard to the annotations provided by the authors of written content [178]; a new method to measure how frequently an annotator agrees with most of the annotators [150]; and Krippendorff's alpha, another statistical measure for inter-rater agreement [64]. In several works [13, 106, 189, 213], a Kappa score is reported, but its type is not explicitly mentioned.

Another important aspect of the ground-truth corpus creation is how to handle disagreements. While the Kappa measure could assess the degree of agreement between multiple annotators and implicitly be a measure of validation, a strategy on how to approach the conflicting annotations still needs to be defined. Previously, such a strategy was mentioned: to use the majority vote when at least three human annotators are involved [10, 14, 150, 213]. However, even for this situation, if the inclusion threshold is not reached a decision is still required. One approach is to discard these problematic units from the ground-truth corpus [13, 14]. Another approach is to have experts to discuss disagreements and decide on final labels [63, 118, 163].

2.3.5 Automatic Annotation of Asynchronous Communication

The automatic solutions to annotate asynchronous communication with speech intentions are further discussed. Several aspects are covered: text processing activities, sampling approaches to balance corpora with uneven number of instances per class, types of automatic approaches and their corresponding machine learning algorithms and feature groups.

Text preprocessing. Segmentation is explicitly mentioned by several works and is used either for sentence delimitation in forum posts [13, 158] or for turn identification in Wikipedia Talk pages [64]. The most common text preprocessing activities are stemming [10, 13, 14, 79, 158, 166, 194] and lemmatization [118, 149], with a stronger preference towards stemming. Regarding stopwords, several works explicitly mention their inclusion as beneficial [10, 109, 158], while others prefer to exclude these function words from text [13, 14, 194]. Placeholders are frequently used to replace domain specific content such as code, files, technical words [13, 29, 166] or even nouns [109]. Then, placeholders are also used to mask urls, numbers and time markers or to encode specific types of function words such as pronouns or 5W1H[11] [13, 29, 107, 166]. Finally, replacing

[11] Who, What, Where, When, Why, How

informal words with their formal version and emoticons with their meaning is also used in [166] and [189]. Although the rest of the studied literature most likely uses text preprocessing, these activities are not explicitly discussed in publications.

Sampling. [64, 79, 189] use sampling methods. Goldstein and Sabin [79] merge subcategories into larger classes. Then, the classes with very few instances are either discarded or completed by manually selecting new instances. In binary classification, when the corpus is transformed in positive and negative instances for a target class, the negatives samples could largely outnumber the positives ones. To overcome this issue, Ferschke et al. [64] randomly select negative instances until their number is equal to the number of positive ones. Grouping less popular classes and using oversampling and under-sampling are other strategies reported by Shen and Kim [189].

Feature groups. An overview of the feature groups used in the supervised and semi-supervised solutions is presented in Table 2.9. The most common feature groups are n-grams or related to punctuation or to the position in conversation. The least common features are based on external lexicons, either domain-specific or general-language, and on text similarity. However, taxonomy-related expressions, which are lexicons with cues for speech act classes, are rather frequent. Further, details regarding each feature group are provided and discussed.

While in most solutions using n-grams these features are extracted from the target unit text, few works also consider the n-grams of the neighboring units, such as of the previous and next turns [64, 189]. Text similarity is computed for several scenarios: to identify the most similar post in a thread based only on post titles or post content [118]; between a conversation title and the post [14]; between a post and the first post [10, 14]; between current and previous posts [10]; and between current post and the complete thread [10, 14].

With regard to time, dates, numbers and urls, their presence or frequency in an unit is extracted. Also the time distance between the current turn and the previous turn in the conversation is used as a feature [10, 64, 150].

The morphological and syntactic features appear under different forms: the presence or frequencies of certain POS tags [40, 100, 213], n-grams of POS tags [106, 149], n-grams of pairs composed of lemma/stem for a word and its POS tag [13, 149] and related to dependency trees [106, 213].

Contextual features encode information about what are the true or estimated classes of the parent unit [10, 13, 28, 100, 118] or of the child unit [13, 28]. The full prediction history—all classes predicted for all the previous posts, is used in [118] too.

As mentioned in the group description, the verb-related features can mark the presence of speech act verbs collected from various theoretical lexicons [79, 163, 213]. Also, verb-related information can be extracted from the unit text such as the presence of modal verbs [10, 149, 163] and of specific tenses [10, 163] or the position of the first verb in the unit [149, 163].

Table 2.9: Percentage of related works using each feature group—only the supervised and semi-supervised machine learning solutions are considered.

Feature group	Percentage	References
bag-of-words (1)	32%	[28, 40, 91, 100, 106, 165, 194]
n-grams (2)	73%	[10, 13, 29, 40, 64, 106, 118, 149, 150, 163, 166, 189, 199, 213, 234, 235]
text similarity (3)	14%	[10, 14, 118]
time, date, numbers, urls (4)	32%	[10, 29, 40, 64, 118, 150, 166]
morphological & syntactic (5)	36%	[10, 13, 13, 40, 100, 106, 149, 213]
contextual (6)	18%	[10, 13, 28, 118]
verb-related (7)	23%	[10, 79, 149, 163, 213]
pronoun-related (8)	32%	[10, 29, 40, 79, 150, 163, 166]
sentiment-related (9)	32%	[10, 14, 150, 166, 213, 234, 235]
taxonomy-related expressions (10)	32%	[14, 29, 79, 100, 150, 163, 166]
punctuation (11)	50%	[10, 14, 79, 100, 118, 149, 150, 163, 213, 234, 235]
unit length (12)	32%	[10, 13, 14, 64, 79, 100, 199]
position in conversation (13)	50%	[10, 13, 14, 64, 100, 106, 118, 149, 163, 189, 199]
author identity (14)	32%	[10, 13, 14, 106, 118, 189, 199]
type-specific (15)	41%	[10, 14, 29, 64, 79, 106, 213, 234, 235]
domain-specific lexicons (16)	9%	[163, 166]
general-language lexicons (17)	18%	[10, 213, 234, 235]

The sentiment-related features are extracted by performing sentiment analysis on the unit text [10, 14, 213, 234, 235] or by considering the types of used emoticons [150, 166, 213, 234, 235].

The taxonomy-related expressions highly depend on the adopted classes and can be very specific to the type of asynchronous communication [79]. However, generic expressions for the theoretical classes of speech acts, such as for commissive [79, 163], directive [14, 29, 150, 166] or expressive [79, 150], are also defined.

The punctuation marks investigated are exclamation and question. Either their presence or frequency in an unit are transformed in features.

Across the studied works, the features with respect to author identity vary. Information regarding his/her profile, such as the role [10, 13] or the authority score [14], or regarding the past interventions, such as if he/she is the initiator of a thread [14, 118, 189], the past number of posts [14] or the profile as a speech intention distribution [118], are extracted as features too.

Communication type-specific features are employed in multiple solutions for modeling: email [29, 79], forum [10, 14, 64, 106] and Twitter [213, 234, 235] conversations.

General-language lexicons are mainly used in solutions designed for tweets and consists of vulgar words [213, 234, 235] or abbreviations specific to online chats [213].

In unsupervised learning, the most frequent features are bag of words and n-grams [107, 109, 109, 158, 171]. Cosine similarity is used to quantify the text similarity between a post

and the initial post [158] and between BoW vectors [109]. With regard to contextual features, Mildinhall and Noyes [138] compute transition probabilities between classes of speech acts, which are further used in clustering, and Paul [154] defines the probability of an intention class to be associated to a text as a log-linear function of the intention classes in the previous block of text. Author identity appears as a common feature group in the proposed unsupervised solutions [107, 109, 158]. Some detailed information about the writer are considered too, apart from his/her identity: the number of posts in a thread created by the author, the number of posts in a forum created by the author, if the same person is the author of the previous post [158] and the tendency of an author to use some specific speech intentions [107]. Other less frequent feature groups are related to POS tags and dependency trees, the unit length, the position in conversation and punctuation marks [109, 158].

Using n-grams and BoW can result in a very large number of features, which may negatively impact the complexity of the machine learning algorithms. For this reason, selecting only the most relevant n-grams or words for each class is tackled by multiple works. Three approaches to assess relevance and perform feature selection are reported: one based on the χ^2 statistical test [10, 64], one based on Information Gain [29, 64, 166, 189] and one performing an exhaustive search over all combinations of features [149].

Automatic modeling approaches. The most common approach to automatically model asynchronous communication with speech intentions is based on supervised machine learning algorithms—60% of the studied works. In Table 2.10, specific algorithms are presented. Often, multiple algorithms are explored in a solution and their performance is compared. Log-linear models, support vector machines and decision trees are very frequent. In most of the works, SVM has a linear kernel, but SVM with polynomial kernel is also used by Qadir and Riloff [163]. Moreover, the hybrid version of SVM, suitable for annotating sequences, is also integrated in multiple solutions. Conditional random fields, another sequence modeling approach that takes into consideration the neighboring instances, also appears among the algorithms explored. Various implementations of the perceptron classifier are evaluated: voted perceptron [40], multi-layer perceptron [14, 194] and single-layer perceptron [194].

Table 2.10: Supervised and semi-supervised algorithms used in the related works.

Algorithm	**References**
Maximum Entropy (MaxEnt), Logistic Regression	[10, 14, 28, 91, 118, 189, 213]
Support Vector Machine (SVM)	[13, 14, 40, 64, 79, 100, 149, 163, 166, 189, 199, 213, 234]
SVM-Hidden Markov Model (SVM-HMM)	[100, 118, 150, 199]
Conditional Random Fields (CRF)	[118, 189, 199]
Bayesian Networks	[14, 64, 91, 194, 213]
Decision Trees and Random Forests	[14, 40, 64, 79, 91, 189, 213]
Perceptron	[14, 40, 194]

Carvalho and Cohen [28] create an iterative approach on top of the predictions provided by maximum entropy classifiers to simultaneously assign speech intentions to all turns of a thread. Qadir and Riloff [163] opt to predict speech intention classes in two steps: first, to identify if an unit is expository and second, if not expository, to predict the classes. A hierarchical classification method is similarly proposed by Omuya et al. [149]. However, in this work, the preference is given to minority classes as follows: if the classifier trained to recognize the minority class predicts true, then assign the class to the instance; else continue with each classifier, in the ascending order dictated by the frequency of their associated classes, until a label is identified.

Three related works report using semi-supervised machine learning [106, 158, 235]. Zhang et al. [235] approach the modeling in a semi-supervised manner first through transductive SVM and second through graph-based label propagation. Jeong et al. [106] enhance a baseline implemented as an AdaBoost classifier with two semi-supervised strategies: bootstrapping and semi-supervised boosting.

Unsupervised solutions are explored in 30% of the cases. The conversation models are implemented using Hidden Markov Models [107, 109, 154, 158, 171], one solution showing also an alternative semi-supervised approach [158]. Clustering is also applied in [138] (the hierarchical clustering and k-nearest neighbor algorithms) and in [109] (graph-theoretic clustering), or as a step before training the HMM on clusters in [158] (hierarchical clustering).

The annotation units in the automatic solutions are maintained the same as reported in the manual annotation.

In supervised approaches, the annotation strategy for instances with multiple labels is to create binary classifiers for each class [10, 13, 28, 64, 91, 163]. Thus, each classifier predicts if a certain instance belongs to the class that the classifier is trained to recognize. The complete label set for an instance consists then in all the labels predicted true by the classifiers. In unsupervised approaches, multiple speech intention classes can be inherently associated to an instance by discovering clusters associated to mixtures of classes [138] or by considering that specific text blocks are generated from multiple states representing the classes [107].

2.3.6 Considering Relations among Speech Intentions

Relations among speech intentions are considered in several related works. Some solutions assume the presence of such relations and exploit them by including information about surrounding speech acts and turns in prediction. Other works aim at discovering these relations either simultaneously with speech acts or separately.

The first work to investigate relations among speech intentions for modeling asynchronous communication was presented by Carvalho and Cohen [28]. Their intuition is that negotiations and task-related email conversations are sequential in nature and exploiting this aspect could lead to improved email act classification. The first step of their study was to derive a transition diagram, showing transition probabilities between the most common email acts. The second step

was to test if indeed using the surrounding speech acts—of both parent and first child turns, was effective in prediction compared to using only content features—bag of words. Although the results were promising for some speech act classes, this approach was recognized to be unpractical because it relied on the first child email act.

Therefore, in a third step, a new solution was proposed which assigned classes simultaneously and iteratively to all emails in a thread. This method was a collective discovery of email acts starting from the classes predicted by the local classifiers trained on content features only and then iteratively updating and inferring confidence factors of email acts by exploiting relational information [28]. The simultaneous prediction of labels for sequences by exploiting unit dependencies has been tackled by other works too using structural learners, such as SVM-HMM and CRF [100, 118, 150, 189, 199].

Relying on the prior work of Carvalho and Cohen [28], which obtained improved prediction results by considering the sequential correlation between speech acts, Arguello and Shaffer [10] approach similarly the modeling of forum conversations. The proposed solution exploits the relations among speech intentions by including in the feature set of the current turn the confidence values for each class being associated with the previous turn. The confidence values are obtained by predicting the speech act classes of the previous turn, without any of these contextual features. In a similar way, Bayat et al. [13] consider as feature the speech act of the previous turn and Kim et al. [118] define a more extensive set of contextual features—the predicted classes of the previous post, of the previous post from the same author and of all the previous posts in the thread.

Apart from discovering speech act classes, Hu et al. [100] create an approach to automatically identify links between emails, which are considered interrelated, by reasoning on their conveyed speech acts. For instance, an initial email can contain a request for action and one of its responding emails can contain a commitment. These links, also referred to as adjacency pairs, do not necessarily imply that the initial and responding emails are literally adjacent. The speech acts of emails are exploited as features in the link prediction. Similar to Hu et al. [100], Kim et al. [118] explore the discovery of links between posts in forums, by exploiting contextual features regarding the speech acts of previous posts.

In the solutions presented so far, the existence of relations among speech intentions was assumed true and was mainly exploited in the speech act prediction by considering the neighboring speech acts as features or implicitly in structural learners. However, these relations can also be discovered either per thread, by automatically identifying the links between interrelated turns [100, 118], or per corpora, by aggregating multiple annotated conversations in a transition diagram featuring regularities among speech act classes [28]. This latter approach was also implemented by two other works [178, 189].

Sappelli et al. [178] studied how email exchanges between two people evolved in time. Specifically, pairs of linked emails were transformed in pairs of email acts. Then, a transition diagram

provided an overview of the overall exchanges by showing common pairs of email acts and how many times they appeared in the corpus. Shen and Kim [189] identified frequent dialogue patterns, where each pattern was represented as a sequence of three states: previous, current and next speech acts. These patterns and their frequencies were studied for two different groups in the context of an online course forum: students versus instructors. A state transition model based on speech act classes and frequent dialogue patterns was also derived to characterize question and answer forum conversations.

Another group of works using unsupervised machine learning with HMM jointly discovered speech act classes and their relations [154, 158, 171]. Transition diagrams were obtained, each node in the diagram representing a cluster of sentences supposed to have in common a speech act class or a mixture of speech act classes. The links between nodes showed transition probabilities.

2.4 Conclusions

This chapter presented a systematic analysis of the literature on modeling asynchronous conversations with speech intentions. In order to answer the proposed research questions, a detailed research process was followed (presented in Section 2.1). Further, the research questions are revisited in Subsection 2.4.1, the study limitations are presented in Subsection 2.4.2, the contribution of this study compared to the related works is summarized in Subsection 2.4.3 and future directions of inquiry on the studied topic are formulated in Subsection 2.4.4.

2.4.1 Research Questions Revisited

RQ_L1. The first research question investigated how the related works automatically modeled asynchronous conversations with speech intentions and targeted two areas:

1. what was the output of the modeling technique, hence what types of speech intentions were identified and if relations among these intentions were also discovered;
2. how various types of asynchronous communication were automatically annotated with speech intentions or modeled as processes of interrelated speech intentions.

For the second area focusing on the method multiple aspects were discussed:

- types of annotation unit—turn, sentence, part-of-sentence;
- types of annotation strategies—single- or multi-label;
- the creation and validation of ground-truth corpora;
- text preprocessing activities, such as lemmatization, stemming and segmentation;
- sampling methods for unbalanced corpora;

- machine learning approaches for the automatic annotation of conversations with speech acts—supervised, semi-supervised and unsupervised;
- types of features defined for the machine learning algorithms.

Corpora belonging to three types of asynchronous communication were used in the related work: email (in particular, before 2010), forums and Twitter (with a growing interest in the last decade). 21 taxonomies of speech intentions were identified in Section 2.3.3.

The annotation unit was mainly a turn, with few exceptions in forums and emails, when sentences or parts of sentence were annotated instead. The annotation strategy varied across the different types of asynchronous communication: mainly single-label for tweets, mainly multi-label for forums and a mixed preference for single- and multi-label for emails. In case of emails and forum posts, the tendency to associate an unit to a single label was observed for units that were sentences or parts of sentence.

In most works using supervised machine learning, multiple human annotators coded the same dataset and the reliability of the concurrent annotations was assessed with Kappa statistic[12]. Thus, the joint creation and validation of the ground-truth corpus was performed.

Unit text was often stemmed and lemmatized. Also, placeholders were used in some works to mask domain-specific content, urls, numbers or different types of function words such as pronouns.

Sampling was rarely used in the studied works and consisted either in merging classes with few instances or in oversampling or under-sampling.

A large number of feature groups was discovered and specific features within each group were described. Some feature groups appeared in most works, such as content-based features (BoW, n-grams), while others were very rare, such as domain-specific lexicons. Then, approaches for feature selection were identified, the most frequent being based on information gain.

Finally, details were provided about each type of the automatic modeling methods: supervised, semi-supervised and unsupervised. Supervised approaches were the most common—in particular using SVM and the hybrid learner SVM-HMM. Unsupervised approaches consisted in conversation modeling with HMM, complemented with other clustering algorithms.

Few works modeled the output as processes of interrelated speech intentions. Processes were represented as transition diagrams and were discovered either post-annotation or simultaneously with the speech act classes in unsupervised solutions.

RQ_L2. The second research question assessed how well the solutions presented in the related works satisfied the goals and properties of an envisioned solution, defined in Chapter 1. The envisioned solution should be automatic, corpus-independent and effective by revealing

[12]Alternative measures to Cohen's Kappa, the most frequently used test, were also discussed together with strategies to handle disagreements, such as experts discussion.

comprehensive and relevant knowledge. The automatically discovered knowledge should be also correct and complete proving the algorithmic performance of the proposed solution.

All the presented methods were automatic. However, not all the methods were corpus and domain-independent because of how the output knowledge was represented or because of the features defined for the machine learning algorithms:

- A part of the speech intentions taxonomies contained classes customized for the type of asynchronous communication or communicative goals (e.g. email [40, 79], tweets [91, 171], forum posts [10, 13, 64, 118, 166]).

- Some feature groups were specialized for specific domains, types of asynchronous communication, taxonomies or specific data. Domain-specific lexicons make solutions relevant only to some domains; email, forum and tweet-specific features challenge the transfer of solutions to other types of asynchronous communication; taxonomy-related expressions make solutions highly dependent on the output representation, which could be customized as previously discussed; author identity requires users to be identified and their data accessible; to some extent, n-grams and BoW features depend on the corpus content too.

Most solutions were not comprehensive, either because the speech intentions were not comprehensive or relations among speech intentions were not identified. A taxonomy was considered limited when it was not able to classify any utterance—for instance, Searle's speech act types are comprehensive with respect to this criterion, being exhaustive. Additionally, a taxonomy was considered limited when it did not provide a more detailed view than the main theoretical types of speech acts emerging from linguistics [184].

In the study of emails and tweets, expressives were scarcely covered, although directives were frequently studied as multiple individual sub-categories. Moreover, commissives were only included mostly in forum and email taxonomies. The scarce taxonomies that were comprehensive and corpus-independent were derived from dialogue acts [106, 150] and thus focused on describing the functional roles of utterances in dialogues, rather than showing utterance illocutions in the pragmatic sense. Dialogue acts permit to differentiate detailed types of answers and questions, but do not necessarily account for all illocutionary forces, being less suited for linguistics-focused research on conversations.

Then, with regard to the process aspect of conversations, few works tackled this by revealing transitions diagrams. While these diagrams were general, they captured mainly sequences and they were derived from adjacency pairs of speech acts. Nonetheless, more complex relations could also exist to represent processes [1]. Also, regarding adjacency pairs, immediate precedence is an assumption rather valid for synchronous conversations than for the asynchronous ones. Asynchronous conversations are threaded and have lengthy turns—thus no single turn sequence or speech act per turn.

The relevance of the proposed solutions was discussed and motivated by the majority of the related works. Finally, because of the high variance in output representations—21 different taxonomies were defined and very few works re-used previous taxonomies—the algorithmic effectiveness of the proposed techniques could not be compared.

2.4.2 Study Limitations

Several threats to validity were identified and mitigated in the current literature study.

A first threat to validity refers to the *constructs* used for data collection and data analysis and interpretation. The aim and research questions formulated in the beginning of this work guided the definition of constructs for these phases. Additionally, to mitigate this threat regarding data collection, comprehensive search terms were defined and continuously updated during the search process until satisfactory coverage was achieved. Every time the search query was refined, it was applied again for all repositories. To mitigate the construct validity in data analysis and interpretation, a conceptual framework was defined starting from the research questions and improved in multiple iterations by studying batches of selected publications. The final version proved exhaustive in mapping all the information reported in the complete set of articles.

A second threat to validity is that of *external* validity, which ensures that the selection of studies is complete under the scope of the current work. By mitigating this threat, it could be claimed that the conclusions reached from this study hold across the selected domain and defined scope. This threat was mitigated by defining a comprehensive search query, which was applied on most of the computer science digital libraries. However, not all the publications which appeared in the search hits could be retrieved, sometimes being constrained by the access of the university to the selected academic repositories. The abstracts of these publications were nonetheless considered in the data evaluation. If the articles seemed highly relevant, but could not be downloaded, the related publications of the authors were searched and retrieved, if available. Consequently, it can be claimed that very few highly relevant articles were not included in the present study. Moreover, a "snowball" search strategy complemented the search query approach.

The third category of threats to validity refers to *internal validity* and could appear either in data evaluation—where the relevant articles from the collected body of works are selected, or in data analysis and interpretation—where the content considered relevant for the analysis and for drawing conclusions is extracted. In order to frame the threats to internal validity in data evaluation, a clear list of inclusion and exclusion criteria was defined, considering the research questions and an initial set of discovered relevant publications. However, although the criteria were discussed with other researchers, the reliability of applying the inclusion/exclusion system by multiple researchers was not statistically tested. Then, to further mitigate the threats to internal validity in data analysis and interpretation, a clear and comprehensive conceptual framework was defined as a classification scheme with multiple facets and categories.

The knowledge and *conclusions* extracted from the application of the conceptual framework

were presented mainly as graphical tables with quantitative information and narrative summaries. Narrative reviews could be significantly affected by subjective judgments, although in the current work most facets of the conceptual framework encoded objective information in their dimensions. An additional threat to conclusion validity still needs to be further addressed: the data from which conclusions were drawn was encoded only by a researcher. Applying the conceptual framework concurrently by multiple researchers and statistically assessing the inter-rater agreement of the codes used per publication would have been a sounder approach.

2.4.3 Summary of Contributions

This chapter is a systematic effort to map and understand the asynchronous communication modeling with speech acts, a research area that has been awhile emerging. The previous literature studies, the most related to the current work, focused mainly on synchronous conversations and on dialogue acts [123, 161, 206, 217]. However, asynchronous communication has its own particularities and dialogue acts are rarely used to annotate this type of corpora. Consequently, this chapter is the first systematic literature survey of this topic to date.

Furthermore, in addition to the knowledge contribution, a conceptual framework to guide an extensive analysis and interpretation of the literature on the target topic was created. The previous literature studies covered only some facets, such as the machine learning approaches in [123, 217] or characteristics of dialogue taxonomies in [161, 206].

The relevance of this work is manifold:

- For the computer science academic community, this systematic survey provides an historical and thematic overview of what it has been done, how it has been done and why it has been done. Understanding existing theories and their relations can help researchers to theoretically ground their work. The methodological review can support researchers with formulating a sound rationale for the methodological strategy of their work. Additionally, gaps regarding both theory and methodology can be identified.

- For the linguistic community, the presented knowledge provides an overview of the taxonomies and the annotation strategies, that have emerged from computer science research. Most taxonomies have been created with reference to theoretical works, although adjustments have been also made empirically to accommodate types of corpora, goals or algorithmic limitations.

- Finally, for practitioners interested in developing intelligent software based on the interpretation of speech intentions in emails, tweets or any other type of posting, this review could be used as a guide that contains the most common steps to achieve this goal.

2.4.4 Directions for Further Research

Multiple directions of research are formulated following this literature survey and a part of them—1, 2, 3 and partly 4, will be tackled in the chapters to follow. These directions are listed below as a list of open questions grouped by problematic:

1. The taxonomies designed for the study of asynchronous communication are very diverse. Domain-specific taxonomies are relevant to the target domain and goals. From a theoretical perspective, discourse can be studied through the same framework—the speech act theory, which has the advantage of being general and thus applicable to any type of corpus or domain. However, this framework is also very high-level. Should the community turn towards more general theoretical approaches and, instead of making adjustments specific to the domain, formulate more detailed, but corpus-independent speech intentions? Is this achievable, in the first place, and still relevant across different domains and goals? Dialogue act taxonomies showed that this was possible for synchronous communication, but they have multiple limitations regarding their application to the asynchronous type.

2. The existence of relations among speech intentions and how these regularities predict conversation unfolding are matters of disagreement in pragmatics. These relations were presumed true in multiple computer science works. However, instead of relying on such still debatable assumptions, what if computer science endeavors support the building of evidence towards one of the sides of the controversy? Could computer science techniques enable the creation of a theory of asynchronous conversations? And would it be beneficial to enrich the relations among speech intentions with representations beyond adjacency-pairs and sequences?

3. Annotating a turn composed of multiple sentences with a single label is a limitation and simplification, theoretically speaking. However, also by following the dialogue theory, each utterance must be segmented in parts, such that each part can be associated to only one speech intention. Empirically, it has been noticed that this approach is troublesome for non-experts or especially for some new types of communication, such as tweets. Also, the manual annotation led to better results by choosing sentences or turns as units and by allowing multiple labels per unit. How these aspects should be tackled in order to consider both theoretical rigor and empirical relevance and ease of application?

4. While unsupervised methods significantly decrease the human effort required for ground-truth corpus preparation, the obtained results are still below those obtained with supervised learning. Where should the trade-off between algorithmic performance and human effort be set? Are semi-supervised algorithms a good compromise between the two for modeling conversations with speech intentions? Additionally, feature engineering is another area requiring significant human effort and sometimes is not transferable to other domains,

applications or corpora. Could simpler and domain- and corpus-independent features be defined for the effective discovery of speech intentions?

5. Three types of asynchronous communication were identified. However, apart from email, Twitter and forum conversations, are there any other types of asynchronous communication, which have very specific characteristics and call for different approaches to model them with speech intentions?

CHAPTER 3

MODELING PUBLIC TWEETS WITH SPEECH INTENTIONS

Epure E.V., Deneckere R., Salinesi C. (2017). Analyzing Perceived Intentions of Public Health-Related Communication on Twitter. In ten Teije A., Popow C., Holmes J., Sacchi L. (Eds), *Proceedings of 16th Conference on Artificial Intelligence in Medicine* (pp. 182-192). Vienna, Austria. Springer.

Contributions: E.E.V. designed and conducted the research and wrote the article. S.C and D.R. provided feedback on the research design and on the article.

The contributions of this chapter are:

- a taxonomy of corpus-independent, comprehensive speech intentions to model the Twitter public communication.
- an automatic method using supervised machine learning to annotate each tweet with multiple speech intentions based on discourse features only. The discourse features are designed to capture characteristics of speech intentions based only on discourse cues; hence, they are domain- and corpus-independent.

Solutions to automatically model tweets with speech intentions were presented in Chapter 2. Speech intentions are valuable to automatically analyze and summarize types of political discourse on Twitter [91]. Moreover, in the context of customer service on Twitter, speech intentions leverage a comparative analysis of strategies of communication between agents and customers and the automatic prediction of customer satisfaction or frustration [150]. Other long-term applications relying on the modeling of tweets with speech intentions were discussed too: the analysis of persuasive tactics and personal traits [165], the enhancement of search engines to allow queries

not only with topic keywords, but also with keywords expressing conversational goals such as advice, information or questions [165, 234], the comparison of written verbal behavior of different communities [234] and the automatic identification of topic shifts and of community influence on individual behavior [234].

In the related works, the speech intentions for tweets are represented through varied taxonomies. The taxonomies defined after applying unsupervised learning techniques, by manually analyzing the obtained clusters, contain classes that show tweet functional roles such as *reference broadcast*, *questions to followers* and *comments* [154, 171]. Additionally, taxonomies designed to capture conversational intentions relevant to specific applications [91, 165] or with a focus on dialogue acts [150] have been proposed too. Domain-independent taxonomies emerging from the speech act theory have been also adopted [213, 234, 235]. Although these latter taxonomies are general compared to most of the previous ones, they are still high-level with only directive speech acts being represented through finer-grained speech intentions.

However, in order to ensure a detailed modeling of the Twitter communication and enable the proposed applications, a more comprehensive taxonomy of speech intentions should be defined. While Oraby et al. [150] do propose such a solution, their classes—built on dialogue acts, are focused on task-oriented dialogues, giving priority to differentiating between functional types of answers and questions and less to differentiating intentions as illocutionary forces [12]. Moreover, in all the presented solutions, there are features heavily extracted from the corpus content [91, 150, 165, 171, 213, 234], specific to Twitter [213, 234, 235] or focused on hand-crafted lexicons [213, 234, 235] and taxonomy-related expressions [150]. These features limit the applications of these solutions to other types of asynchronous conversations.

Consequently, the objective of this chapter is to propose an improved approach to model Twitter communication. The research questions addressed in this chapter are: **RQ1**—*How to formalize conversations with comprehensive and corpus-independent speech intentions?* and **RQ2**—*How to automatically discover the proposed speech intentions from asynchronous conversations independently of the domain and corpus characteristics?*. Specifically, the investigation is focused on creating a method to automatically discover corpus-independent but finer-grained speech intentions from tweets. The speech intentions are considered perceived because they are interpreted from the stance of the readers. Public tweets are the scope for multiple reasons. This is the most common type of tweets accessible to researchers for investigating the Twitter communication. Compared to private tweets, the public ones most often convey specific conversational intentions—for instance, commissive speech acts are very rare, as shown by all related works. Thus, effort is put into defining common speech intentions for public tweets to allow for detailed investigations of this type of asynchronous communication.

To answer the research questions first, a corpus-independent and more comprehensive taxonomy of speech intentions for public tweets is proposed in Section 3.1. Second, an automatic method to annotate tweets with the defined representation is created (Section 3.2). Supervised

machine learning is chosen as it has proven to be effective in the past works. Discourse features, independent of domain, type of asynchronous communication, external lexicons and content are designed and put to test for the automatic discovery of speech intentions, in the supervised machine learning setup. Experiments are conducted to validate and evaluate each of these steps. Finally, a discussion on the external validity and other limitations is reported in Section 3.3. Also, the relevance of the proposed solution to medicine, the application domain selected for illustrating the current contribution, is discussed.

3.1 A Speech Intention Taxonomy for Public Tweets

Human behavior is intrinsically intentional as thoroughly discussed in philosophy [21] and psychology [5]. Though, behavior is not necessarily linked to only physical human acts, but also to language. Generally, people communicate with various intentions. Utterances are considered thus as acting through words while their leading intentions are called speech intentions [184]. For example, "This hospital has a nonstop emergency service" asserts the speaker's belief about the world, while "Could you please give me a painkiller?" requires the listener to act. As a reminder, Searle [184] proposed five types of speech acts:

1. *Assertive* is used to state true or false information about the state of affairs in the world.
2. *Expressive* is used to express the speaker's feelings towards the state of affairs in the world.
3. *Directive* implies the listener carrying out an action as a result of the speaker's utterance.
4. *Commissive* denotes the engagement of the speaker to a future course of action.
5. *Declarative* is the type of utterances, changing the world state such as firing someone.

The speech act theory emerged over time as a frequent adopted framework to model or predict language behavior from text.

Daniel Vanderveken [212] studied how speech acts are lexicalized in contemporary English vocabulary, proposing a detailed list of 300 verbs. Compared to Searle's speech act types, this classification has the advantage of being highly fine-grained. Although, as it is, 300 classes are too many to be applied in manual annotations and in automatic modeling solutions.

However, the classes are organized in five trees (see Appendix E). Each tree root corresponds to one of the Searle's speech act type. Each other intermediate or leaf node in the tree contains a speech intention considered a specialization of the speech intention found in the parent node. For instance, directive can be of type *request*, which can be further decomposed in *invitation* and *question*. Consequently, this organization, from a general to a highly detailed view, allows for the adoption of speech intentions at any level of detail in the defined range. Once the level decided, speech intentions could be further specialized or aggregated through their common parent.

Table 3.1 presents the main types of speech acts and the speech intentions adopted in the current work. The current taxonomy emerged from the Vanderveken's theoretical work [212], by selecting speech intentions from the immediate children of the tree roots. In this way, corpus-independent classes, more detailed than the main speech act types were selected. Additionally, their number was kept low to enable facile manual annotation—needed for the creation of the ground-truth corpus.

Then, the refinement of the taxonomy was based on manual corpus analysis. Thus, empirically it was noticed that *assertive* and *directive* account for most speech acts expressed in public tweets. These findings are not surprising considering that public tweets less often expose personal feelings (*expressive*) or personal goals (*commissive*). Such communication takes place rather privately. Moreover, *declarative* speech acts are very rare even in live conversations, being used in very particular settings. These observations are aligned with the previous works on modeling tweets with speech acts [91, 213, 234]. Additionally, a change was also brought to the speech intentions representing the children of the assertive tree root (see Appendix E): *guess* and *hypothesize* were merged in *hypothesize* because they were found very similar during the manual analysis of the public tweets.

Table 3.1: Identified speech intentions for the Twitter public communication.

Speech act type	**Intention**	**Tweet Example**
Assertive	**assert**	New study reveals autoimmune/inflammatory syndrome triggered by HPV vaccine URL
Assertive	**hypothesize**	Vitamin B1 may help relieve fatigue in Hashimoto's thyroid patients URL
Directive	**propose**	The gut microbiota and inflammatory bowel disease #microbiology #autoimmune URL #gutmicrobiota
Directive	**direct**	Are you a Cure Champion? Sign up for the Walk to Cure Psoriasis in a city near you...URL
Directive	**advise**	How To Avoid Holiday Autoimmune Flares URL
Directive	**warn**	Why you shouldn't be going from competition to competition URL #thyroid #metabolism #autoimmune

The instructions on how to associate the proposed speech intentions with tweets and the interpretation of each class are presented below:

- An *assert* is a tweet that conveys information clearly; it could be news or public personal declarations. Though an *assert* tweet may contain an url, the linked resource appears with the role to sustain or further detail the communicated message.

- A *hypothesize* is a tweet containing weak assertions, such as probable statements or hypothetical questions.

- A *propose* is a tweet that always references an external resource, which must be accessed in order to consume the message compared to the assertive speech acts. The *propose* tweets usually provide key or opinion words to give some clues about the resource content. Thus, *propose* is associated with weak attempts to make the reader access the message. A similar class, *directing to information*, is identified in [91] too.

- A *direct* is a strong attempt to make the reader act or reply following demands, requests, invitations, encouragements or questions—in contrast to *propose* tweets.

- *Advise* and *warn*, which are very similar, are directive speech acts suggesting actions to be followed or resources to be consumed. The suggestions, if followed, are supposed to be either good (*advise*) or bad (*warn*) for readers. The opinionated effect that this type of tweets might have differentiates them from *direct* tweets.

3.1.1 Experiments

The proposed taxonomy emerged from two linguistic works [184, 212]—as described in Section 3.1. The referenced linguistic works are relevant to study any type of communication. Thus, through design, the proposed speech intentions are corpus-independent. Moreover, compared to the related solutions to model public tweets with speech acts, the taxonomy has a higher level of granularity, as it proposes 4 directive speech intentions and 2 assertive speech intentions.

However, as the speech intentions are perceived, experiments are required to validate the fact that indeed, humans consistently interpret the same tweets through these classes. With other words, the question is: just by relying on the definition of the classes and personal intuition regarding communication, would two human annotators associate the same speech intention(s) with a tweet? Therefore, it could be claimed that the speech intentions taxonomy is *valid* as it was consistently applied by human annotators, showing alignment in the tweet perception. Also, this ensures experimental reproducibility and future application to other corpora.

In Chapter 1, it was stated that a taxonomy is comprehensive not only if it is fine-grained, but also if it is exhaustive when classifying utterances. Through empirical analysis of the corpus and aligned with the previous works, it was observed that commissive, expressive and declarative speech acts were very rare or even absent in this type of communication. Consequently, the defined classes are considered exhaustive in relation to public tweets. However, experimental validation of the taxonomy exhaustiveness is also necessary.

To validate the taxonomy, an experiment consisting in the manual annotation of a public tweet corpus by two human coders was conducted. Specifically, data was collected, the experimental setup was designed and the strategy to measure the validity was defined. Each of these steps is presented below.

Data Collection. Two sets of public tweets were collected via Twitter Streaming API . The first

one consists of 2714 tweets and the second one of 43153 tweets. The first set contains tweets fetched from the Twitter stream by applying the filter keyword "autoimmune". The second set was collected by filtering the Twitter stream by commonly used medical terms and jargon, lexicon proposed by Ridpath et al. [170] (see Figure 3.1).

Figure 3.1: Medical terms and jargon, lexicon proposed by Ridpath et al. [170].

The reason to restrict the domain to medicine is to analyze the public health-related communication on Twitter and illustrate the relevance to one application domain. Even though the communication on autoimmune diseases was initially targeted, the second set was included to ensure the diversity of the medical domain input tweets in evaluation.

Further, 600 tweets were randomly sampled from the first set and 500 tweets from the second set, making sure that messages with the same format were not duplicated and retweets were excluded. The selected 1100 tweets were used in both the taxonomy validation and the creation of the ground-truth corpus for the machine learning experiment (described in Section 3.2.1).

Manual Annotation Experiment. The selected corpus was tagged in parallel by two annotators: one the researcher who defined the taxonomy and the other seeing the taxonomy for the first time (also researcher in computer science). The taxonomy was only briefly introduced as the external annotator was expected to rather rely on verbal intuition when associating speech intentions with tweets.

Similar to all the previous works, the tweet was the unit of annotation, but contrary to most of these solutions, multiple labels were allowed in the annotation process (maximum 3). Also, the annotation was not supposed to be exhaustive—to find all the possible speech intentions associated with a tweet. The instructions to the external annotator stated that at least one label had to be chosen per tweet and more labels per tweet were advised in several cases: when the tweet had multiple sentences; when the tweet appeared to express multiple conversational goals;

or when one speech intention could not be clearly conveyed, hence choosing multiple labels was also a mechanism to handle the incertitude in annotation.

Exploratory annotations and past works [124, 150] showed that single-label annotation had some negative consequences: the process of choosing only one label among multiple often equally legitimate possibilities could be challenging, resulting sometimes in arbitrary choices by different annotators. Additionally, from a theoretical perspective, single-label classification of utterances is limiting too because of the co-presence of locutionary and illocutionary acts [12] and of indirect speech acts [184]. For these reasons, multi-label annotation was chosen in the current work.

Further, to experimentally evaluate if the taxonomy is indeed exhaustive, a class *other* was introduced to be used in the manual annotation, when none of the proposed speech intentions was a suitable choice.

Validity and Representativeness Measures. To measure the validity post-annotation, the Cohen's Kappa statistical test was used (see Equation 3.1). The Cohen's Kappa test considers the degree of agreement over all the pairs of speech intentions and the probability of agreement by chance. A score of the Cohen's Kappa test over 0.6 is considered a good result, specifically between 0.61 and 0.8 substantial and between 0.81 and 0.99 almost perfect [27]. The statistical significance of the results is computed with *z-test*.

$$\kappa = \frac{p_a - p_e}{1 - p_e} = 1 - \frac{1 - p_a}{1 - p_e}. \tag{3.1}$$

where:

- p_a represents the observed agreement among raters; it is equivalent to accuracy.
- p_e represents the chance agreement and is calculated with Equation 3.2. k is the code for a speech intention, N is the total number of tweets and n_{ki} is the number of times the rater i predicted the speech intention k.

$$p_e = \frac{1}{N^2} \sum_k n_{k1}, n_{k2}. \tag{3.2}$$

As multiple tags were allowed per unit, an alignment of the sets from the two annotators was necessary. The order of the labels was not informative. Consequently, the alignment consisted in first forming pairs of speech intentions from the intersection of the two label sets. Then, the elements, which were different between the two unit sets, were randomly paired. For instance, a first set {*assert*, *advise*} and a second set {*advise*, *assert*, *propose*} would be aligned as: {*assert-assert*, *advise-advise*, *propose*-empty_choice}. The processing of the raw tagged data for the Cohen's Kappa test is summarized in Algorithm 1.

Algorithm 1 Function to transform raw, labeled data for computing the interrater agreement.

Require:

$inputFile$—the input file containing the tweets and the labels assigned by the two raters;

$perSpeechAct$—a flag to mark if the transformations is made for computing the interrater agreement per speech intentions or per speech act types. If $false$, speech intentions are considered by default.

Ensure:

$Labels$—the list of pairs of labels($label_1$, $label_2$).

function TRANSFORM_FOR_INTERRATERAGREE($inputFile$, $perSpeechAct$)
    $Labels \leftarrow []$
    for each $row \in inputFile$ **do**
        $setRater1, setRater2 \leftarrow$ CREATE_LABEL_SETS(row, $perSpeechAct$)
        $agreeLabels \leftarrow$ GET_AGREEMENTS($setRater1$, $setRater1$)
        $agreePairs \leftarrow$ GET_AGREE_PAIRS($agreeLabels$)

        $disagreeLabels1 \leftarrow setRater1 \setminus agreements$
        $disagreeLabels2 \leftarrow setRater2 \setminus agreements$
        $disagreePairs \leftarrow$ GET_DISAGREE_PAIRS($disagreeLabels1$, $disagreeLabels2$) ▷ random pairing

        extend $Labels$ with $agreePairs$ and $disagreePairs$
    return $Labels$

To measure the exhaustiveness of the proposed taxonomy post-annotation, the frequency of the class *other* was computed:

$$E = \frac{N_{"other" \in labels}}{N}, \tag{3.3}$$

where $N_{"other" \in labels}$ is the number of tweets labeled with "other" by any of the annotators and N is the total number of tweets.

3.1.2 Results

RQ1 addresses the definition of a valid corpus-independent and comprehensive taxonomy of speech intentions for public tweets. Thus, a solution was proposed and, through experiments, the validity and exhaustiveness of the proposed speech intentions for Twitter public communication were assessed.

The condition of being exhaustive could be considered fulfilled through the artifact design. Relying on theory, the taxonomy is linguistically exhaustive for the written utterances belonging to the assertive and directive speech act types. Relying on corpus analysis, the taxonomy is mapped on the Twitter public communication. Also, a high frequency of the class *other* in the annotated corpus could experimentally denote a lack of exhaustiveness. However, *other* was used

only in 0.005 of the cases by both annotators, supporting thus the intuition behind the taxonomy design. *Other* appeared to replace, for instance, *expressive* speech acts such as greetings.

The Cohen's Kappa (κ) test was performed to assess the taxonomy *validity* and the results are presented in Table 3.2. The overall Cohen's Kappa score for the speech intention classes is 71.5% (z-score=40.9, p-value=0.001) being considered a *substantial* result. The scores for all speech intentions, apart from *advise* and *other*, are substantial too. The overall Cohen's Kappa score for the main speech act types is 78.3% (z=28.1, p=0.001). The fact that this score is higher than the previous one shows that mismatches occurred sometimes between speech intentions belonging to the same speech act type.

Table 3.2: κ scores for speech act types and intentions. At least substantial results are underlined.

Assertive:		**Directive**:				**Other**:
0.8		0.8				0.41
assert	**hypothesize**	**propose**	**advise**	**direct**	**warn**	**other**
0.77	0.78	0.70	0.49	0.68	0.76	0.35

In order to get more insights into the common disagreements, the confusion matrix is projected in Table 3.3. Most frequently, *advise* tweets are interpreted as *propose* and *direct* tweets, while *other* class is misclassified as *direct*. Few cases of misclassifying *direct* as *propose* are also encountered. Between the assertive and directive speech acts, the most common confusion includes *assert* and *propose* tweets. Often these tweets contain news, the difference between the two types being the extent to which the presented message is self-contained or an url has to be followed in order to understand the message. However, it appears that this definition may be subjective and leads to divergent interpretations sometimes.

Table 3.3: Confusion matrix showing the agreements (the values on the diagonal) and the disagreements between two annotators in using the proposed speech intentions.

	assert	**hypothesize**	**propose**	**advise**	**warn**	**direct**	**other**
assert	**423**	9	62	9	1	6	6
hypothesize	3	**39**	3	0	0	2	1
propose	29	0	**310**	24	3	8	3
advise	0	0	7	**27**	2	0	0
warn	2	0	2	0	**13**	0	2
direct	5	3	13	10	0	**121**	18
other	0	0	1	0	0	0	**9**

The related works manually annotating tweets with speech acts report similar or lower overall results: moderate and substantial Kappa scores [91, 150, 213, 215]. Some of the classes used in the related works overlap on the classes of the current taxonomy. Therefore, a comparison of the inter-rater agreement scores for these classes is possible. *Assert* tweets are moderately discovered in [150] and almost perfectly in [91]—the current experiments show results in be-

tween the two, namely substantial. The manual classification of *questions* and *requests* led to moderate and substantial Kappa scores in [150]. The equivalents of *direct* and *propose* tweets in [91]—requesting action and directing to information, respectively—obtain similar inter-rater agreements scores to the current work (around 0.70). In conclusion, it could be argued that the obtained results prove the validity of the taxonomy and are aligned with previous literature results.

3.2 Automatically Annotating Tweets with Speech Intentions

The automatic annotation of tweets with speech intentions is formulated as a supervised machine learning task. Given a ground-truth corpus and a defined set of features, four types of classifiers (*Logistic Regression*, *Linear SVM*, *Random Forest* and *Multinomial Naive Bayes* [17]) were explored to annotate tweets with the defined speech intentions and their performance is compared. Both single-label and multi-label setups were assessed. A detailed presentation of the experimental setup is provided in Section 3.2.3.

Before the machine learning experiment, several steps were required first. The ground-truth corpus was created based on the dataset manually annotated in the experiments for the taxonomy validation. This step is detailed in Section 3.2.1. Then, features of speech intentions were defined and extracted from tweets for being used in the automatic classification. In order to normalize the text before feature extraction, text processing and natural language processing were applied. Tweet NLP [151] was used for part-of-speech tagging of the original tweets. Further, the tweet text was conversed to lowercase characters and lemmatisation was applied.

Two types of features were defined: *Content* and *Discourse* features. *Content* features consist of standard text mining features: *BagOfWords* and *OpinionKeywords*. *BagOfWords* is a dictionary of tokens and their frequencies. The tokens were extracted from the preprocessed Twitter corpus, before the classification. *OpinionKeywords* include the frequencies and ratios of negative and positive opinion words from a predefined lexicon [129]: *freqPositiveWords*, *ratioPositiveWords*, *freqNegativeWords*, *ratioNegativeWords*.

Discourse features are novel and defined by considering linguistic means to express speech intentions, without relying on the overall corpus content, corpus domain, characteristics of the Twitter communication or external lexicons. They are described in Section 3.2.2.

3.2.1 Ground-truth Corpus

Multiple decisions were made for the creation of the ground-truth corpus based on the the dataset annotated in the previous experiment. All the tweets with classes upon which the two annotators agreed were automatically included in the final corpus. Then, the strategy to handle disagreements was devised after a detailed manual analysis of the common mismatches.

The class that was more frequently confused with other classes than being consistently applied by the annotators is *advise*. Specifically, it was confused with *propose* in 60% of the disagreements; with *direct* in 19% of the disagreements; with assert in 17% of the disagreements; and with warn in 4% of the disagreements (computed based on the confusion matrix in Table 3.3). By further manual analysis, it was concluded that a tweet was an advice either because it redirected the reader to an external webpage that actually contained an advice or the tweet was in itself an advice. The first case corresponded to mismatches involving *propose* (e.g. "4 Steps to Heal Leaky Gut and Autoimmune Disease URL"), while the second case to mismatches regarding *direct* (e.g. "#TipTuesday: Vitamin D deficiency is linked to autoimmune diseases. Add mushrooms to Thanksgiving!") and *assert* (e.g. "The quicker you receive treatment, the better your chances for a good recovery from #Stroke are"). It might be surprising that an advice is stated as an assertion but this is an example of indirect speech acts.

Nevertheless, what emerged was that an *advise* tweet could ultimately be related to also *propose*, *direct* or *assert*. Considering its low Cohen's Kappa score, the validity of the definition of this class was questioned. Therefore, the decision to transform the *advise* tweets in their secondary speech intentions for the creation of the ground-truth corpus was made. Nonetheless, the class is still maintained in the taxonomy, but an update of its definition and more manual annotation experiments are required in order to validate it empirically.

Moreover, the *warn* tweets were also transformed because of insufficient instances—2% of the tweets were annotated as warnings–and for maintaining consistency with *advise*. The final speech intentions for *advise* and *warn* were set by the author, by considering the label sets from the manual annotations too. Additionally, a discussion took place between the original annotators for the rest of the disagreements in the corpus and for the more challenging cases regarding *advise* and *warn* and a collective final decision was made.

In the final corpus[1], 89% of the tweets are single-label and the rest of them have associated 2-3 classes of speech intentions. The most common pairs of co-occurring speech intentions to interpret a tweet are: *direct* and *assert* (5.1% of the tweets); *assert* and *propose* (3.8% of the tweets); *direct* and *propose* (2.4% of the tweets).

3.2.2 Discourse Features of Speech Intentions

A person could use Twitter to address utterances to the community or to specific users. In live conversations, the meaning of utterances are implicitly understood from the content and form of utterances, but also from non-verbal cues such as speaker's gestures and voice characteristics. Though not as rich as this case, the tweets as a form of written communication have also characteristics that convey speech intentions. In the current work, these cues are named *Discourse features* and are summarized in Table 3.4. The rationale behind these features is further provided.

[1]http://tinyurl.com/hk9t83y

PronominalFeatures are frequencies of various pronouns, considering also their objective, subjective, possessive and reflexive forms. The first person singular (*freq1stPersonSg*) is chosen because it could be a sign of personal declarations specific to *assert* tweets (e.g. "I'm for vaccination."). Similarly, the third person (*freq3rdPerson*) could be linked to *assert* tweets when reporting facts and stories or for descriptions. By contrary, the first person plural form (*freq1stPersonPl*) or second person (*freq2ndPerson*) could be cues of *direct* tweets (e.g. "We must vaccinate our kids! You should too!!").

Table 3.4: Proposed discourse features for discovering speech intentions.

Group	Name	Description
PronominalFeatures	*freq1stPersonSg*	frequency of *I*, including its other forms.
	freq3rdPerson	frequency of third-person pronominal forms.
	freq1stPersonPl	frequency of *we*, including its other forms.
	freq2ndPerson	frequency of *you*, including its other forms.
PunctuationFeatures	*hasExclamation*	presence of exclamation marks.
	hasQuestion	presence of question marks.
	hasEllipsis	presence of ellipsis—at least two sequent dots.
	hasColon	presence of colons.
	hasQuotes	presence of quotes—all possible forms considered.
5W1HFeatures	*freq5W*	frequency of *what*, *when*, *where*, *why*, *who*.
	freq1H	frequency of *how*.
EmoticonCues	*freqEmoticons*	frequency of ASCII and Unicode emoticons.
TitleCues	*freqTitleWords*	most tweet words have the first letter uppercase.
VerbFeatures	*hasVerb*	presence of verbs, except present and past participles; negative forms considered too.
	hasImperative	presence of imperative verbs; negation included.
	hasCan	presence of *can* and its short or negative forms.
	hasCould	presence of *could* and its short or negative forms.
	hasMust	presence of *must* and its short or negative forms.
	hasMay	presence of *may* and its short or negative forms.
	hasMight	presence of *might* and its short or negative forms.
	hasShould	presence of *should* and its short or negative forms.
SyntaticConstructs	*hasNV*	presence of noun-verb constructs.
	hasNN	presence of noun-noun constructs.
	hasAN	presence of adjective-noun constructs.
	hasNComma	presence of nouns followed by punctuation.
	hasPN	presence of nouns preceded by a prepositon, postposition or subordinating conjunction.
	hasDN	presence of nouns preceded by determiners.
	hasVN	presence of verbs followed by nouns.
	hasCommaU	presence of punctuation marks folllowed by urls.
	hasNP	presence of nouns followed by a prepositon, postposition or subordinating conjunction.

Further, *PunctuationFeatures* are indicators of the discourse functions: the presence of ex-

clamation (*hasExclamation*), interrogation (*hasQuestion*), ellipsis (*hasEllipsis*; for omissions, hesitations); colon (*hasColon*; for titles, explanations) or quotes (*hasQuotes*).

5W1HFeatures complements the punctuations for revealing discourse functions (e.g. questions). The frequency of "what", "when", "where", "why", "who" (*freq5W*) is separated from that of "how" (*freq1H*) as *how* is also used for advice or proposals (e.g. "How I Quited Smoking URL").

The frequency of *EmoticonCues* (*freqEmoticons*) could be inversely correlated with impersonal reporting; hence, possibly linked to *assert* and *propose* tweets. Both ASCII and Unicode emoticons were checked.

TitleCues (*freqTitleWords*) seems to be an often marker for news, being thus a potential discriminator for the *assert* and *propose* tweets (e.g: "New Releases in Science URL"). It is computed as a relative frequency: $N_{UpperCaseWords}/N_{words}$.

VerbFeatures (*hasVerb*, present and past participles not considered) could be an indicator of weak directives when false. These tweets often lack the subject-predicate form. Features regarding the modals are also defined: (*hasCan*, *hasCould*, *hasMust*, *hasMay*, *hasMight*, *hasShould*). Modals could convey various intentions: *hypothesize*, *advise*, *direct* etc. For all verb features, negative forms of the verbs were identified too.

A feature related to verb moods is also defined: *hasImperative*, which is frequent in requests and demands, thus being an indicator of *direct* tweets. To identify imperatives, a rule-based algorithm was created, having as input the tweet tagged with part-of-speech (POS) information (see Algorithm 2). The algorithm does not cover all possible cases; thus, it may miss imperative utterances sometimes—for instance, when formed with modals. However, its precisions is high. Experiments on the dataset regarding the autoimmune diseases revealed a precision of 84%, the false positives being caused mainly by errors in parsing. For instance, in the tweet "Sleep Hydration Tip to Improve Detoxification URL #thyroid #metabolism #autoimmune", "Sleep" is tagged as verb, which makes the algorithm to falsely yield true for this instance. Other example of false negatives are:

- "HOPE FOR HEALING .... We all know someone who is affected by an autoimmune disorder .... allergies , asthma"—"HOPE" here can be interpreted as both verb and noun;
- "#Autoimmune_neuromuscular_disorders : treatment optimization . Live webcast in 5 minutes (12:30 pm ET) . URL"—"Live" here is wrongly identified by the parser to be a verb.

SyntaticConstructs features are created with the goal of incorporating syntactic characteristics of the discourse. The assumption is that tweets with the same speech intention might share similar discourse form. The POS tags were analyzed in order to dynamically discover representative POS-related features. The followed steps are further reported:

1. The corpus was annotated with POS tags using a dedicated tweet NLP parser [151].

Algorithm 2 Function to identify if a tweet contains verbs having an imperative mood.

Require:

$tweetTokens$—a list with the tweet words without any preprocessing.

$tweetPOSes$—a list of POS tags corresponding to the tweet words; $tweetPOSes[i]$ is the POS tag of $tweetTokens[i]$. The codes for POS tags are from Tweet NLP [151].

Ensure:

$hasImperative$—returns 1 if at least one imperative verb is found and 0 otherwise.

function HAS_IMPERATIVE($tweetTokens$, $tweetPOSes$)
    verify the correctness of the input, throw exception if incorrect
    $hasImperative \leftarrow 0$
    $N \leftarrow tweetTokens.length$

    $modalV \leftarrow$ LOAD_MODALS(environ.Language) ▷ negative forms of modals included too
    $auxV \leftarrow$ LOAD_AUXILIARY(environ.Language) ▷ e.g. does, haven't etc.
    $infinitiveV \leftarrow$ LOAD_DICT_VERBS(environ.Language) ▷ a dictionary with verbs

    for each $t \in tweetTokens$ **do**
        $ind \leftarrow$ index of t in $tweetTokens$
        $pos \leftarrow tweetPOSes[ind]$

        if $pos \neq$ 'V' **then**
            continue
        if t ends in "ing" or $t \in modalV$ **then**
            continue ▷ a verb cannot be imperative with an "-ing" form; modals can be used for imperative, but not always—they are currently excluded to avoid false positives.

        ▷ an imperative verb appears at the beginning of a sentence, However, tweets can be ungrammatical and may miss the correct punctuation. So, other cases should be tackled too. The final condition is: either the verb starts the tweet or follows after a punctuation mark (','), an url ('U'), an emoticon ('E') or a number ('$').
        if $ind = 0$ or $tweetPOSes[ind-1] \in \{',', 'U', 'E', '\$'\}$ **then**
            if $(t \in auxV)$ **then**
                if $ind + 1 < N$ and $tweetTokens[ind+1] \in infinitiveV$ **then**
                    $hasImperative \leftarrow 1$ ▷ e.g. "Don't buy this..", "Do check again!";
                    break ▷ Stop the search, an imperative already found
                else if $t \in$ {"have", "do"} **then**
                    $hasImperative \leftarrow 1$ ▷ e.g. "Do as I told you!"
                    break ▷ Stop the search, an imperative already found
            else if $t \in infinitiveV$ **then** ▷ check if the verb has an infinitive form
                $hasImperative \leftarrow 1$
                break ▷ Stop the search, an imperative already found
    return $hasImperative$

2. The normalized frequencies of each two consecutive POS tags were computed from the POS-tagged corpus.

3. The pairs of consecutive POS-tags with a score of at least *0.5* per intention were selected.

The final syntactic features are: *hasNV*, *hasNN*, *hasAN*, *hasNComma*, *hasPN*, *hasDN*, *hasVN*, *hasCommaU*, *hasNP*. The encoding of these features is: *N* nouns; *V* verbs, *A* adjectives, *Comma* punctuation, *P* pre-, post-position or subordinating conjunctions, *D* determiners and *U* urls. Compared to the original output of the parser [148], two changes were made. *N* incorporates also proper nouns (the symbol ^) and pronouns (the symbol *O*). Moreover, the keyword *Comma* replaces the symbol ",", used for punctuation by the tweet parser.

3.2.3 Experiments

In order to answer the second research question (**RQ2**) and assess the effectiveness of the discourse features for speech intention classification, which are domain-independent compared to the content ones, a machine learning experiment was set-up. As previously mentioned, Logistic Regression, Linear SVM, Random Forest and Multinomial Naive Bayes were the selected classifiers. These classifiers were chosen as they were frequently used in the related works. Also, as the ground-truth corpus contained individual tweets and not conversations, structural learners—other common machine learning classifiers in the related works, were not an option. These four classifiers are briefly introduced below.

A classification task can be formalized as: given a set of M instances $\{\boldsymbol{x}^{(m)}\}$, where $m = 1,2,...,M$, with the associated target classes $c^{(m)} \in C$ (C is a finite set of categories), the objective is to predict for a new instance $\boldsymbol{x}$ its corresponding class $c \in C$. Each instance is represented as a vector $\boldsymbol{x} = (x_1, x_2, ..., x_N)$, where each x_i is a numerical value specific to a defined feature and N is the total number of features.

Logistic Regression. Let's assume a binary classification task, where the goal is to predict if a new instance belongs to a class c or not. The Logistic Regression classifier learns a function $h(\boldsymbol{x}) \in \{0,1\}$ from the training instances [17]. A value of $h(\boldsymbol{x}) = 1$ denotes that $\boldsymbol{x}$ belongs to the class c, while $h(\boldsymbol{x}) = 0$ means that $\boldsymbol{x}$ cannot be associated with c. In Logistic regression, the simplest target function can be a linear combination of the input variables:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = w_0 + w_1 x_1 + ... + w_N x_N \tag{3.4}$$

This function is also a linear combination of the parameters $w_0, w_1, ..., w_N$, which needs to be estimated. Equation 3.4 could be further extended as a linear combination of non-linear functions of the input as follows:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = w_0 + \sum_{j=1}^{N} w_j \phi_j(\boldsymbol{x}) \tag{3.5}$$

where $\phi_j(\boldsymbol{x})$ are called basis functions and could have other definitions apart from $\phi_j(\boldsymbol{x}) = \boldsymbol{x}$ as in Equation 3.4. w_0 is known as a bias parameter; in order to simplify Equation 3.5, a basis function corresponding to w_0 can be defined as $\phi_0(\boldsymbol{x}) = 1$ and integrated. The target function becomes:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{j=0}^{N} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}} \phi(\boldsymbol{x}) \tag{3.6}$$

being a model with $N+1$ parameters that need to be estimated.

Further, as the classifier has to generate for a new instance either a value of 0 or 1, then the output $h_{\boldsymbol{w}}(\boldsymbol{x})$ is constrained to this interval: $0 \leq h_{\boldsymbol{w}}(\boldsymbol{x}) \leq 1$. A common approach is to consider $h_{\boldsymbol{w}}(\boldsymbol{x}) = g(\boldsymbol{w}^T \phi(\boldsymbol{x}))$, where g is the sigmoid function:

$$g(\boldsymbol{w}^T \phi(\boldsymbol{x})) = \frac{1}{1 + exp(-g(\boldsymbol{w}^T \phi(\boldsymbol{x})))} \tag{3.7}$$

The sigmoid function crosses 0.5 at the origin. When used with Logistic regression, a yielded value $h_{\boldsymbol{w}}(\boldsymbol{x}) \leq 0.5$ is interpreted as 0 ($\boldsymbol{x}$ does not have associated the class c), while $h_{\boldsymbol{w}}(\boldsymbol{x}) \geq 0.5$ is interpreted as 1 ($\boldsymbol{x}$ has associated the class c).

In order to estimate the parameters $\boldsymbol{w}$, Logistic regression uses the following overall cost function, which has to be minimized:

$$J(w_0, w_1, ..., w_N) = \frac{1}{2M} \sum_{i=1}^{M} (h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2 \tag{3.8}$$

where $y^{(i)} \in \{0, 1\}$ is the true label of the instance $\boldsymbol{x}^{(i)}$, as given in the training set. If the function $Cost(h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}), y^{(i)}) = \frac{1}{2}(h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2$ is defined to represent the cost of an individual instance, similar as in Equation 3.8, then the overall cost function could be re-written:

$$J(w_0, w_1, ..., w_N) = \frac{1}{M} \sum_{i=1}^{M} Cost(h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}), y^{(i)})) \tag{3.9}$$

It is challenging to use this cost function for parameter estimation because it is non-convex. This means that using the gradient descent method on this type of function risks to converge to a local minimum, instead of the global one. To address this issue, a new cost function for an individual instance, which is convex—hence, it has a global minimum, is defined as follows:

$$Cost(h_{\boldsymbol{w}}(\boldsymbol{x}), y)) = \begin{cases} -\log h_{\boldsymbol{w}}(\boldsymbol{x}), & \text{if } y = 1. \\ -\log(1 - h_{\boldsymbol{w}}(\boldsymbol{x})), & \text{if } y = 0.. \end{cases} \tag{3.10}$$

Finally the gradient descent method is applied on Equation 3.9 with the cost function defined as in Equation 3.10. Specifically, the parameters are repeatedly updated using a learning rate α:

$$w_j = w_j - \alpha \frac{\partial J(w_0, w_1, ..., w_N)}{\partial w_j}, \text{ simultaneously update all } w_j \tag{3.11}$$

The number of iterations in gradient descent varies, depending on the problem. The rule is to continue to simultaneously update all w_j until the overall value of $J(\boldsymbol{w})$ does not seem to decrease much more in the current iteration compared to the previous one. The learning rate α could be set experimentally by plotting the value of $J(\boldsymbol{w})$ versus the number of iterations and observe how quickly it converges to a global minimum (e.g. possible values 0.001, 0.01, 0.1, 0.5 etc.).

The generalization of this classifier type from binary classification to the classification with multiple classes is called multi-class Logistic Regression. A common approach to handle this situation is to train a Logistic Regression classifier for each unique class. Hence, the multi-class classification task is reduced to having multiple binary classification tasks. This strategy is also known as "one-versus-all" or "one-versus-the-rest".

Linear Support Vector Machines (SVM). Considering also a binary classification task, the main idea of SVM is to find a hyper-plan that maximizes the margin between the two types of instances: those belonging to the class c ($y^{(i)} = 1$) and those that do not belong to the class c ($y^{(i)} = -1$).

Let $\boldsymbol{w}$ be a vector, of any length, perpendicular on this hyper-plan. The decision rules for associating a certain instance $\boldsymbol{x}$ with the class c can be formulated:

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x}^{(i)} + b \geq 1, & \text{for } i,\ y^{(i)} = 1. \\ \boldsymbol{w}^T\boldsymbol{x}^{(i)} + b \leq -1, & \text{for } i,\ y^{(i)} = -1. \end{cases} \tag{3.12}$$

which can be further combined in a single decision rule:

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1, \forall i \tag{3.13}$$

The support vectors are those instances with the following properties: $\boldsymbol{w}^T\boldsymbol{x}_+ + b = 1$ for positive instances ($y = 1$) and $\boldsymbol{w}^T\boldsymbol{x}_- + b = -1$ for the negative instances ($y = -1$). The margin between the two hyper-plans to which these types of instances belong is then:

$$margin = (\boldsymbol{x}_+ - \boldsymbol{x}_-)\frac{\boldsymbol{w}}{||\boldsymbol{w}||} \tag{3.14}$$

where $||\boldsymbol{w}||$ is the magnitude of the vector $\boldsymbol{w}$, thus the ration $\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$ gives the unit vector. Replacing the support vectors in Equation 3.14, the margin becomes:

$$margin = \frac{2}{||\boldsymbol{w}||} \tag{3.15}$$

Maximizing the margin means minimizing $||\boldsymbol{w}||$ under the condition $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1,\ \forall i$. To find the extreme of a function under constraints, the Lagrangian multipliers can be used. Writing Equation 3.15 in the Lagrangian formulation results in a new expression for which the extreme has to be found:

$$L = \frac{2}{||\boldsymbol{w}||} - \sum_i \alpha^{(i)}[y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) - 1], \alpha^{(i)} \geq 0 \tag{3.16}$$

Further, Equation 3.16 has to be minimized with regard to $\boldsymbol{w}$ and b:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_i \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)} = 0 \Rightarrow \boldsymbol{w} = \sum_i \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)} \tag{3.17}$$

$$\frac{\partial L}{\partial b} = \sum_i \alpha^{(i)} y^{(i)} = 0 \tag{3.18}$$

By replacing Equation 3.17 in Equation 3.16 and doing the simplifications, the function that must be minimized is rewritten as:

$$L = \sum_i \alpha^{(i)} - \frac{1}{2} \sum_{i,j} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \boldsymbol{x}^{(j)T} \boldsymbol{x}^{(i)}, \text{ with } \alpha^{(i)} \geq 0 \text{ and } \sum_i \alpha^{(i)} y^{(i)} = 0 \tag{3.19}$$

This problem is quadratic and could be solved with multiple methods, including the gradient descent, which was briefly introduced before in the presentation of Logistic Regression. Multiple measures are taken for the case when the positive instances cannot be linearly separating, such as relaxing the constraints in Equation 3.12. Versions of SVM to handle the multi-class classification are also available or an one-versus-all approach could be used.

Random Forest. Random Forest is an ensemble classification method based on multiple individual Decision Tree classifiers. Each Decision Tree is trained separately on a sub-sample of the training instances. An approach to generate each subsample is to randomly select instances from the initial dataset with-replacement. The "with-replacement" strategy entails that once an instance is selected, it is not excluded from the initial dataset. Thus, this specific instance could be selected again for being part of the same sub-sample. In order to make a final prediction, the Random Forest classifier averages the prediction results of all individual Decision Trees. Random Forests are often preferred to simple Decision Trees because they better address over-fitting and appear to give more accurate predictions.

Further, the Decision Tree classifier is introduced. The main idea is to infer a tree from the training data, which will be used after to predict the class c for new instances $\boldsymbol{x}$. The nodes of this tree, except from the leaves, represent features of the instances. Each branch from a certain node correspond to values, or value ranges, of the feature represented by the node. The leaf nodes contain the classes to be predicted. Once a leaf node is reached when assessing the attributes of an instance $\boldsymbol{x}$, that class is associated with the instance.

Multiple strategies to construct the trees exist. All of them rely on the idea that at each step in the iterative learning of the tree, the best feature for classifying instances, from the ones not selected yet, should be picked to become a node. Then, its possible branches are generated. The best feature is the most informative among the existing ones. In order to quantitatively evaluate how informative a certain features is, first the entropy of a dataset is defined:

$$E = \sum_{c \in C} -p(c) \log_2 p(c) \tag{3.20}$$

where $p(c)$ is the probability of the class c, derived from the training dataset. The entropy can be also computed for a class c, conditioned by the existence of a feature f with the value v, as $E_v = \sum_{c \in C} -p(c|v)\log_2 p(c|v)$. Based on the entropy, the information gain of a feature f, showing how informative the feature is, is defined:

$$IG_f = E - \sum_v p(v)E_v = -\sum_{c \in C} p(c)\log_2 p(c) + \sum_v p(v) \sum_{c \in C} p(c|v)\log_2 p(c|v) \tag{3.21}$$

where v are all possible values of feature f and $p(v)$ is the probability of the feature value v in the dataset. The feature with the highest information gain is chosen at each step in the tree construction. Multiple algorithms for inducing a decision tree based on the information gain measure exist: ID3, C4.5, C5 [155].

There is also another strategy to construct the decision trees based on the GINI index and GINI split. The GINI index is defined for a feature value v as:

$$GINI(v) = 1 - \sum_{c \in C} [p(c|v)]^2 \tag{3.22}$$

where $p(c|v)$ is the relative frequency of class c, given the feature value v. If the node corresponding to a feature f is split in V partitions, the GINI split is computed as:

$$GINI_{split}(f) = \sum_{v \in V} \frac{n_v}{n} GINI(v) \tag{3.23}$$

where n_v is the number of instances at child corresponding to v and n is the number of instances at node f. In order to measure how informative a feature is, the one with the lowest GINI split is chosen at each iteration. This approach is implemented by the CART algorithm [155].

Multinomial Naive Bayes. A Naive Bayes classifier is based on Bayes Theorem, which states that the probability of a class c, given an instance $\boldsymbol{x}$ is:

$$P(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c)P(c)}{P(\boldsymbol{x})} \tag{3.24}$$

where:

- $P(\boldsymbol{x}|c)$ is the probability of an instance $\boldsymbol{x}$, given the class c (posterior probability of $\boldsymbol{x}$)
- $P(\boldsymbol{x})$ is the probability of $\boldsymbol{x}$ (prior probability of $\boldsymbol{x}$)
- $P(c)$ is the probability of c (prior probability of c)

If there are multiple classes, for a new instance $\boldsymbol{x}$, the classifier will predict the class c if and only if $P(c|\boldsymbol{x}) > P(c_i|\boldsymbol{x})$, $\forall i, c_i \neq c$. With other words, the class c is chosen for which $P(c|\boldsymbol{x})$ is maximized. Considering the theorem in 3.24, maximizing this is equivalent with maximizing $P(\boldsymbol{x}|c)P(c)$ as $P(\boldsymbol{x})$ is constant. Estimating $P(\boldsymbol{x}|c)$ is computationally expensive. However, an

assumption is introduced, namely that the features probabilities conditioned by the class c, $P(x_i|c)$, are independent:

$$P(\boldsymbol{x}|c) = P(x_1, x_2, ..., x_N|c) = P(x_1|c)P(x_2|c)...P(x_N|c) \tag{3.25}$$

These probabilities, $P(x_1|c), P(x_2|c), ..., P(x_N|c)$ and $P(c)$, can be now computed from the observations in the training dataset. Multinomial Naive Bayes classifier, compared to the Naive Bayes one, assumes that the conditional probabilities regarding the features have a multinomial distribution and it is often used in text classification [17]. Another remark is that the Naive Bayes classifier and its variations are inherently multi-class.

The *scikit-sklearn* implementations of the classifiers [155] were used with default parameters. Multinomial Naive Bayes was selected as its library implementation allows continuous features too. Naive Bayes and Random Forest inherently handle multi-label classification, while Logistic Regression and Linear SVM in an on-versus-all approach—hence, implementing binary classification for each speech intention.

Various configurations of features were evaluated: discourse features only (*DiscF*), content features only (*ContF*), all features (*AllF*). Also, the features were normalized beforehand to the 0-1 range. Finally, assuming no interaction in models, the most predictive features for each speech intention were found, by analyzing the weights of the best classifier, trained on standardized feature values [75].

The metrics for performance evaluation of the single-label corpus are *Precision* (P), *Recall* (R) and *F1-score* (F1), which are defined as follows:

$$P = \frac{T_p}{T_p + F_p}; R = \frac{T_p}{T_p + T_n}; F1 = 2\frac{PR}{P + R} \tag{3.26}$$

where T_p is the number of true positive instances, T_n is the number of true negative instances and F_p is the number of instances falsely predicted as positive by the classifier. These are computed as *macro scores* (with their average weighted by the support of each intention) averaged over the 10 folds. Same evaluation decisions were applied within each fold in cross-validation.

The *Hamming loss* was used to evaluate the accurate prediction of the *multi-label* corpus:

$$HammingLoss = \frac{1}{M|C|}\sum_{c \in C}\sum_{\boldsymbol{x}} T_{\boldsymbol{x},c} \oplus P_{\boldsymbol{x},c} \tag{3.27}$$

where M is the number of instances, $|C|$ is the number of classes, $T_{\boldsymbol{x},c}$ is a boolean value showing if the true labels of the instance $\boldsymbol{x}$ contains the class c, $P_{\boldsymbol{x},c}$ is a boolean value showing if the predicted labels of the instance $\boldsymbol{x}$ contains the class c, $\oplus$ is the XOR operation. The Hamming loss shows the fraction of incorrectly predicted labels to the total number of labels. For this, the single-label corpus was the train set and the multi-label corpus the test set.

Finally, the obtained results were tested for statistical significance with the *two-tailed t-test*.

3.2.4 Results

The overall results obtained by the classifiers with various configurations of features are summarized in Table 3.5. The results for each speech intention class are presented in Table 3.6.

Table 3.5: Overall results of the classification experiment using various feature sets and only the single-labeled tweets. The measures are macro-weighted. The best results are in bold.

	Logistic Regression			**Random Forest**			**Linear SVM**		
metric	DiscF	ContF	AllF	DiscF	ContF	AllF	DiscF	ContF	AllF
P	0.76	0.65	0.78	0.72	0.57	0.70	0.78	0.63	0.78
R	0.76	0.65	0.78	0.72	0.59	0.70	0.78	0.64	0.78
F1	0.76	0.62	**0.78**	0.72	0.55	0.68	**0.78**	0.62	**0.78**

Statistically, Logistic Regression and Linear SVM are comparable ($p > 0.5$, two-tailed t-test) and both outperform Random Forest. Multinomial Naive Bayes is not reported as yielded results similar to Random Forest for *DiscF* and *ContF* and worse for *AllF*. It can be noticed that the *DiscF* group systemically leads to similar results as *AllF* (apart from *hypothesize* with Logistic Regression) and to improved results over *ContF*—up to 5 times better with Random Forest.

For multi-label classification, the minimum Hamming loss scores are obtained when using *DiscF* with Logistic Regression (0.287) and Linear SVM (0.264).

Vosoughi and Roy [213] also reported that the Logistic Regression classifier led to the best results: a weighted average F1-score of 0.70. A similar F1-score was obtained by Zhang et al. [234] using a SVM classifier with linear kernel, similar to the current work.

Further, an analysis of the feature predictive power in relation to each speech intention is made based on the feature weights estimated for Linear SVM:

- For *assert*, the most predictive discourse features are related to reporting information or making public declarations: *hasVerb*, *hasNV*, *hasQuotes*, *freq3rdPerson*, *freq1stPersonPl*, *freq1stPersonSg*. The most representative content features appear to be related to news and particularly to the scientific ones ("implications", "future", "bacteria", "influence").

- For *hypothesize*, both discourse and content features reveal the importance of modal verbs ("might", "may" and "could"). In addition, the most important discourse feature is *hasQuestion*.

- For *direct*, *hasImperative* is the most discriminatory feature. This also emerges from top predictive content features that contain multiple verbs ("know", "learn", "read", "check"). Then, features linked to requests, encouragements and questions follow in importance in the classification of *direct*: for discourse features–*hasQuestion*, *hasExclamation*, *freq2ndPerson*, *freq1H*, *freq5W*; for content features –"what", "please", "let".

- For *propose*, the top most important discourse features are related to news summaries (*freqTitleWords*, *freq5W*, *hasColon*) and advice or warnings (*freq1H*, *hasMust*, *hasShould*,

Table 3.6: The results of the classification experiment using various feature sets for each class and only the single-labeled tweets. The support of each class is presented in the first column, under the label. The measures are macro-weighted. Best results for each speech intention and for each type of classifier are in bold. Best results overall for each speech intention are underlined.

		Logistic Regression			**Random Forest**			**Linear SVM**		
	metric	DiscF	ContF	AllF	DiscF	ContF	AllF	DiscF	ContF	AllF
assert sup.=416	*P*	0.75	0.67	0.77	0.70	0.57	0.66	0.77	0.69	0.79
	R	0.83	0.71	0.84	0.79	0.64	0.86	0.82	0.66	0.83
	F1	0.79	0.69	**0.80**	**0.74**	0.60	**0.74**	**0.80**	0.68	**0.81**
hypothesize sup.=49	*P*	0.62	0.62	0.77	0.65	0.50	0.60	0.71	0.55	0.75
	R	0.41	0.10	0.47	0.49	0.06	0.12	0.76	0.33	0.55
	F1	0.49	0.18	**0.58**	**0.56**	0.11	0.20	**0.73**	0.41	0.64
direct sup.=119	*P*	0.71	0.66	0.78	0.70	0.59	0.67	0.75	0.50	0.77
	R	0.79	0.18	0.75	0.70	0.11	0.35	0.82	0.27	0.79
	F1	0.75	0.28	**0.76**	**0.70**	0.18	0.46	**0.78**	0.35	**0.78**
propose sup.=391	*P*	0.80	0.63	0.79	0.74	0.57	0.75	0.80	0.62	0.78
	R	0.72	0.79	0.75	0.66	0.71	0.70	0.72	0.76	0.76
	F1	0.76	0.70	**0.77**	0.70	0.64	**0.72**	0.76	0.68	**0.77**

freq2ndPerson). These features are correlated with the content ones, which incorporate interrogation cues ("why", "how"), impersonal scientific words ("epidemic", "lupus") and suggestion-related words ("step", "recipes", "good").

Finally, the POS-related features appear highly predictive for all speech intentions, in particular for *assert*, *direct* and *hypothesize*.

The obtained results show that an automatic method to satisfactorily identify the proposed speech intentions from tweets could be created based on supervised machine learning. Moreover, the defined discourse features are highly effective in the classification. The proposed discourse features improved significantly the discovery of speech intentions over the content ones. Additionally, they are also correlated with the top most important content features. However, the discourse features benefit of being much fewer (30 vs. 4608) and corpus- and domain-independent, allowing thus reproducibility on other English corpora of asynchronous communication. Thus, an answer to **RQ2**, addressing the automatic modeling of public tweets with speech intentions, independently of domain and corpus, was provided.

3.3 Conclusion

An approach to analyze the perceived speech intentions in the Twitter public communication was proposed. The speech intentions taxonomy for public tweets is corpus-independent and finer-grained compared to the related works. Additionally, its automatic discovery proved effective, with F1-scores between 0.73 and 0.8, using Linear SVM with discourse features only. The solution is corpus-independent, as the speech intentions are general linguistic classes, empirically adjusted to the Twitter public communication. Also, the discourse features are discourse-related, being suitable for any corpus type or domain. Finally, the most predictive content features were often related to the discourse ones and to the speech intentions, acting as a positive feedback for the proposed taxonomy and for the designed discourse features.

The relevance of the solution to medicine, the selected application domain, has not been illustrated yet. The next section discusses how the current solution could support the creation of high-level, behavioral models of health information consumers and disseminators, relevant to studies in narrative medicine and health information seeking and dissemination.

3.3.1 Relevance of Designed Approach to Medicine

The Internet has nurtured a highly available and accessible environment for disseminating health information. Nowadays, the massive dissemination and consumption of health information have also been impacted by the tremendous adoption of social media [157]. This has led to the creation of communities, fostered by the interpersonal interactions, with acknowledged advantages such as anonymity and 24-hour availability [39].

Social media allows to disseminate health information, express beliefs and feelings about health matters and react to existing content. One's beliefs are built on or altered by the information to which he or she is exposed [39]. This could be further reflected in new behavioral intentions and eventually new behaviors [5]. Social media has even a stronger influence on people because of social norms: the attitudes and behaviors of a community towards health matters are transparent in this online environment. This aspect could be exploited to promote healthy behavior such as quitting smoking. However, inaccurate information, available to a very large and generally vulnerable target—people impacted directly or indirectly by chronic diseases, could become harmful and have mass consequences [39].

Therefore, there is an interest in health care to analyze social media communication as a form of behaving in order to be able to identify interactions that could lead to behavioral changes and to identify messages reporting new behaviors. For this, suitable automatic methods are necessary. The link between the solution proposed in this chapter and modeling or predicting behavior is as follows.

1. The perceived speech intentions are related to the written verbal behavioral of health information disseminators. A large-scale analysis focused on the perceived speech intentions of public, health-related tweets can be further performed. Information dissemination strategies in different communities (e.g. communities for various diseases) can be compared.

2. The current solution allows to create automatic techniques to measure the impact of various message formulations on health information consumers, similar to message framing [148]. The public's reactions to the same, but differently formulated messages can be analyzed (e.g. which formulation is the most re-tweeted?). Also, is an individual more likely to read an online, health-related article if its link is tweeted as a hypothetical question or as an information?

3. As collective narratives concerning health care can be strong drivers for behaviors, their automatic discovery can be enhanced with an intentional perspective [99]. Automatic techniques already exist for identifying components of the crowd narratives, such as topics or events [162]. However, disseminator disposition towards the presented events is necessary for a thorough narrative representation [142]. The disseminator disposition can be conceptualized through perceived speech intentions.

Finally, relevant knowledge for medicine could be already discovered from the analyzed corpus. The most popular intentions are *assert* (46%) and *propose* (41%). Consequently, it appears that Twitter is publicly used for information dissemination. However, it is quite interesting that these strategies are different, half of the tweet messages being self-standing (*assert*), while the other half requiring redirection and consumption of an external resource (*propose*). The ratio of *direct* tweets (18%) shows also a significant expected reaction from consumers, by replying or following advice, warnings, requests or invitations. Similarly, Godea *et al.* [77] identify tweet purposes in

health care (advertising, informational, positive or negative opinions). However, the focus of the current solution is on pragmatic speech intentions, being thus a more general approach.

3.3.2 Study Limitations

The current study includes several research decisions that may have had an impact on validity. Internal, external, construct and conclusion validity aspects are further discussed.

Potential threats to the *internal validity* concern both the manual and automatic annotations. Manually assigning labels by humans can be biased, although this is a common approach in text classification. To reduce bias and quantitatively assess the taxonomy validity, a pair of expert and external human annotators labeled the same dataset. Then, Cohen's Kappa, a widely spread statistics for inter-rater agreement, was computed for each speech intention. Further, the ground-truth corpus was obtained by intersecting the two label sets for each unit. The goal was to obtain a correctly, but not exhaustively labeled corpus. To mitigate the internal validity threats emerging from the ground-truth corpus, the problematic speech intentions (*advise* and *warn*), were transformed in their secondary intentions. Also, when the intersection contained no labels, the expert and the external annotator decided on at least one final label.

Further, another threat regarding the automatic annotation was related to the frequency of each class in the corpus. While some classes were popular, others were not sufficiently represented. Thus, the results were reported as macro-weighted to account for each class support. Another potential threat to internal validity is overfitting, mitigated by using a cross validation setup However, parameter tuning was not implemented, using default configurations. The experimental setup was detailed for replication in Section 3.2.3.

Regarding the *external validity*, first, more manual annotation experiments must be conducted with other human coders and datasets too. Second, Twitter private communication should be researched too, including speech intentions from the *expressive* and *commissive* categories. Another potential threat to external validity emerged from the limited number of labeled instances in the ground-truth corpus, all belonging to the same domain—medicine. However, to minimize this threat, discourse features independent of the domain and corpus were defined and proved effective in the automatic annotation.

Additional threats to external validity originated from the use of specific libraries implementing supervised classification and natural language processing (scikit-learn [155] and nltk [132], respectively) and specific lexicon for sentiment analysis [129]. Nevertheless, these Python libraries have been frequently used in machine learning experiments and proved their robustness in a variety of applications. Also, their accessibility facilitates the replication of the presented experiments. Finally, a small but representative selection of machine learning algorithms based on different modeling mechanisms was used, similar to other related works.

The threats to *construct validity* in the current study could stem from the conceptualization of speech intentions and the evaluation of the predictive models. To mitigate the former threat,

linguistic theories and empirical analysis have been considered in designing the taxonomy. Then, an experiment aimed at validating the application of the proposed classes with two human annotators was conducted. The results revealed that the *advise* tweets could not be reliably identified. This may be because of the class definition. However, more experiments with multiple external human annotators should be conducted in order to enforce the taxonomy validation and check if some confusions systemically take place. Furthermore, with regard to the evaluation of the classifiers, metrics commonly used in machine learning experiments were employed, such as precision, recall and F1-score.

Conclusion validity has been supported by the use of statistical tests measuring the confidence of the conclusions (p-value established to 0.05, common significance level), selected according to the properties of the data. The conclusions in the machine learning experiment were reached by controlling the validation bias through cross-validation and separating train-test sets and by using suitable performance measures (macro-weighted scores).

3.3.3 Directions for Further Research

The proposed taxonomy, although corpus-independent and more comprehensive than the classes found in the related works, could not be entirely validated—the *advise* class did not reach satisfactory inter-rater agreement. Therefore, the taxonomy should be updated to tackle this issue, but also to become comprehensive for any type of asynchronous communication. This latter point implies that fine-grained classes specific to *commissive*, *expressive* and *declarative* speech acts should be included too. Additionally, for an extensive validation, a manual annotation experiment with multiple external human annotators and datasets is necessary.

The discourse features led to very promising results for the classes present in the ground-truth corpus (*assert*, *hypothesize*, *propose*, *direct*). Nonetheless, are these features still sufficient to discover other types of speech intentions? Or should other discourse characteristics be identified as cues of speech intentions? Moreover, would the algorithmic performance be improved through feature selection, parameter tuning or sampling techniques? Then, some frequent pairs of speech intentions were identified in the corpora. Given this aspect, would strategies for multi-label classification lead to improved results?

Finally, the process view of conversations, namely the relations among speech intentions, could not be tackled. This was a limitation of the dataset, which contained only individual tweets and not conversations. Therefore, other datasets with conversations must be considered and a solution to reveal processes of interrelated speech intentions from annotated conversations still needs to be designed in order to meet all the solution goals and properties, as defined in Chapter 1. Also, the relevance of the proposed approach to medicine asks for stronger validation.

CHAPTER 4

A Structural Taxonomy of Speech Intentions

Compagno, D., Epure, E.V., Deneckere-Lebas, R., Salinesi, C. (2018). Exploring digital conversation corpora with process mining. *Corpus Pragmatics, 2*(2) 193-215.

Contributions: E.E.V. and C.D. designed the speech intentions taxonomy and wrote the article, E.E.V. designed the manual annotation experiment, C.D. and E.E.V deployed the experiment, E.E.V. analyzed the collected data, S.C and D.R. provided feedback on the experimental design and on the article.

The chapter presents the following contribution:

- A corpus-independent and comprehensive speech intention taxonomy to model any type of communication. The taxonomy is an extension of the previous one to model public tweets, presented in Chapter 3. This time, 18 classes are defined to conceptualize speech intentions and a structure based on oppositional traits of the proposed speech intentions is created to guide annotators trough the manual application of the taxonomy on corpora. Extensive experiments with external annotators prove the validity of most speech intentions and that the use of the taxonomy in manual annotations by non-experts is feasible.

While Searle's classification [184] has multiple advantages, including relative simplicity in interpretation and annotation of data and high popularity in the computer science community, only five types of intentions are too broad to capture detailed insights into verbal behavior [72]. More granular taxonomies also exist in linguistics [212, 226]. Vanderveken [212] exemplifies a detailed list of speech acts through 300 English verbs, organized in trees (see Appendix E). The root of each tree is one of Searle's speech act types [184] and the children at each level are specialized speech intentions of their parent node [212]. Although the level of detail is very high,

the exact use of these trees in empirical work is problematic, despite their relative organization. The fine differences between intentions and the large number of classes would make the manual annotation very challenging and slow. However, as shown in Chapter 3, these trees of speech acts could support the definition of new taxonomies for empirical works, with a level of granularity established on the tree structure—as desired by the originators or allowed by the context.

Works in computer science have also proposed intention taxonomies. Often ad-hoc classes of conversation units have been created to fit a specific domain (e.g. email management [40, 178], online tutoring [10, 13, 166]) or a given corpus (e.g. Twitter messages [234], emails [79]). Other works used Searle's types [184] as a basis from which they deviated in various ways: for example, they focused on some speech act types only [163, 234], covered a speech act type only partially [100, 165] or introduced super-classes related to multiple speech act types [14, 64]. General taxonomies for dialogue analysis have been created too [110, 196]. Although aligned with linguistic theories, these classes capture functional aspects of dialogues, such as types of answers and questions, rather than pragmatic speech acts. Also, their use in practice is targeted mainly at experts, as the class number is large and particularities of spoken dialogues are addressed.

Leveraging the advantages and disadvantages of the theoretical works and empirical computer science solutions, a new taxonomy of speech intentions was proposed to analyze the Twitter public communication in Chapter 3. Corpus-independent and more detailed than in the related works, the taxonomy proved representative for public tweets. Nonetheless, the adopted speech intentions are not exhaustive compared to the speech act theory because commissive, expressive and declarative types are missing. Also, its validation was preparatory, but not sufficient. Consequently, the goal of the current chapter is to tackle the limitations of the previous related works—stated in the first two paragraphs, and of the solution provided in the first iteration—stated in the previous paragraph, and, in a second iteration, to propose an improved representation of written communication. Specifically, the first proposed research question is revisited: *How to formalize conversations with comprehensive and corpus-independent speech intentions and process relations?* (**RQ1**).

To this end, a structured, corpus-independent and fine-grained classification of speech intentions is defined and presented in Section 4.1. Additionally, its grounding on the speech act theory and conversation analysis is exposed. Further, an experiment consisting in manually annotating a corpus of digital, asynchronous conversations is designed and conducted—presented in Section 4.2. The annotation is sentence-by-sentence, multi-label and contextual. The results are discussed in Section 4.3 and the study limitations and directions for future work in Section 4.4.

4.1 Taxonomy Design

This section starts with presenting the taxonomy design rationale. Specifically, it motivates the use of certain theoretical linguistic works [12, 184, 212], the first steps taken towards the creation

of the taxonomy and their motivation. Then, in the second part, the taxonomy consisting in 18 speech intentions, organized according to some oppositional traits within each speech act type, is presented and discussed. Additionally, concrete examples are provided for each class.

4.1.1 Design Rationale

When modeling communication and verbal interactions, speech acts are often used to discretize utterances and enable further analyses in many domains including pragmatics, conversation analysis, computer science and computational linguistics. Searle's original classification of speech acts [184] differentiates five main groups of speech acts: assertive, directive, commissive, expressive and declarative. These classes can conceptualize the "overt aim" of the actions produced by speakers or writers with utterances [197]. The overt aim is the pragmatic meaning, attributed to language production in real-world contexts, emerging from the mutual understanding of communicative intentions between speakers or writers and listeners or readers.

Searle's typology [184] is a milestone in pragmatics and has multiple merits. It emerges from philosophical research, but retains its simplicity in interpretation and application. It allows to identify discrete classes of speech acts; hence, it is frequently used in empirical studies relying on corpus annotation. Moreover, these classes are general, thus corpus-independent and not linked to any particular domain, discourse genre or other specific cases of language production. In the course of time, Searle's classification has become highly popular and has been used by many research communities, including computer science.

The five main types of speech acts—assertive, expressive, directive, commissive and declarative, are also the basis of the current design. However, five general classes are not sufficient for a detailed analysis of conversations and of communication, in general [72]. The trees defined by Vanderveken [212] to lexicalize speech acts with common English verbs appeared as a promising direction to refine the speech act types, as presented in Chapter 3. The organization of the trees such that each node, except for the leaves, represents a speech intention more general than the ones contained by all its children, motivated the inclusion of this work in the current taxonomy design. For instance, commissive speech acts are all children of the root *commit* and a path can be traced in the tree to more specific commissive speech acts such as *accept* and *promise*.

The trees corresponding to the assertive, directive, commissive and expressive speech acts have 21 classes as direct children of the roots. These general, but more detailed speech intentions could ensure a higher level of detail for the analysis of discourse, while their number would make the manual annotation still manageable. The current taxonomy is derived from the 21 first-level children of the Vanderveken's trees [212]. However, as the focus of the taxonomy is to be used in empirical works, which frequently imply manual annotations, a further improvement is brought to the taxonomy in order to make it more understandable and applicable. Specifically, a layer of organization is added to these general classes, consisting in several oppositional traits that are defined through conceptual and empirical analyses. The final classification has 18 speech

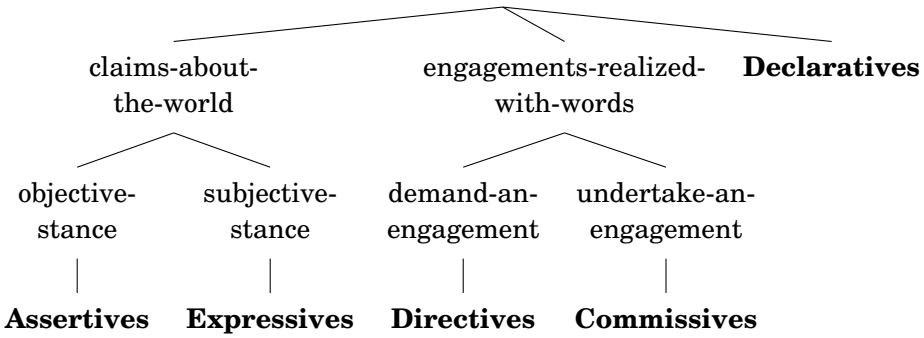


Figure 4.1: The proposed structure for the five main types of speech acts.

intentions, as some of the 21 speech acts were merged or split[1]. These interventions took place in multiple iterations of discussions and manually annotating corpora with a linguistic researcher. The taxonomy evolved in these iterations, the changes being triggered also by the inter-rater agreement scores obtained in the manual annotations of the selected datasets by its originators.

4.1.2 Deriving a Structural Speech Intentions Taxonomy

Two larger categories can be identified from Searle's five types of speech acts (see Figure 4.1). The first category includes assertives and expressives, while the second category consists of commissives and directives. The main difference between the two is that the speech intentions in the former are claims about the world, while the speech intentions in the latter realize social engagements with words. For instance, this difference is emphasized through the following utterances: "The salt is on the table." (claim about the world) and "Could you pass me the salt?" (engagement with words). Declaratives, the remaining speech acts, are considered a special type.

The first group, claims about the world, is further divided in two sub-categories that reflect different degrees of subjectivity used in the enunciation of utterances: the objective-stance sub-category composed of assertive speech acts and the subjective-stance sub-category composed of expressive speech acts. These relations are nonetheless different than the Searle's view on assertive and expressive speech acts [185]. Searle considers that expressive speech acts are intentions conceived as internal mental entities, while assertive speech acts are instead related to states of affairs "out there" in the world.

Searle's interpretation of the speech act theory and of Austin's work [12] is not the only one possible [182]. In fact, the proposed precise ontological differentiation between assertive and expressive may not be needed for empirical linguistic research. Consequently, in the current work, these classes are both considered as claims about the world—although this interpretation is slightly forced in the example "I feel sad", and their stance is judged only by the presence or

[1]Merged speech act classes: *hypothesize* and *guess*; *congratulate* and *rejoice*; *condole* and *complain*; *subscribe*, *undertake*, *pledge*, *threaten* and *engage*; *warn* and *advise*; *propose* and *direct*. Splitted speech act classes: *request* and *require*; *assert* and *sustain*.

absence of emotional cues that the speaker or writer conveys in the message. The preliminary annotation of real-world written corpora with the proposed taxonomy also revealed the presence of a continuum between assertive and expressive speech acts, instead of a clear divide. Their identification appears linked to how words are read and how punctuation or other enunciation cues are interpreted. Several examples are provided further: "Two people died and several were injured yesterday night." is an assertive; "What a horrifying event!" is an expressive; "Sadly, two people died and several were injured yesterday night." can be each of the two.

The second group is also composed of two sub-categories— commissive and directive speech acts, which complement each other. *Direct* utterances ask for engagements and their acceptance or refusal can be expressed through *commit* utterances. The relation between these sub-categories is based only on being complementary, compared to the previous sub-categories considered in a continuum. Also, ideally, in a conversation each *direct* utterance has associated a *commit* utterance and the other way around. Some examples are: "Would you prepare the presentation for tomorrow?" (requesting an engagement), "No, I will for Tuesday." (undertaking an engagement).

For the analysis of written asynchronous communication, declarative speech acts are considered a special case, very rare in these settings. Conceptually, declarative and assertive speech acts could be positioned in a continuum relation too, where an *assert* utterance introduces facts that are already true in the speaker's world, while a *declare* utterance brings about facts in the world only after the utterance enunciation. However, as for now, the current taxonomy is created for the annotation of asynchronous communication, aiming at simplicity, and declarative speech acts are very rare in this type of corpora, the oppositional trait between assertive and declarative is not included. A well-known example of a *declare* utterance is "I declare you husband and wife".

Further, the description of the 18 classes and their structure with oppositional traits is provided and summarized in Figure 4.2. Assertive speech acts can be independent from or linked to previous utterances in conversation. In the latter group, two intentions in speech are defined: *agree* and *disagree*. If independent from previous speech acts, assertives could be further divided based on the strength of utterances. A writer can simply *assert* something—with a medium strength, state something confidently by sustaining it with arguments and examples (*sustain*, with a strong strength), state something while showing limitations of the uttered beliefs through probability, possibility and doubt (*guess*, with weak strength). Examples for these classes are:

- *agree*: "Yes, you are completely right."
- *disagree*: "I don't entirely agree with what you said."
- *assert*: "The weather is warm today."
- *sustain*: "I am sure it will rain today because the weather app I use is always reliable."
- *guess*: "It may be true, but there is room for doubt."

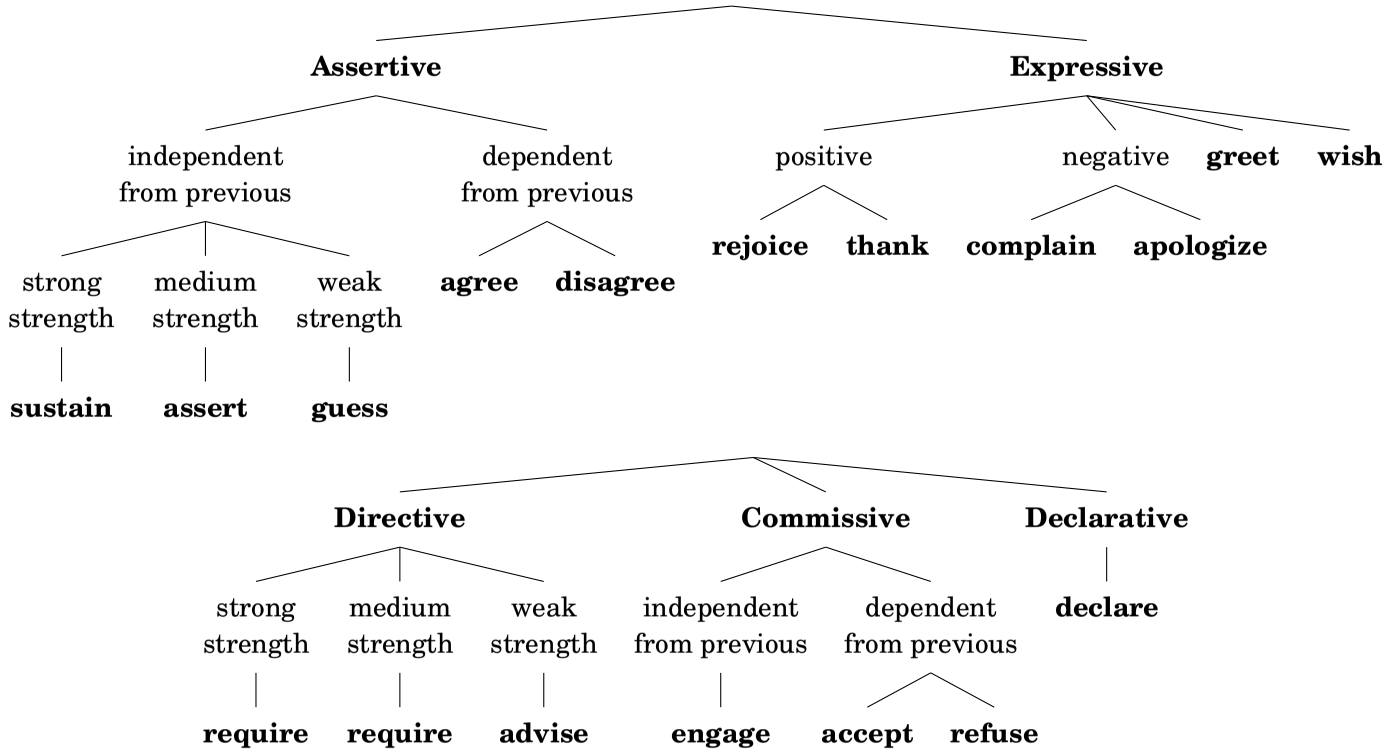


Figure 4.2: The speech intention taxonomy with speech act types in the first-level nodes, speech intentions in the leaves and oppositional traits in the intermediary nodes [41].

The expressive speech acts are differentiated between general and special types. The general types could be mapped on utterances expressing either positive feelings and opinions (*rejoice*) or negative feelings and opinions (*complain*). An utterance which contains neither positive, nor negative emotions is expected to be associated to assertive—let's notice again the continuum relation between assertive and expressive. Sarcasm can also influence the separation into positive and negative expressive as, for instance, a positive utterance may be in fact negative, being sarcastically used. However, as long as the context is taken in consideration in manual annotations, these cases could be correctly discovered and the proposed classes are still applicable.

The special types of expressive speech acts identified for analyzing asynchronous communication are *greet*—contains any form of salutation, *thank*—to state someone's merit for something and express a positive opinion about it, *apologize*—to state one's fault for something and express a negative opinion about it, and *wish*—to express positive or negative emotional states about the future. The fact that *wish* refers to the future is the main difference with *rejoice* and *complain*. Examples of expressive utterances are further presented:

- *rejoice*: "That's great"!
- *complain*: "This is much worst than I expected..."
- *greet*: "Bye bye!"
- *thank*: "I really appreciate what you did for me."
- *apologize*: "I'm sorry to hear about that."
- *wish* "I hope you will feel well soon."

The directive speech acts are divided in three groups, depending on their strength with which they attempt to engage the listeners or readers. Strong-strength directives are represented by the class *require*, medium-strength directives by the class *request* and weak-strength directives by the class *suggest*. Requirements could be orders and injunctions. Requests include also questions and invitations. Suggestions could be either recommendations, advice or warnings; hence, they are facts or suggested actions supposed to be good or bad for listeners or readers. As emphasized by Vanderveken [212], in order to discriminate between *request* and *require*, an analysis of the expected reply could be made. Specifically, the content of *require* utterances could be obeyed or disobeyed, while *request* utterances can be replied to with acceptance or refusal. Compared to these, suggestions expect a much weaker response, if any, such as thanking. Examples of directive speech intentions are:

- *require*: "Finish this by tomorrow!"
- *request*: "Can you help me with understanding this paragraph?"

- *suggest*: "You should avoid stress."

Similar to the assertives, the commissive speech acts can be dependent or independent from previous utterances. If dependent, these previous utterances are associated with directive speech acts and their corresponding commissive could be *accept* or *refuse*. The strength of the complementary directives does not influence the type of commissive responses and only these two classes cover responses to all types of directives. If a commissive utterance is standalone, then the class *engage* is proposed. Examples of commissive classes are:

- *accept*: "Fine, I'll do it"
- *refuse*: "Sorry, I can't do it."
- *engage*: "I will attend this course from Standford next summer."

As already discussed, declarative speech acts are too rare in asynchronous communication. Therefore, an internal categorization is not provided.

The proposed classes and oppositional traits may appear to some extent fuzzy. However, this broad structure and the class descriptions are intentionally indistinct as annotators are expected to also rely on their intuitive interpretation of language production when classifying utterances. The taxonomy is also supported by examples, considered typical for each class. The overall approach has been designed with the aim to make it as applicable as possible to corpora. Defining an extended or exhaustive list of features of each class would be impracticable and even impossible, given that each class may contain different heterogeneous utterances.

4.2 Experiments

The goal of the experiments was to validate the taxonomy. The taxonomy is considered valid if it is *consistently* applied by human annotators, showing thus an alignment in the perception and also implying experimental reproducibility and its potential for future applications to other corpora. The taxonomy consistency was quantitatively evaluated by applying it to a corpus of digital conversations [207]. A dataset of *Reddit* threads was used[2]. The taxonomy consistency does not imply that the taxonomy must be used to exhaustively annotate conversations. Instead, the consistency implies that the classes, when used by multiple annotators for the same unit, are aligned.

The taxonomy used for the experiments went through three changes:

1. A class *other* was added to be used when other classes did not fit the unit annotation.

[2]Reddit is a website organizing threads of discussion in sections called "subreddits" according to thematics.

2. *Require* and *request* were merged in the class *direct* as empirical evidence, collected in the preliminary evaluation with the originators only, showed that *require* utterances, such as orders, were absent in Reddit conversations.

3. For same reasons as before, the class *declare* was excluded, giving the possibility to cover declarative instances, if present, with the class *other*.

The proposed speech intentions taxonomy is corpus-independent, fine-grained and, compared to the previous attempt in Chapter 3, designed to be complete. The completeness of the classification, meaning that the taxonomy allows the interpretation of utterances with all types of speech acts, is build upon the completeness shown by the linguistic works referred for its design [184, 212]. However, this could be also proved experimentally by observing how frequent annotators resort to using the class *other*—a high frequency denoting incompleteness.

Further, the steps of the validation experiment, data collection, the manual annotation setup and deployment and the assessment of the classification consistency, are discussed.

Data Collection. The dataset consists of 21 anonymized conversations, collected from a subreddit about autoimmune diseases[3]. The complete subreddit was crawled, but only 21 conversations were randomly and automatically selected[4]. This specific "subreddit" was chosen first for maintaining the same line of proving the relevance for at least one application domain, the medicine, as in the previous iteration presented in Chapter 3. Second, an exploratory analysis of these conversations revealed that most speech intentions could be used to characterize this type of conversation utterances. When sharing and discussing their experiences with diseases, Reddit users may ask for or give advice, thank for it, request specific information, express positive or negative feelings about personal experiences and other situations, provide information and so on.

Each conversation starts with a first post and continues with comments to the initial post or to the other comments. The selected conversations have multiple threads, thus their individual structure is a tree. A conversation was composed in average of 8 turns and 45 sentences. The posts had between 1 and 13 sentences and the comments between 1 and 7 sentences.

Manual Annotation Experiment. Conversations of different lengths from the selected subreddit were automatically parsed into grammatical sentences by considering the punctuation and randomly assigned to pairs of annotators. There were two types of annotators: 2 experts—the originators of the taxonomy, and 10 external annotators—subjects who did not have previous knowledge about the taxonomy. The external annotators were mostly researchers, with varied

[3]The subreddit used in the current work is: `https://www.reddit.com/r/lupus/`

[4]The reason why only 21 conversations were selected is related to the number of annotators who agreed to take part in the experiment. For each group of annotators, a csv (comma-separated values) file with 3 conversations was generated using a Python script, such that each conversation among those not already selected was randomly chosen and all 3 conversations together did no contain more than 200 sentences (annotation units). Additionally, 3 datasets were annotated by two groups, instead of one.

academic ranks, from PhD students to professors, and with varied backgrounds—computer science, finance, sociology of communication and anthropology. English was not the native language of any of them, but they all used it on a very frequent basis in their work.

Each external annotator was expected to annotate sentences belonging to 3 conversations, selected such that each dataset had between 100 and 200 conversation units. Thus, compared to the previous iteration where entire tweets were annotated, this time the unit of annotation was not the turn, but each turn *sentence*. A set of instructions introducing the task and the taxonomy, together with multiple examples were provided to each external annotator by email. Appendix F presents all the resources exchanged during the deployment of the experiment . The experts were paired with external annotators and proceeded with the same task independently.

The external annotators were asked to identify the aim(s) of sentences by considering the sentence linguistic form, their intuition, but also the sentence context—the entire turn and conversation. It was also mentioned that more than one speech intention could be associated with an unit—and up to three, in case if the unit appeared to have more than a single aim or the annotator had doubts about multiple choices. Thus, this was a *multi-label* manual annotation.

The reasons for designing a multi-label manual annotation, which takes in consideration the context, emerged from exploratory pre-tests. Exploratory annotations revealed that the sentence context could improve the classification by being realized faster, perceived as easier and leading to more precise results—in terms of inter-rater agreement. Also, as already mentioned in Chapter 3, single-label annotation proved challenging because the process of choosing only one class among multiple, often equally legitimate possibilities could be cumbersome, long and resulting sometimes in arbitrary choices. An example of a sentence realizing more than one speech intention is: "Thanks, I will certainly try them!" (*thank*, *accept*).

Additionally, the fuzzy definition of the proposed structural classification intentionally opens the possibility of having multiple classes associated with an utterance. It could be argued that the multi-label classification is a more suitable setup for manually annotating real-world corpora for theoretical reasons too. When establishing the aim of an utterance in linguistics, a differentiation is made between locutionary and illocutionary acts [12] and between direct and indirect speech acts [184]. However, in the instructions on how to associate the classes to an utterance, references to these perspectives on speech acts were not provided, aiming to keep the procedure simple for the taxonomy application by non-expert, external annotators.

Assessing Classification Consistency. To measure the classification consistency, the inter-rater reliability was computed for each class and speech act type, considering the annotations provided by all teams composed of external annotators and experts. The procedure was similar to that described in Chapter 3 for assessing the consistency in the tweet annotation experiment.

As multiple labels per unit were allowed in the manual annotation, several transformations

of the label sets were needed[5]. For each unit, the labels proposed by different annotators were aligned such that the matching classes were paired, despite the order in which they were provided. Thus, after this transformation, between one and three pairs of speech intentions were produced for each conversational unit. For example, if an expert labeled an utterance with {*assert*, *complain*, empty_choice} and the external annotator with {*complain*, *sustain*, empty_choice}, the resulting pairs were *assert-sustain* and *complain-complain*. Cases when one annotator used more labels than the other were also present. For example, if an expert labeled the utterance with {*assert*, *request*, empty_choice}, and the external annotator with {*assert*, empty_choice, empty_choice}, it was considered that both annotators agreed on the use of *assert*. However, the pair *request*-empty_choice was not taken into account in the calculation of the agreement.

This procedure allowed the measurement of how many times experts and external annotators used the same speech intentions to describe conversational units. Then, inter-rater reliability was computed with Fleiss' Kappa, which is similar to Cohen's Kappa, but it could be used when more than two annotators labeled each unit [66]. Considering N the number of annotators, n the numbers of units to be annotated, C the number of classes and n_{ij} the number of times the unit i is assigned to the class j, Fleiss' Kappa is computed with the following equations:

$$\kappa = \frac{P - P_e}{1 - P_e} \tag{4.1}$$

$$P = \frac{1}{Nn(n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{C} n_{ij}^2 - Nn) \tag{4.2}$$

$$P_e = \sum_{j=1}^{C}(\frac{1}{Nn}\sum_{i=1}^{N} n_{ij})^2 \tag{4.3}$$

The function "kappam.fleiss" from R package "irr" was used. The null hypothesis (H_0)—the inter-annotator agreement is due to chance, was tested with the z-test (p-value established to 0.05). The interpretation of the Fleiss' Kappa result is the same as for Cohen's Kappa: fair results (0.21-0.4), moderate results (0.41-0.6), substantial results (0.61-0.8) and almost perfect results (0.81-0.99).

4.3 Results

Table 4.1 presents the Kappa statistics for the Searle's five original types of speech acts. This computation is possible because the proposed 18 classes are specifications of the Searle's types. For instance, *sustain* and *assert* are both specifications of assertives. Each row shows the Kappa score for each speech act type obtained by each group. As mentioned, there were 10 groups. Occasionally, some of the groups shared the same dataset: groups 1 and 2; groups 3 and 4; and

[5]More than one label was used in about 16% cases of the total conversation units.

groups 8 and 9. The last column presents the aggregated Kappa score obtained by each group. The second to last row shows the median Kappa score obtained for each speech act type. The reason for choosing the median instead of the mean is that the score distributions for some classes are skewed. However, the median is not influenced by extreme values and is argued to be more interpretable [70]. Finally, the last column shows how often each speech act type is used by all annotators.

Assertive classes were most often used in the annotation experiment, followed by expressive and directive classes. On the contrary, the classes specific to commissive and *Other* were very rarely employed in the classification. The Kappa score is almost perfect for 5 annotator groups, substantial for 4 annotator groups and moderate for one annotator group. The agreement is very high for assertives, expressives and directives, substantial for commissive and moderate for Other. However, in case of the Kappa score for Other, the p-value is sometimes larger than the significance value of 0.05, which means that there is not enough evidence to conclude that the inter-rater agreements are different from what would have been obtained by chance.

Table 4.1: Fleiss' Kappa scores (κ) obtained for each speech act type and overall, computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The second to last row contains the median (Med) of all κ scores per speech act type across datasets. All p-values are 0 except for Other—groups 3, 4, 5 and 6. The last row presents how often each speech act type was used in the experiment, computed by considering all labels provided by each annotator.

Group		**Assertive**	**Expressive**	**Directive**	**Commissive**	**Other**	*all*
1	κ	0.82*	0.77	0.76	0.66	0.72	0.78
2	κ	0.91*	0.90*	0.87*	0.70	0.72	0.87*
3	κ	0.72	0.75	0.88*	1*	-0.01	0.75
4	κ	0.88*	0.87*	0.85*	0.66	-0.02	0.85*
5	κ	0.81*	0.81*	0.88*	0.66	0	0.82*
6	κ	0.53	0.51	0.78	0.65	-0.01	0.57
7	κ	0.82*	0.80	0.88*	0.80	0.84*	0.84*
8	κ	0.72	0.74	0.80	0.80	-	0.75
9	κ	0.78	0.85*	0.86*	0.66	-	0.82*
10	κ	0.72	0.88*	0.74	0.48	-	0.77
Med	κ	0.80	0.81*	0.86*	0.66	0.36	0.80
freq.		1780	754	463	83	48	

Further, Tables 4.2, 4.3 and 4.4 present the Kappa statistics calculated for each class belonging to assertive, expressive, and directive and commissive, respectively. Moreover, the results are reported for each annotator group and the median for each speech intention is also shown. The scores for the classes *thank* (Table 4.3) and *direct* (Table 4.4) show almost perfect agreement. On the contrary, the use of *accept* and *refuse* (Table 4.4) seems to be completely unaligned among annotators. The fact that the p-values for these classes across all annotator groups are greater

than 0.05 shows that the null hypothesis cannot be rejected. Then, apart from *sustain*, *disagree* and *engage*, which show fair and moderate inter-rater agreements, there is substantial agreement in the classification of all the other classes.

The low Kappa scores obtained for some speech intention classes reveal some potential issues of the taxonomy. Unclear definitions and instructions for the use of these classes may be the cause of the issues. However, the very low frequencies of some of these classes in the corpus may be an influential factor, which also has to be considered. The classes *accept* and *refuse* have the lowest agreement in the experiment, but they are also the least represented classes in the corpus (see the last row of Table 4.5 for the frequency of each class). However, the class *sustain* is very frequent, but still shows a low agreement.

Table 4.2: Fleiss' Kappa scores (κ) obtained for *assertive* speech intentions computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The last row contains the median (Med) of all κ scores per speech intention. All p-values < 0.05, except for *disagree*—groups 1, 5, 7, 8 and 10, and *agree*—group 8.

Group		**assert**	**sustain**	**guess**	**agree**	**disagree**
1	κ	0.68	0.36	0.42	0.38	0
2	κ	0.73	0.41	0.71	0.66	0.50
3	κ	0.46	0.31	0.34	0.56	0.49
4	κ	0.52	0.37	0.87*	0.32	0.66
5	κ	0.64	0.47	0.68	0.66	0
6	κ	0.51	0.13	0.58	0.39	0.56
7	κ	0.51	0.29	0.51	0.79	0
8	κ	0.66	0.52	0.68	0	0
9	κ	0.78	0.53	0.82*	0.66	1
10	κ	0.61	0.19	0.65	1*	-0.02
Med	κ	0.63	0.37	0.67	0.61	0.26

In order to get more insights into the common disagreements, the confusion matrix is projected in Table 4.5. Overall, the class *assert* appears to be the fallback choice:

- The classes *rejoice* and *complain* are more frequently confused with *assert* (about 20% of their occurrences; for instance, in the utterance "I was ill for 3 entire months!") than with any other class (2% of their occurrences). This result can be seen as an evidence for the continuity between the assertive and expressive speech act types, claimed in Section 4.1.

- Within the assertive type, *sustain* was very challenging to differentiate from *assert* (42% of *sustain* occurrences; for instance, in "It was clearly just a conjecture, though, there's no data out there to confirm this"). This confusion may be caused by unclear definitions of *sustain* and of the oppositional trait differentiating *sustain* from *assert*.

Table 4.3: Fleiss' Kappa scores (κ) obtained for *expressive* speech intentions computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The last row contains the median (Med) of all κ scores per speech intention. All p-values < 0.05, except for *rejoice*—group 6, *wish*—group 6, *apologize*—group 1, *thank*—group 1, and *greet*—group 5.

Group		**rejoice**	**complain**	**wish**	**apologize**	**thank**	**greet**
1	κ	0.56	0.81*	0.20	0	0	1*
2	κ	0.66	0.91*	0.23	-	1*	1*
3	κ	0.90*	0.67	1*	0.66	1*	0.56
4	κ	0.85*	0.89*	1*	0.66	1*	0.80
5	κ	0.90*	0.76	0.74	-	1*	0
6	κ	0.15	0.43	0	0.49	1*	0.66
7	κ	0.66	0.39	0.50	0.66	0.91*	-
8	κ	0.56	0.78	0.89*	-	1*	0.39
9	κ	0.76	0.75	0.65	-	1*	0.66
10	κ	0.77	0.79	0.91*	1*	1*	0.66
Med	κ	0.71	0.77	0.70	0.66	1*	0.66

Table 4.4: Fleiss' Kappa scores (κ) obtained for *directive* and *commissive* speech intentions computed for each group of annotators. The scores are rounded to two decimal places. "-" denotes that the class has not been used in labeling the dataset. At least substantial results are underlined. Almost perfect results are marked with "*". The last row contains the median of all κ scores per speech intention. All p-values < 0.05, except for *engage*—groups 5 and 8, *accept*—groups 1, 2, 5, 6, 7 and 9, and *refuse*—groups 1, 2, 3, 4, 8 and 10.

Group		**direct**	**suggest**	**engage**	**accept**	**refuse**
1	κ	0.52	0.65	0.53	0	-0.01
2	κ	0.70	0.71	0.53	-0.01	-0.01
3	κ	0.90*	0.74	1*	-	0
4	κ	0.94*	0.77	1*	-	-0.01
5	κ	0.91*	0.73	-0.01	0	1*
6	κ	0.60	0.69	0.43	-0.01	-
7	κ	0.89*	0.87	0.50	0	-
8	κ	0.96*	0.47	-0.01	1*	0
9	κ	0.89*	0.54	0.39	0	-
10	κ	0.52	0.85*	0.66	0.49	0
Med	κ	0.89*	0.72	0.52	0.01	0.01

Table 4.5: Confusion matrix for the proposed speech intentions. For the sake of brevity, only the first three letters of each class are used to denote the class. The pairs without the empty choice, for all groups of annotators, were taken in consideration for computing the matrix. The last row presents how often each class was used in the experiment, computed by considering all labels provided by each annotator.

	ass.	**sus.**	**gue.**	**agr.**	**dis.**	**rej.**	**com.**	**wis.**	**apo.**	**tha.**	**gre.**	**dir.**	**sug.**	**eng.**	**acc.**	**ref.**	**oth.**
ass.	**415**	56	16	4	2	6	11	1	1		2	4	5	1	1	2	3
sus.	63	**82**	7	1	2	3	9	3		1		2	2	4	1	4	
gue.	16	7	**67**	3	2		2	1	1		1	1	2	2			
agr.	3	2		**17**		1					1						1
dis.	6	2			**6**	1									1		
rej.	17	4	1	2	1	**77**	1		1				1				1
com.	24	12	3	1		1	**147**	2	1			2	1	2		1	1
wis.		1	3					**21**			1	1		2			
apo.									**5**								
tha.						1				**34**							
gre.	2	2		1	1	2					**12**						
dir.	6	2	2		1	1	1	7			1	**93**	6				1
sug.	17	8	2	3		1	2					3	**98**				
eng.	3		1					2					3	**13**			
acc.	1	1			1									5	**2**		
ref.									2							**1**	
oth.	2	5				1	1					1	5				**16**
freq.	1096	351	227	70	36	204	366	64	20	65	35	211	252	54	18	11	48

- The classes *agree* and *disagree* were often taken for other types of assertive (18% and 31% of occurrences, respectively; for instance, in "Same symptoms, too."). The reason here may be the fact that the annotators possibly did not consider the preceding sentences in conversation during the experiment. The annotation was performed in a Google spreadsheet. Perhaps, another annotation interface could have emphasized better the context.

- Finally, in some situations, *suggest* was confused with assertives (for example, in "For the sore joints, I try to workout twice a week to keep my body from feeling too stiff"). This shows misalignment in identifying indirect suggestions.

4.4 Conclusion

The contribution of this chapter is an original taxonomy that meets multiple requirements, not all covered by any of the related works. It is finer-grained, corpus- and domain-independent, with a manageable number of classes and an associated structure. Thus, an answer to **RQ1**, with a focus on speech intentions, was provided in a second iteration.

The taxonomy has 18 classes of speech intentions, which are structured with reference to multiple oppositional traits. This implies that the classes can be identified during annotation by following some explicit criteria. The aim of this taxonomy design is to ensure that the taxonomy application is feasible in any empirical study on conversation corpora. Moreover, the taxonomy is created to be used by both experts and non-experts, encouraging thus its application across domains and allowing to further prove the alignment in the perception of speech intentions.

The taxonomy emerged from linguistic theories [12, 184, 212]. It is a diversification of Searle's speech act types, while keeping the number of classes smaller than in other linguistic taxonomies [196, 212]. A similar approach to expand Searle's speech act types into a finer-grained taxonomy for empirical corpus exploration has been followed by Garcia [72]. However, Garcia's classification is specialized to the teaching context.

Compared to the taxonomy proposed for the study of public tweets in Chapter 3, the expressive and commissive speech act types are included and diversified in the current version. Moreover, the assertive type contains now *sustain* too; *hypothesize* and *guess* classes are equivalent. With regard to the directive type, advices and warning are merged in the class *suggest*, which shows a high consistency; also the class *propose* is no longer included because preliminary annotations did not reveal its presence. However, when analyzing what types of utterances were classified using *other*, it was observed that often these contained links, thus supporting the re-introduction of *propose* in directives, apart from *direct* and *suggest*.

The validation of the taxonomy in the experiments confirms reproducibility and interpretation consistency for most classes, but *accept*, *refuse*, *disagree*, *sustain* and to a lesser extent *engage*. More annotations may help with identifying if the same issues persist and also their causes.

Additionally, more annotations need to be performed in order to prove if the taxonomy is corpus-independent empirically or to show if, eventually, it is more suitable for some particular cases.

4.4.1 Study Limitations

To mitigate the threats to *construct validity*, Fleiss' Kappa, a common measure to evaluate the inter-rater agreement, was used. The results of this statistical test could have been influenced by the proposed way to align the sets of labels provided by different annotators. However, the alignment of the label sets could, at most, influence the confusion identification, but not the consistency score of each class use. The common classes of each two label sets were always aligned first—hence, the agreements were always identified. Also, as in the experiments, only 16% of the sentences were labeled with more classes, it can be considered that the confusion matrix manages to capture the most significant trends. Finally, the classes were created starting from previously accepted and well-known linguistic works to conceptualize speech intentions.

With regard to the *internal validity*, various measures to support it were taken. First, an experimental protocol emerged during three iterations of manually annotating corpora internally. Then, in order to enforce the confidence of the relationship between the input—the proposed structural taxonomy, and the output—the consistency in applying the classes, multiple datasets were annotated by different pairs of experts and external annotators in an experiment. However, some results indicate that the use of a spreadsheet for annotation may have led to ignoring the utterance context. An external factor that could have influenced the result, the agreement by chance, was counteracted through the use of the Kappa statistics and z-test (p-value=0.05).

The *conclusion validity* was also ensured by statistically testing the results, with z-test and an established confidence level of 95%.

Finally, the main threat to the *external validity* is the use in the experiments of conversations on the same topic and from the same online discussion board. However, apart from this, different conversations were labeled by most of the groups. Additionally, two different experts and ten different external annotators participated in the experiments.

4.4.2 Directions for Further Research

A corpus-independent and detailed taxonomy of speech intentions was designed to conceptualize asynchronous conversations. However, it was validated by being applied to only one corpus representing one type of asynchronous communication. Consequently, future work should be mainly targeted on assessing the taxonomy application on very diverse corpora and on identifying its potential for practical applications, apart from the more general role it could play in the linguistics and conversation analysis domains.

Then, although most of the proposed classes show at least substantial agreement among annotators, issues were also identified for some speech intentions. Thus, improving the class definitions and more annotations to identify other possible causes, also on different corpora, are

required. Additionally, the annotation instructions should be more clearly formulated and another annotation tool that could better emphasize the context should be explored. These aspects were also emphasized by the subjects through the provided feedback form.

A further step is to use machine learning to automatically associate speech intentions to sentences. Once automatized this, larger and less specialized conversation corpora can be exploited for empirical research in linguistics, conversation analysis and other concerned domains.

CHAPTER 5

MODELING CONVERSATIONS WITH SPEECH INTENTIONS

Epure, E.V., Compagno, D., Salinesi, C., Deneckere, R., Bajec., M., Zitnik, S. (2018). Process Models of Interrelated Speech Intentions from Online Health-related Conversations. *Artificial Intelligence in Medicine*. doi = https://doi.org/10.1016/j.artmed.2018.06.007

Contributions: E.E.V. designed the method and the experiments and wrote the article, Z.S. provided feedback on the experimental design, E.E.V. and Z.S. implemented the machine learning experiments, E.E.V. analyzed the results, S.C, D.R., R.B., Z.S. and C.D. provided feedback on the article.

The contribution of this chapter is:

- A corpus-independent, automatic method to annotate sentences of forum conversations with multiple speech intentions from the proposed taxonomy, with supervised machine learning. Compared to the related works, the solution provides the most detailed representation of forum conversations as each sentence of a turn is annotated with multiple labels and the proposed taxonomy has the largest number of classes. Additionally, extensive experiments are conducted: to compare multiple scenarios for classification; to assess the effectiveness of varied groups of features regarding the content, the discourse form and the conversation structure; to test different strategies to improve the classification of some labels challenging to identify; and to address the external validity by discovering a subset of speech intentions from two other, heterogeneous corpora.

A speech intention taxonomy consisting of 18 classes, specializations of the Searle's five types of speech acts, and multiple oppositional traits was proposed and presented in Chapter 4. The taxonomy was designed starting from the theoretical linguistic works, but also empirically by

analyzing various corpora. Through design, the taxonomy is corpus- and domain-independent and comprehensive—as in both finer-grained and exhaustive. Through experiments, it was proven that, apart from four classes—*accept*, *refuse*, *disagree* and *sustain*, all the others were consistently perceived and used by multiple human annotators.

In addition, these results were obtained with non-expert annotators, demonstrating the potential of the taxonomy to be used in diverse fields and its consistent perception by different people. Therefore, although improvements and more experiments are still needed, the taxonomy manages to satisfy the solution requirements formulated in Chapter 1 and Chapter 2, being an improved alternative for the study of asynchronous communication compared to the related works. Moreover, compared to most of these related works, the taxonomy was extensively validated.

The aim of the current chapter is to investigate to what extent speech intentions and speech act types can be automatically predicted from asynchronous conversations. Hence, the second proposed research question is addressed in a second design iteration: *How to automatically discover the proposed speech intentions from asynchronous conversations independently of the domain and corpus characteristics?* (**RQ2**).

Specifically, the feasibility to automatically label conversation utterances with the taxonomy and Searle's speech act types is tackled. To achieve this aim, supervised machine learning is chosen as it is highly popular and proven effective in the existing literature, as motivated in Chapter 3 too. Nonetheless, supervised classification is challenging because it requires a validated ground-truth corpus obtained through manual annotation. Also, features as cues of speech intentions must be proposed for use by classifiers.

This chapter is a continuation of the previous contribution chapters as now the extended classes are used for annotation and corpora covering other types of asynchronous communication than tweets are explored. Moreover, several questions raised at the end of Chapter 3 about the impact of the machine learning experimental design on the results are addressed here as well. The relevant aspects identified in this regard are the selection of features, tuning the parameters of the classifiers, sampling to overcome class imbalance and exploring multi-label classification.

In Section 5.1, medicine is revisited as an application domain. Specifically, the potential improvement opportunities brought to this domain if the current aims are achieved are discussed. Also, a better understanding of the theoretical foundation supporting these claims is provided by examining other existing theories on behaving through language.

Then, in the sections to follow, the ground truth corpora used in the subsequent experiments are described (Section 5.2) and the extended set of features to be extracted from conversations for classifiers is presented in Section 5.3. Further, the complete experimental setup to address the research aim is detailed in Section 5.4 and the results are summarized in Section 5.5. Finally, the threats to validity and directions for future research are discussed in Section 5.6.

5.1 Medicine as Application Domain

The massive spread of online communication is at the same time an opportunity and a threat for public health [39]. The reason is that online communication could be a drive for behavioral changes [39, 88]. This is positive and desirable [36, 102] when leading to healthy habits or to the adoption of constructive attitudes towards the management of chronic diseases. However, online communication can also have a negative and undesirable influence when inducing unhealthy behavior for citizens [74]. As every person or organization can disseminate information about health, contents are created at a very fast pace, making digital sources hard to monitor and regulate [39, 180]. Consequently, erroneous and misleading information reaches a very large and often vulnerable audience, resulting in high risks for those directly or indirectly exposed to it.

Obtaining a better understanding of behavioral changes resulting from and through online communication becomes important for medicine in order to exploit the opportunities and address the issues. This need has been already acknowledged and tackled by academic communities dealing with health information seeking and dissemination [39, 180], persuasion and compliance gaining, including persuasive technologies for health care [36, 102], and narrative medicine [34, 35]. In order to be able to model and predict behavioral changes, meaningful insights into online communication, as a form of written traces of behavior and human interaction, should be gained, so to better understand the potential factors of change.

Further, it is motivated why an approach centered on *speech intentions* is a suitable choice for the conceptual framework and practical aim, according to current theories in psychology, sociology, philosophy and linguistics. As already described in Chapter 1, the term "speech intention" stands for the perceived conventional purpose of utterances—also referred to as speech act [12, 184, 197].

Theoretical Aspects on Behavioral Changes. The theory of planned behavior states that humans as agents act consistently with behavioral intentions, which are adopted on the basis of personal beliefs, shared attitudes and social norms [5]. *Social norms* in particular are related to how other people accessible to the agent perceive a certain behavior. Norms could be thought of as accepted rules of behavior, tacitly or explicitly emerging in a certain community. Hence, a behavioral change is triggered, not only by a change in beliefs, but also by the social circle acknowledging its acceptance [23, 31].

One theoretical approach to describe how changes in beliefs take place is the so called *narrative paradigm* [65]. Researchers referring to such a theoretical approach sustain that people use stories to explain their experience of the world. Personal stories emerge from individual experiences, while others are conveyed collectively by the community. Several kinds of "good reasons" may induce agents to embark on a story and so to adopt some beliefs. "Good reasons" does not necessarily mean rational, formally valid reasoning; the expression refers to what is *perceived as meaningful* and consistent with one's past experiences or the opinions of others [65]. Thus, the narrative paradigm implicitly acknowledges the importance of social norms.

Persuasion and compliance gaining are two other concepts relevant to the study of behavioral changes [221]. *Persuasion* is defined as the intentional act of changing the beliefs and attitudes of others. *Compliance gaining* is defined as the intentional act of changing the behavior of others. As shown by the theory of planned behavior [5], persuasion and compliance gaining are related.

The picture of human action drawn by these references highlights the importance of social, and especially discursive factors in adopting and modifying one's behavior. Philosophy and linguistics in the 20th century put forward how important speech is for taking action and influence each other [9, 12, 184]. Language production is one of the main ways through which humans realize acts and bring about their intentions. Such intentions are expressed with words, susceptible of philosophical and scientific analysis. As an example, the utterance "You should try this treatment! It really worked for me." is acting in itself associated to certain intentions that can be subsumed by the speech acts called *advise* and *assert*. Moreover, a reply such as "I will, thanks for the suggestions!" shows an *acceptance*. This verbal exchange exposes how behavioral influence takes place. The identification of processes of speech intentions associated to these utterances could support the identification of potential changes in behavior, but also of potential changes in beliefs. Consequently, the study of language shows its potential relevance for providing insights into behavior and behavioral changes.

Envisioned Applications. Given that speech realizes intentions and that intentions could influence planned behavior [9, 21], it becomes evident that the analysis of online text as digital behavioral traces can potentially be useful to the aims of medical research in fields such as narrative medicine, health information dissemination and seeking and technologies for health persuasion and compliance gaining.

Narrative medicine states that effective health practices should "recognize, absorb, interpret and act" on the stories shared through the experience of illness [34]. Patients and their close circles are encouraged to ease their way through the disease with stories. Telling should be used as a way to unburden oneself, to bond with others and to share. Practitioners should learn about the situations of patients through stories and identify new views on illnesses. Hence, the therapeutic relationships are expected to strengthen and new knowledge to be gained.

Narrative medicine practices seek to understand the shared "chaotic or formless experience" of diseases [35]. When trying to derive meaning from a story, not only the content is important but also "how the text is configured" [35], that is, which strategy is adopted by the author in order to be understood by an envisaged or "model" reader [49]. Readers cooperate with textual traces and cues in order to infer a comprehensive meaning. Moreover, the interpretation is often realized through actual replies and readers may end up "asking for witness, for presence, for answer" [35]. By modeling online communication, the ground for the massive automatic analysis of stories and of the attested reactions to them is set. Specifically, narrative knowledge can be extracted by attributing meaning to communication through symbolic means [34], namely through speech

intentions. In this way, insights into how different narrative stances are constructed can be gained (e.g. pro- and anti-vaccination communities might use diverging strategies of communication) and into how to identify when people embark on these stories and commit to act as a result.

Health information seeking and dissemination research looks into how health information is disseminated online and how health information consumers engage with it [39]. Multiple benefits and threats revolve around this topic, with strong implications for the health care system. As advantage, the Internet creates a highly available, anonymous and accessible environment where information can be found and human interaction can take place at one's fingertips. This results in better support and education in health matters [25, 39]. As disadvantages, first, the ways of presenting information, often via technical terms or complex system interfaces, can be problematic. Second, the most important issue is the lack of feasible methods to assess the quality of online information [39, 180]. Cline and Haynes [39] state that modeling interpersonal interactions and communication supported by the Internet is necessary in order to further understand the effects on health-related beliefs and behaviors. By modeling communication through meaning as perceived by readers [49, 51, 184], more suitable message framings for some given audiences may be enabled; the health information seekers could be helped to locate what they are looking for, not only by topic, but also by the type of message (e.g. question, advice etc.); and content with potential triggers for behavioral intentions or beliefs could be identified.

Persuasive technologies are computer systems, applications and devices created with the aim to alter one's attitudes and behaviors for improved health habits [36, 102]. Getting insights into how attitudes and behaviors are adopted is a priority for the designers of such technologies [36]. Ultimately, communication and its effects on people in terms of persuasion should be investigated [36, 102]. The theoretical basis integrated in the creation of persuasive technologies can be complemented with knowledge learned empirically from large scale, online communication. Identifying different strategies to suggest behavior or to gain compliance could help to diversify the ways the communication is addressed to specific audiences. Further, inferring human intent—objective at the core of this work—is explicitly highlighted as an important direction of investigation by researchers in persuasive technologies [102].

5.2 Ground-truth Corpora

Three corpora are used in the current work. The first is created starting from the datasets used in the experiment for the taxonomy validation (presented in Chapter 4). It consists of Reddit conversations labeled with the intention taxonomy (hereinafter referred to as *Reddit*, 2280 instances). The other two corpora are external, labeled with different classes of speech intentions. One is composed of forum conversations, released by Bhatia et al. [14] (referred to as *Bhatia*, 461 instances). The other is a corpus of synchronous conversations (referred as to *SWBD* [110, 196], 94950 instances). The external corpora are used for assessing the classifier external validity and

only some of their instances are considered, as it is explained at the end of this section.

In Table 5.1, statistics for the three corpora are presented. As a reminder, a turn in any type of conversations can be composed of multiple utterances (grammatical sentences). In asynchronous communication, a turn is referred to as a post, which can be a starting post or a comment. Frequently, a conversation can be modeled as a tree—more generally as a graph or as a forest of graphs, with each turn being a node. In the context of a tree model, a conversation thread is then a path from the root—the starting post, to a leaf—a last-level comment.

Table 5.1: Statistics of the *Reddit* corpus and of the two external corpora, *Bhatia* and *SWBD*. *Bhatia* is released as threads of conversations and each turn is labeled. *SWBD* is composed of synchronous conversations and does not contain threads; each utterance or part of it is labeled. *Reddit* is labeled at the utterance level (grammatical sentences).

Dataset	**#Turns**	**#Utterances**	**#Conversations**	**#Tokens**	**#Threads**
Reddit	395	2280	51	35790	204
Bhatia	556	-	-	35585	100
SWBD	-	223606	1155	2073791	-

As already mentioned, the *Reddit* corpus is derived from the datasets used in the experiments for validating the taxonomy, plus from some other datasets annotated in parallel by pairs of experts, post-validation[1]. The 17 classes used in the manual annotation experiments are presented in Table 5.2. The goal of the taxonomy validation was to discover if different human annotators applied the intention classes consistently to a real-life corpus and thus attributed similar meaning to utterances. Specifically, Reddit conversations about autoimmune diseases were segmented in sentences and tagged by 10 pairs of annotators, each pair composed of a non-expert and an expert. Each annotator could choose up to 3 speech intentions to annotate each sentence—multi-label annotation. Applying the Kappa (κ) statistical test to measure the inter-rater agreement, at least moderates scores were obtained for most intentions—moderate κ for *engage*; substantial κ for *assert*, *guess*, *agree*, *rejoice*, *complain*, *wish*, *apologize*, *greet*, *suggest*; almost perfect κ for *thank* and *direct*. Additionally, almost perfect agreement was obtained for all Searle's speech act types, apart from commissive which had substantial score.

The lowest scores were obtained for *sustain*, *disagree*, *accept* and *refuse*. The *sustain* class was among the most often used labels. However, its application revealed high disagreement between annotators, being mainly confused with *assert*. Thus, it did not pass the validation. With regard to *accept*, *refuse* and *disagree*, their very low occurrence may have influenced their low κ scores, as shown also by the p-value ≥ 0.06 computed with z-test, making the results not conclusive. In consequence, to create the ground-truth corpus for the current experiments, *sustain* was merged with *assert*; *disagree* and *agree* to a new, more general class named *agree*; *accept* and *refuse* with

[1]21 conversations were annotated in the validation experiments by pairs of experts and non-experts. Post-validation, pairs of experts annotated other 27 conversations. The process to establish the final labels for each sentence of the 48 conversations was the same: the intersection of the two label sets was chosen and, if not available, two experts decided the labels through discussions.

Table 5.2: Description of the speech intentions. The frequency of each class is presented under the name—if missing, the class is not used in the automatic experiments. *Assert* includes also *sustain*, *agree* stands for *agree* and *disagree*, *engage* includes also *accept* and *refuse*.

Class	Description	Example(s)
assert sup.=1193	Plain statement.	*The labs billed my insurance for the initial battery of tests.*
sustain	Statement that increases the confidence of a claim with explicit reasons or examples.	*We have a $1000 deductible plan, so I'd rather use the money for medications.* *I strongly believe it's the good thing to do.*
guess sup.=146	Statement that weakens the certainty of the claim by showing doubt, possibility or probability.	*I don't know.* *So I would say it's kind of both?* *I say it as one who's not through this daily!*
agree sup.=78	Agreement with a previous statement or positive answer.	*He's absolutely right.* *Yes.*
disagree	Disagreement with a previous statement or negative answer.	*No you aren't being needy.* *Doesn't sound like lupus to me.*
rejoice sup.=152	Positive attitude about something or someone. It includes positive attitudes towards negative cases.	*Great!* *He always believes me and it's so nice because everyone else doesn't.*
complain sup.=397	Negative attitude about something or someone, including dissatisfaction, blame, condolences, grieving.	*My doctor should have told me earlier...* *I feel sorry for your loss.* *I'm tired of this medication.*
wish sup.=75	Desire for future or possible events. It includes negative situations too.	*Hopefully, you have a net of support.* *Good luck to you both.*
apologize sup.=21	Excuse for some fault.	*I apologize in advance for my redundancy.*
thank sup.=54	Expression of gratitude, acknowledgement, appreciation.	*Thanks in advance.* *I highly appreciate your help.*
greet sup.=20	Salutations of all kinds.	*Hi there, from Colorado USA!*
direct sup.=184	Statement that expects the reader to reply, accept or refuse. It includes orders, requests, questions and invitations.	*How can I be of help to her?* *Any advice is appreciated.* *Go home and decide by tomorrow.* *Subscribe to this group for information.*
suggest sup.=225	Statement aiming to influence reader's acts by implying what is good/bad. It can be advice, warning, suggestion, recommendation.	*The best now is to be his pillar of support.* *Careful with the online advices.* *Go and buy an aspirin, you'll feel better!* *Always trust your sensations.*
engage sup.=82	Promise, commitment or intent of the speaker to act in the future.	*I will follow a new treatment next summer.* *I'm returning the book tomorrow.*
accept	Acceptance to comply with some request, invitation, proposal.	*I'll definitely consider your suggestions.*
refuse	Refusal to comply with some request, invitation, proposal.	*I don't think this solution would be the best for me.*
other sup.=46	Utterance not fitting any other class.	*http://www.cdc.gov/flu/protect/vaccine/general.htm#side-effects*

engage. The 13 resulting classes are marked in Table 5.2, by showing for each one its support. For now, the granularity of the taxonomy was sacrificed to some extent, but it still incorporates a high level of discriminating detail: 3 classes of assertive speech acts, 6 classes of expressive speech acts, 2 classes of directive speech acts and 1 class for commissive. Clearly and as also discussed in Chapter 4, annotating more conversations containing the other merged intentions would enable their individual study. Future manual annotation experiments could help to identify the causes of disagreements in order to eventually improve the taxonomy or the tagging instructions and to collect more instances for the least represented classes.

Further, as each annotator could label an utterance with up to three classes, each utterance had finally associated the intersection of the proposed label sets—after applying the intention merging, as described in the previous paragraph. Where no common labels existed, two experts discussed the reasons for disagreement and made an ad-hoc decision. The disagreement sentences could have been excluded from the classification experiments. However, the representation of each turn had to be complete in order to later support the accurate discovery of processes from conversations. Finally, let's notice that the labels associated to each utterance in the ground-truth corpus are not necessarily complete and do not cover all possible perceived intentions for an utterance. Similar to the previous experiment on Twitter corpora, the main goal is not to obtain an extensively labeled corpus—that is hard to obtain also for conceptual reasons such as the presence of indirect speech acts, but to ensure the creation of a precise and reliable corpus—that is when the utterance is associated with a speech intention, this association is indeed true.

The external corpora are included in the experiments for validating the performance of the trained classifiers on datasets that differ from *Reddit* in several perspectives: the application domain, the conversation format and the labeling units.

Bhatia dataset consists of forum threads discussing issues and solutions related to the Ubuntu operating system [14]. The dataset is labeled at the post level—where a post contains multiple utterances, with their ad-hoc taxonomy designed for forum analysis (classes in Table 5.3).

SWBD, standing for Switchboard-DASML dataset, contains transcribed, spontaneous phone conversations on varied topics [196]. Each utterance or part of it is labeled with one of the 42 mutually exclusive labels, named dialogue acts (see Table 5.3 for class examples).

The external corpora contains different classes, but a part of these can be aligned to some of the proposed taxonomy intentions. Only the aligned classes are used in the experiments, the mapping being presented in Table 5.3. Also, the classes are not semantically identical. The current taxonomy intentions are sometimes subsets or supersets of classes from other corpora (e.g. *direct* is part of *Bhatia*'s *Clarification*).

In the experiments, each instance of the *Reddit* corpus is one sentence, while in *Bhatia* the whole post is an instance and in *SWBD* a sentence or a part of it. Also, the *Reddit* dataset has a hierarchical structure, while the external corpora are released as sequences of turns. The obtained classification results will likely be poorer because of these heterogeneities. Nonetheless,

Table 5.3: Mapping of the taxonomy classes on *Bhatia* and *SWBD* classes. Not all classes could be aligned. The number of instances for each class from the external corpora is in parentheses.

Reddit	*Bhatia*	*SWBD*
assert	Further details, Solution (262)	Statement-non-opinion (76627)
guess	no mapping	Hedge, Maybe/accept-part, Other answers (1660)
agree	no mapping	Yes answers, Affirmative no-yes answers, No answers, Negative non-no answers (4344)
direct	Question, Clarification (163)	Yes-no-question, Wh-question, Declarative Yes-no-question, Backchannel in question form, Tag question, Declarative Wh-question, Open question, Action-directive (12041)
engage	no mapping	Offers, Options Commits (115)
rejoice	Positive feedback (25)	no mapping
complain	Negative feedback (11)	no mapping
thank	no mapping	Thank (82)
apologize	no mapping	Apology (81)

it will be a good indicator of the robustness of the classifiers and the generalization of the results.

5.3 Feature Engineering

Similar to the previous work on modeling public tweets with speech intentions, *Content* and *Discourse* features are defined. *Content* features are standard text mining features derived from words. *Discourse* features incorporate linguistics means interpreted as potential cues to express speech intentions. Compared to the features presented in Chapter 3, *Content* features are now *n-grams*. Also, *OpinionKeywords* features are no longer considered as *Content* features, but as *Discourse* because they are cues of expressive speech acts. Moreover, as more intentions and speech acts types are considered now, more *Discourse* features are also defined to capture various attributes of the text-based speech conveying them. Also, a new group of features exploiting the whole structure, *Conversation*, is proposed as in [14]. The complete feature list is in Table 5.4.

The process to discover the features was the following:

1. Related literature on speech act and dialogue act discovery was studied and a first list of features was identified [10, 14, 28, 106, 110, 124, 133, 196, 234].

2. Only the features that appeared highly predictive or popular in the literature were kept.

3. By analyzing theoretical and empirical works in linguistics [12, 183, 184, 187, 212], this list was augmented with some new features (e.g. the imperative mood).

4. The list of features was tested in the previous experiments on a Twitter corpus, where the rationale behind the *Discourse* features was detailed too (see Chapter 3).

5. As new intention classes were introduced and the type of data changed, additional *Discourse* features and the new *Conversation* features were added. These new features were identified after a second scan of the literature focused on asynchronous communication and on all speech act types [10, 14, 28, 133].

Table 5.4: Proposed features (3 main groups): 1–*Content*, 2–*Discourse*, 3–*Conversation*; when "highly correlated" features are mentioned, these are selected with the χ^2 statistical test.

Name	**Description**
1.n-grams	sequences of 1-3 words; all and those highly correlated.
2.PronomCues	frequencies of 1st person singular, of 1st person plural, of 2nd person and of 3rd person pronouns; all forms—subjective, objective, possessive and reflexive—are considered.
2.PunctMarks	presence of "?", of "!", of ":" and of ellipsis (starting with "...").
2.QuestionCues	frequency of "what", "when", "where", "who", "why" and "how".
2.ExpressCues	frequencies of emoticons—both ASCII and Unicode considered, of interjections, and of negations; frequencies and ratios of negative and positive opinion keywords [129].
2.VerbCues	frequency of each modal verb; presence of verbs in general, of future tense, of past tense, of imperative mood, of noun before verb and of verb in "-ing"; for all these features negated verbs are also included.
2.TextLength	total and unique-word length of the utterance—as is and after stemming.
2.TextForm	presence of urls, of multiple, sequent words with first letter upper case, of words entirely upper case, of repetitions of words.
2.POSCues	frequencies of various sequences of two consecutive POS tags—all and those highly correlated; POS n-grams statistically correlated with the classes (up to length 3).
3.ConversPos	utterance absolute and normalized position; post position.
3.QuotedPost	current utterance is quoted from another post.
3.SimilarScore	cosine similarity to previous and to initial posts, to the whole conversation.

Further, the rationale of the newly introduced features is presented (for the others, refer to Chapter 3, Section 3.2.2). Multiple verb-related features are defined. The presence of future verbs and of verbs ending in "-ing", preceded by the first person pronouns can be cues of *engage* (e.g. "I will start this new treatment next month"). The past tense could be mainly related to either assertive or expressive, as it refers to reporting in the past, which could have a neutral or opinionated stance. Regarding the expressive classes, the presence of negations could be correlated with *complain* as well as the presence of consecutive uppercase words. Interjections are signs of expressive speech acts, in general. The length of utterances is relevant because shorter turns can be associated with *accept*, *agree*, *greet* and *thank*, while longer turns can be *assert*—in particular the integrated *sustain*, *rejoice* or *complain*.

Syntactic constructs are slightly differently extracted compared to the previous approach, presented in Chapter 3. Now, they are dynamically generated from the corpus as POS n-grams. Also, all POS n-grams and only those highly correlated with the speech intentions—based on the χ^2 statistical test—are considered. The position of a sentence in a turn is relevant because, for instance, expressives tend to occur at the beginning or end, while directives mainly at the end for starting posts. This intuition is also outlined in [163]. The position of a turn in conversation is suggested in [14] to be an attribute of various classes. For instance, in the current corpus, *direct* utterances are expected to be in the starting posts, *suggest* utterances in the comments replying to the starting posts, *engage* and *thank* utterances in the comments replying to the previous comments. The quotation of a previous utterance in a turn could have associated various types of intentions, such as *direct*—when new questions are asked, *suggest* and *assert*—for providing an answer either as an advice or as an information, *rejoice*—for expressing positive attitudes towards the initial post and so on. Additionally, as suggested in [14], the similarity between posts could be relevant because questions and answers may share similar words.

5.4 Experiments

The aim of the experiments is to assess the feasibility of annotating sentences, extracted from turns of asynchronous conversations, with the proposed speech intentions. Specifically, the overall results and the results for each speech intention obtained by different classifiers with different configurations of features are compared. Additionally, features are analyzed with respect to the importance they have in the correct classification of each speech intention. The main corpus in the experiments to address these objectives is *Reddit*, which is composed of 2280 instances. Finally, in order to assess its external validity, the best classifier trained on *Reddit* was used to predict the sub-selection of intention classes on the external corpora (*Bhatia* — 461 instances, and *SWBD* — 94950 instances; see Table 5.3 for the class alignment).

Multiple text-processing tasks were applied to prepare the data for the experiments. First, each utterance was tokenized and then enriched with lemmas and part-of-speech (POS) tags. Further, each conversation was transformed into a tree data structure, with the initial post—composed of a sequence of utterance objects—being the root node. All the other posts directly replying to the initial post are direct children of the root and the same rule applies for all the descendants, accordingly. Then, in the next step, the features defined in Table 5.3 were extracted for each utterance and normalized to the $0-1$ range.

In the research field of speech act modeling, multiple classes of speech intentions are often used, but each utterance may have one single class associated (*multi-class* classification; e.g. [14, 196, 234]). However, some works employ a *multi-label* annotation schema where an instance is labeled with multiple speech intentions. The approaches in this situation are to train binary classifiers to recognize each speech intention and then aggregate the true predictions of all

classifiers to give the final labels for an instance [10, 28, 91] or to train binary classifiers to directly recognize pairs of speech intentions [150]—maximum two labels can be then predicted for an instance . Similarly, in the current work, multiple machine learning setups for both multi-class and multi-label classification were implemented and their results compared.

In the first scenario, several types of classifiers were trained to recognize for each sentence the most likely speech intention (multi-class). The selected types of classifiers were *Logistic Regression*, *Linear SVM* and *Random Forest*, which are models commonly appearing in the literature and also proven the most effective in the previous experiment on modeling public tweets (see Chapter 3). Additionally, in this multi-class scenario, the data needed to be transformed because a part of the instances had multiple labels. Inspired by a previous work [149], which prioritized the minority class with improved results, the label representing the class with the lower support in the corpus from the three options was selected for each utterance.

In the second scenario, binary classifiers were trained and evaluated for each speech intention. The type of classifier was chosen based on the best results obtained in the previous setup. Moreover, compared to the previous setup, where one single label needed to be chosen for an instance, this time all labels were relevant. Before training a classifier for a speech intention, each instance in the corpus was marked with 0 or 1, depending on the absence or presence of the target class among its labels. The transformation was made for each class (one-versus-all).

In the third scenario, a multi-label setting was created by using a problem transformation method, namely *Binary Relevance* [168]. First, each instance of the corpus had the label set represented through a binary array marking the presence or absence of each defined speech intention. For the current context and depending on if it is a classification of speech intentions or of speech act types, the binary array has 13 slots corresponding to the 13 speech intentions or 5 slots corresponding to the 4 speech act types plus *other*[2]. Then, the best performing type of classifier according to the first scenario was selected. A classifier of this type was trained in an-one-versus all approach for each class to recognize if the class could characterize the sentence. This scenario was an extension of the previous to measure the extent to which all known classes of an utterance were identified.

The *scikit-learn* [155] implementations of Logistic Regression, Linear SVM and Random Forest were used. The classifier parameters were set automatically in order to optimize the F1-score using internal cross validation on the training data. The Logistic Regression classifier required also the specification of a number of iterations for optimal performance. A 10-fold cross validation setup was adopted for the first and second scenarios, where the training and the testing sets were created through a stratified sampling approach. For the third scenario, two thirds of the data were used for training and one third for testing—the split was in a stratified

[2] As an example, considering that the first position in the binary array marks the presence or absence of assertive speech acts, the second position—of expressive speech acts, the third position—of directive speech acts, the fourth position—of commissive speech acts and the fifth position—of *other*, then a sentence labeled with {*assert*, *complain*} has associated the array $[1,1,0,0,0]$ and another one labeled with {*engage*, *thank*}, the array $[0,1,0,1,0]$.

manner, as before. The initial training and evaluation of the classifiers were performed on *Reddit*. Further, the already trained classifiers were evaluated on the external corpora too. Moreover, various feature selection methods were also explored: *ReliefF* [120], *Randomized Lasso* [216], *Recursive Features Elimination* [86] and selection of features based on the Linear SVM model [155]. Due to the class imbalance of the *Reddit* corpus—e.g. *assert* represents roughly a half of the data, while some classes refer to only few dozens of examples, the experiments were performed also by oversampling the instances for the least represented target classes.

The measures were *macro* scores, meaning that each score (precision, recall, F1-score, hamming score) was computed for each individual class and all the obtained scores were averaged to give a final result. In this way, the classifier performance was assessed by giving equal weight to each speech intention, despite the imbalanced number of instances belonging to each class. Moreover, for the first and second scenarios, the Kappa score was also computed. For statistically comparing the results obtained by the classifiers, t-test was used for each pair of classifiers, across each fold and feature group combination. The significance level was set to 0.05.

5.5 Results

With regard to the optimization of the classifier parameters, the best predictions with Random Forest were achieved using 15 trees. The regularization parameter was optimized for the other two classifiers and the best results were obtained when the regularization parameter was set to 0.5 for Linear SVM and to 3.5 for Logistic Regression (with the optimal number of iterations between 8 and 10). Further, the feature selection methods did not improve the results significantly. Similarly, no improvement was noticed with oversampling. Hence, the results are further reported without any feature selection or oversampling methods.

Table 5.5 presents the classification results obtained by the selected classifiers for all speech act types and for all speech intentions on the *Reddit* dataset. Thus, it corresponds to the first setup consisting of multi-class classification. The reported scores are macro precision (P), recall (R) and F1-score (F1) in percentage, and Kappa score (κ). The first column marks different combinations of feature groups. Then, for each feature group, the first line contains the classification results for all speech intentions, while the second line presents the classification results for all speech act types. For instance, the F1-score obtained by Logistic Regression using only the Discourse features (the row starting with {1}) for all the speech intentions is 41.5 and for all speech act types 62.2. The results show that the best results obtained with Logistic Regression and Linear SVM are similar and significantly better than the best results obtained with Random Forest—best results for each classifier are in bold. For this reason, for the second and third scenarios, Logistic Regression was chosen. A possible explanation for the lower performance of Random Forest is that this classifier may tend to overfit the training data if the classes are not well separated.

The best overall Kappa score for speech intentions ($\kappa = 0.54$) is similar to the average Kappa

Table 5.5: Precision (*P*), Recall (*R*) and F1-score (*F1*) —macro values, in percentage—and Cohen's Kappa score (κ) obtained by Logistic Regression, Linear SVM and Random Forest in classifying intentions and speech act types on the *Reddit* corpus. For each cell, the first line contains the score for intention classification and the second line for speech act type classification. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best F1 and κ scores of each classifier in predicting the intentions and speech act types are in bold. Best overall F1 and κ scores are underlined.

Features	**Logistic Regression**				**Linear SVM**				**Random Forest**			
	P	*R*	*F1*	κ	*P*	*R*	*F1*	κ	*P*	*R*	*F1*	κ
{1}	42.5	41.8	41.5	0.40	47.2	41.3	43.0	0.34	52.9	35.5	40.2	0.27
	72.3	56.3	62.2	0.40	67.6	62.4	63.5	0.44	58.9	51.4	53.9	0.42
{2}	69.0	49.7	55.4	0.44	68.8	50.3	55.6	0.42	69.3	47.5	**52.2**	0.40
	67.6	55.9	60.2	0.44	68.5	52.6	58.0	0.31	73.0	56.1	62.1	0.43
{3}	8.3	8.6	6.0	0.01	8.3	8.6	6.0	0.01	23.0	16.7	18.1	0.10
	20.2	22.0	17.6	0.07	20.4	22.6	18.5	0.09	43.9	31.1	34.3	0.09
{1,2}	68.9	54.0	**56.9**	**0.51**	60.3	54.9	55.6	0.53	59.7	43.3	46.2	0.43
	9.5	66.1	**71.3**	0.57	75.8	67.4	70.8	0.49	84.4	64.2	**70.8**	**0.54**
{2,3}	67.1	51.1	**56.9**	0.47	62.2	49.6	53.4	0.45	54.6	39.7	44.5	0.34
	72.5	57.2	61.8	0.47	65.3	57.1	60.2	0.37	67.5	62.5	63.2	0.45
{1,3}	49.0	35.4	40.1	0.32	43.6	41.7	42.0	0.29	48.1	38.1	40.5	0.33
	72.0	60.2	64.8	0.43	71.7	66.6	67.9	0.46	56.7	51.4	53.3	0.40
{1,2,3}	71.6	53.1	**56.9**	0.49	59.9	56.5	**57.4**	**0.54**	72.4	45.0	50.0	**0.44**
	79.6	66.0	**71.3**	**0.58**	77.4	68.4	**72.1**	**0.52**	70.3	56.0	60.1	0.45

score obtained in the manual classification experiments ($\kappa = 0.57$), while per speech act type is much lower (0.58 compared to 0.80). Additionally, it could be noticed that systematically, the scores for speech act types are higher that those for speech intentions. Possible reasons may be:

- the complexity introduced by the larger numbers of speech intentions (13) compared to the number of speech act types (5),
- the low support of instances associated with some speech intentions and
- the inability to differentiate some speech intentions belonging to the same speech act type based on the defined features.

Some speech intentions frequently co-occur, including speech intentions belonging to the same speech act type (e.g. *complain* and *assert* in 31% of the occurrences of *complain*; *agree* and *assert* in 14% of the occurrences of *agree*). Hence, their differentiation is more challenging. Although not presented in Table 5.5, the micro scores were also analyzed and they were much higher compared to the macro ones, suggesting that indeed the least popular labels were poorly classified. The obtained confusion matrices in this multi-class classification also revealed that common confusions occurred between *assert* and *complain* or *rejoice*, and between *assert* and *agree*.

Furthermore, *Discourse* and *Content* features appear as highly more predictive than the *Conversation* ones. However, compared to the previous experiments for annotating public tweets with speech intentions presented in Chapter 3, *Discourse* features do not longer seem to lead to better results than the *Content* ones. Though, the combination of both results in classification performance similar to the best scores obtained when all feature groups are used.

The previous results are not informative with regard to the identification of each speech intention. However, the second scenario consisting in training binary classifiers for each class enabled this analysis and the results are presented in Table 5.6. The macro F1-scores (in percentage) and the κ scores are reported for Logistic Regression, trained with different configurations of features (marked in the first row), in an one-versus-all approach. The best F1-score varies between 54.5 (for *agree*) and 100 (for *thank* and *other*) and the best κ score between 0.52 (for *complain*) and 1 (for *thank* and *other*). At least substantial κ scores are obtained for 10 out of 13 speech intentions and the rest are moderate.

When compared to the manual classification, the automatic annotation leads to similar κ scores for most classes. Significant differences are observed too: the classifiers perform worse than the human annotators for *complain* and *suggest*, but outperform them for *engage* and *wish*. The results also show that *Discourse* and *Content* features are often the strongest predictors. However, for some speech intentions (*agree*, *guess*, *apologize*, *thank* and *greet*), just the *Content* features alone could lead to the best results. Also, for some other speech intentions (*assert*, *rejoice*, *thank*, *direct*, *suggest*), using only Discourse features achieves results close to the best ones. Another interesting aspect is that the combination of *Content* and *Conversation* features lead to the best results for two speech intentions: *engage* and *suggest*. Thus, as it was intuited in the feature design, the position of turns in conversation and the turn text-based similarity may be attributes of some speech intentions.

Finally, when compared to the previous multi-class scenario, the results obtained by training separate binary classifiers are much higher. Several reasons are identified. Let's assume that an utterance is labeled with *suggest* and *rejoice* and the class *suggest* is less frequent than *rejoice*. In this situation, the utterance will have associated just *suggest* for the first scenario because the strategy is to select the least represented class in order to transform the corpus in a single-label one. Then, let's suppose that, in the multi-class classification with Logistic Regression, a F1-score of 0.65 is obtained for *suggest* and a F1-score of 0.70 for *assert*. Then, the utterance will be labeled with *assert* (0.70 > 0.65), although a binary classification would have yielded true for *suggest* too (0.65 > 0.5, the threshold of the logistic function). Additionally, let's assume another situation. The multi-class Logistic Regression predicts instead *rejoice* as the best result. However, this label will be counted as a false positive because the utterance is known as being only *suggest*, despite the fact that, in reality, the utterance is *rejoice* too. Consequently, training binary classifiers for each speech intentions appears as a better approach among the two.

In the dataset, each sentence has associated up to three labels. Out of 2280 sentences, 1904

Table 5.6: Per-intention binary F1-scores (in percentage) and κ scores obtained by the Logistic Regression classifier in predicting speech intentions on the *Reddit* corpus, in an one-versus-all setup. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best scores for each speech intention are in bold.

Features	{1}		{2}		{3}		{1,2}		{2,3}		{1,3}		{1,2,3}	
	F1	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ
assert	77.1	0.54	78.7	0.55	67.6	0.14	**80.5**	**0.60**	79.3	0.56	76.1	0.53	79.8	0.58
agree	20.0	0.19	**54.5**	**0.54**	0.0	0.0	**54.5**	**0.54**	40.0	0.39	22.2	0.22	**54.5**	**0.54**
guess	38.1	0.36	**61.5**	**0.59**	0.0	0.00	53.8	0.51	**61.5**	**0.59**	30.0	0.28	51.9	0.49
complain	44.8	0.36	50.8	0.44	0.0	0.0	**57.6**	**0.52**	52.3	0.45	42.9	0.33	48.5	0.40
rejoice	56.0	0.54	59.3	0.57	0.0	0.00	75.0	0.74	66.7	0.64	58.3	0.56	**80.0**	**0.79**
apologize	0.0	0.00	**66.7**	**0.66**	0.0	0.00	**66.7**	**0.66**	**66.7**	**0.66**	0.0	0.00	**66.7**	**0.66**
thank	90.9	0.91	**100**	**1.0**	0.0	0.0	**100**	**1.0**	**100**	**1.0**	90.9	0.91	**100**	**1.0**
wish	50.0	0.49	83.3	0.83	0.0	0.00	**85.7**	**0.85**	83.3	0.83	61.5	0.60	83.3	0.83
greet	0.0	0.00	**66.7**	**0.66**	0.0	0.00	**66.7**	**0.66**	**66.7**	**0.66**	0.0	0.00	**66.7**	**0.66**
direct	85.7	0.85	75.0	0.73	10.5	0.10	87.2	0.86	66.7	0.65	85.7	0.85	**88.2**	**0.87**
engage	57.1	0.56	66.7	0.66	0.0	0.00	66.7	0.66	**76.9**	**0.76**	61.5	0.60	66.7	0.66
suggest	53.7	0.49	55.8	0.51	0.0	0.00	59.5	0.56	**63.6**	**0.60**	50.0	0.45	61.1	0.58
other	75.0	0.75	66.7	0.66	0.0	0.00	75.0	0.75	57.1	0.57	**100**	**1.0**	**100**	**1.0**

sentences have only one label, 359 two labels and 17 sentences three labels. It is known now that the binary classifiers outperform the multi-class classifiers. However, a question is still left: to what extent all known speech intentions of an utterance can be identified? To answer this, the third scenario was designed and, in Table 5.7, the results of the speech intention classification obtained in this multi-label setup are presented.

The *Hamming loss* can be interpreted as a label-based accuracy. Similar to the previous results presented in Table 5.5, *Discourse* and *Content* features contribute the most to the correct prediction. Also, the Hamming loss obtained only with the *Discourse* features are similar to those obtained with the *Content* ones. The best F1-score obtained for the feature group {1,2} is slightly lower than the one in the multi-class settings (previous best F1-score 57.4 compared to the current one of 51.6). Overall, it appears that the identification of all speech intentions known for a sentence is satisfactory, but can still be improved. This performance may have been influenced by the fact that the model in the multi-label classification is much more complex: counting only all existing combinations of 1 and 2 labels results in 65 possibilities; also, some unique combinations contain just few instances.

Table 5.7: Hamming and macro F1-scores obtained by Logistic Regression in predicting intentions on the *Reddit* corpus, in a multi-label setup. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best Hamming and F1-scores are in bold.

Feature group	{1}	{2}	{3}	{1,2}	{2,3}	{1,3}	{1,2,3}
Hamming loss (%)	46.7	46.6	33.8	52.9	47.5	46.1	**53.5**
F1-score (%)	37.4	47.8	4.8	**51.6**	49.5	36.4	49.9

The related works proposing supervised solutions to annotate forum conversations with speech intentions expose also a high variance in results depending on the class. The F1-scores presented by Bayat et al. [13] vary between 0 and 0.84, with an overall result of 0.74. Bhatia et al. [14] obtain an overall F1-score of 0.70 and, per class, between 0.15 and 0.85. The solution proposed by Qadir and Riloff [163] has the lowest F1-score of 0.21 and the highest of 0.94. In all these related works, it is not specified if the F1-scores are reported as micro or macro. Kim et al. [118] report the best micro-averaged F1-score of about 0.75, obtained with structural learners (CRF and SVM-HMM). Ferschke et al. [64] have the best macro F1-score of 0.73 and, for each class, between 0.52 and 0.93. Apart from [118] and [64], who predict a number of classes comparative to the current taxonomy, all the others target much less—around 4-8 classes.

In Chapter 3, an analysis of the most predictive features related to each speech intention was conducted. Those results showed that *Discourse* features led to better performance very frequently and they were also correlated with the *Content* ones. Nonetheless, in the current work, the *Discourse* features alone appear to be less often the strongest attributes of the classes and, actually, lexical cues expressed as n-grams are frequently more useful. For getting more insights, the top predictive features for each speech intention were extracted using the feature selection algorithm ReliefF [120] with Logistic Regression. The results are very informative and with high

potential for future work:

1. For all speech intentions, the majority of the most predictive *Discourse* features are related to POS n-grams. Apart from these, the length of utterances, the frequency of pronouns, the features related to sentiment analysis and to verbs (e.g. past tense verbs, verb position) are also highly predictive.

2. Some *Discourse* features, a part of these appearing also as strong predictors in the previous experiments on public tweets, are important in the current experiments (e.g. the presence of imperative verbs or question marks for directive utterances, the presence of future verbs for *wish* or of verbs ending in "-ing" for *engage*).

3. Among the top *Content* features, domain-generic cues are identified to convey the specific speech intentions. For instance, for *greet* "hi, hey, hello" are selected; for *guess* "possible, possibly, guess, not sure, could be", for *thank* "thanks, thank you, appreciate", for *rejoice* "happy, glad, I love, awesome", for *wish* "hope, wish you, hopefully, good luck", for apologize "sorry, apologize". However, apart from these, there are also terms specific to the domain (e.g. "lupus", "chamomile", "butterfly/malar rash"). Future work could tackle the inclusion of the generic cues in *Discourse* features as they comply with the definition, being linguistics cues of speech intentions.

The identification of most speech intentions achieves satisfactory results ($\kappa \geq 0.6$ for 10 out of 13 classes, in the second scenario). However, improvements are also searched for the rest. Two directions to ameliorate the classification for *agree*, *guess* and *complain* are further investigated. First, in the corpus, several speech intentions very often occur with others. Therefore, it is hypothesized that trying instead to predict pairs of classes may improve the results. For this, the results of the multi-label classification for adjacency pairs are analyzed. *Guess* can be similarly or better identified when is part of an adjacency pair with *thank* (F1-score=71.2), with *wish* (F1-score=61.1) and with *direct* (F1-score=62.2). Other slight improvements are observed for *assert* with *thank* and for *greet* with *thank* (F1-score=82.7 and 68.5, respectively). Nonetheless, this direction does not seem to show any advantage for the identification of *agree* and *complain*.

Therefore, a second strategy is formulated. Inspired by the work of Jeong et al. [106], the corpus is augmented with instances from external datasets and the experiments are re-run. The results are presented in Table 5.8. *Agree*, *guess* and *engage* instances come from *SWBD* dataset, while *rejoice* and *complain* from *Bhatia*. An instance in *SWBD* could be an utterance similar to *Reddit* or part of an utterance, while in *Bhatia* an instance is composed of multiple sentences[3]. In this case, the classification of *agree* and *guess* becomes almost perfect and of *engage* is significantly improved. However, for *complain* and *rejoice*, similar results are obtained.

Table 5.9 presents the performance of Logistic Regression trained on *Reddit* and evaluated on *Bhatia* and *SWBD*. The obtained results are lower than those reported in the previous works

[3] In this case, the post is considered to have associated the target speech intention among others.

Table 5.8: Per-intention binary F1-score (in percentage) and κ score obtained by the Logistic Regression classifier in predicting the selected speech intentions on the *Reddit* corpus augmented with instances from *Bhatia* and *SWBD*. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best scores for each intention are in bold.

Features	{1}		{2}		{3}		{1,2}		{2,3}		{1,3}		{1,2,3}	
	F1	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ
agree	95.5	0.86	96.1	0.88	98.9	0.97	97.6	0.93	99.5	0.99	99.3	0.98	**99.7**	**0.99**
guess	93.6	0.88	95.0	0.91	96.8	0.94	95.6	0.92	98.1	0.96	98.6	0.97	**99.2**	**0.98**
complain	46.7	0.40	50.0	0.44	0.0	0.00	53.1	0.46	48.6	0.39	46.9	0.39	**58.0**	**0.51**
rejoice	51.9	0.49	57.1	0.55	0.0	0.00	75.0	0.73	59.5	0.56	48.0	0.45	**81.1**	**0.79**
engage	78.9	0.77	88.9	0.88	77.4	0.76	89.5	0.89	92.3	**0.92**	90.0	0.89	**92.7**	**0.92**

Table 5.9: Macro F1-scores (in percentage) obtained by the Logistic Regression classifier in predicting intentions on *Bhatia* and *SWBD* corpora. From these corpora, only the classes that could be aligned with the current taxonomy were selected. For each cell, the first line contains the score for speech intention classification and the second line the score for speech act type classification. Feature groups {1,2,3} correspond to *Discourse*, *Content* and *Conversation*, respectively. Best scores are in bold.

Features	{1}		{2}		{3}		{1,2}		{2,3}		{1,3}		{1,2,3}	
	F1	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ	*F1*	κ
Bhatia	9.8	0.11	5.6	0.03	**39.8**	**0.54**	8.1	0.13	7.1	0.06	9.1	0.12	9.2	0.14
	26.1	0.12	23.2	0.06	**39.3**	**0.46**	26.1	0.12	22.5	0.06	26.9	0.15	27.2	0.16
SWBD	14.6	0.19	**29.0**	**0.30**	12.2	0.04	27.3	0.28	19.5	0.26	11.1	0.15	19.9	**0.30**
	30.6	0.33	28.3	0.21	10.7	0.00	**32.5**	**0.34**	19.5	0.07	27.1	0.30	29.0	0.32

using these datasets. In *Bhatia* case, this may be due to the presence of some other features such as the user's number of comments and authority score, which appear also among their most predictive features. However, this type of features were avoided in the current work because they are challenging to acquire, in particular, when the environment does not enforce authentication, such as in many discussion boards, or when these statistics are personal (e.g. the case of emails). Moreover, it is not clear if the reported F1-scores are micro or macro—the current micro scores are also higher because the most frequent classes are much better classified. Similar to the results reported by Bhatia et al. [14], the standalone *Conversation* features are the most predictive. This appears in contradiction with the results reported on *Reddit*. It may be that the unit choice, sentence versus complete turn, plays a role and suggests that the *Conversation* features may become relevant if instances are turns instead of sentences.

Similar to the previous discussion, it is claimed that the results for the *SWBD* corpus could have been improved by including other types of features used in the related works, referring to acoustics, prosody or the identity of the speaker. However, the fact that *agree*, *guess* and *engage* intentions are almost perfectly identified (Table 5.8) when examples from *SWBD* were included in classifier training shows that the current corpus may not sufficiently capture the diversity of expressing certain intentions; thus, the classifiers are unable to recognize intentions expressed much more differently. Furthermore, the *Discourse* and *Content* appear as the most predictive, similar to the experimental results on the *Reddit* corpus—in *SWBD* an instance is also an utterance.

5.6 Conclusion

In conclusion, varied supervised machine learning experiments were implemented in order to assess the extent to which sentences from conversations could be correctly annotated with the extended classes of speech intentions. A new type of asynchronous communication was thus explored in the current work: forum conversations, which are very similar to email conversations. New features were identified, compared to those proposed in Chapter 3, and integrated in the classifiers. Various setups were designed and their performance was compared: multi-class, one-versus-all and multi-label classifications. The highest scores for each speech intention were obtained in the one-versus-all setup. Overall, the solution led to promising results–the κ scores for most speech intentions were at least substantial, although some intentions still required further effort for a better identification. Specifically, two strategies to improve the identification of the poorly classified classes were proposed. The first strategy was to use a multi-label classification for frequently co-occurring intentions. The second strategy consisted in augmenting the training corpus with instances of the least represented speech intentions from external, even heterogeneous corpora.

Multiple *Discourse* features proved relevant for the discovery of several speech intentions.

However, the group alone was not sufficient to obtain the best results. Instead the best classifier performance was derived when the *Discouse* features were used together with the *Content* ones. The detailed feature analysis revealed though that the most predictive *Content* features were mainly generic and dependent only on the language, but not on the corpus or on the domain. The *Conversation* features were highly useful when predicting instances represented as turns, although they did not show predictive power when instances were sentences.

The generalization of the classifiers is rarely addressed in literature, mostly because of a lack of other compatibly labeled corpora. However, in the current work, to the extent to which it was possible, the generalization and robustness of the best classifier was assessed on other corpora, including synchronous communication. These experiments revealed satisfactory robustness of the models for the discovery of a subset of intentions, in particular, as the models were trained on a specific health-related Reddit topic, with sentence-based instances and then evaluated against other domains and conversation formats. However, testing the intention classifiers on external corpora also showed that the generalization could not be claimed yet, as the corpus did not capture high variance in expressing some intentions and as the choice of instance representation—part of sentence, sentence, turn—had an impact on the importance of different feature groups.

Finally, compared to all related supervised works to automatically annotate asynchronous communication with speech intentions, the current solution reveals the most detailed output. Each sentence is annotated with multiple labels. Similarly, another work [163] labeled sentences with multiple classes from the Searle's speech act types. However, the resulting annotation from the current work is not more detailed only because of the multi-label classification at the sentence level, but also due to the fine-granularity of the proposed classes to represent speech intentions.

5.6.1 Study Limitations

The threats to validity are mostly the same as those presented in Chapter 3, Section 3.3.2. Additional threats to validity are discussed further. Also, strategies to mitigate threats not handled in the previous iteration are described.

The *internal validity* could have been influenced by the characteristics of the ground-truth corpus and the experimental setups. The ground-truth corpus was prepared in a similar approach as for the experiments on public tweets. Similarly, the classes had a number of instances varying from few dozens to few hundreds. To better assess the impact of class imbalance on the results, instances of the least popular classes were added from external corpora, the experiments were then repeated on the augmented data and the results compared. In this way, it was shown that the classifiers performed similarly for some speech intentions, even though they were exposed to more data (e.g. *complain*), while for some other speech intentions the initial results were not conclusive, more data leading to better classification results (e.g. *agree*, *engage*).

Moreover, with regard to the experimental setup, this time, macro scores were used to equally account for each class despite its support and the Kappa statistical test was computed too. In

order to avoid reaching conclusions biased by overfitting, cross validation was implemented again. Additionally, an improvement was brought by including a randomized stratified data selection scheme to cross-validation. In this way, the test set had instances of each speech intention, in each fold. Then, the parameters of the classifiers were tuned on the training data of each fold, compared to the previous experiments in Chapter 3, when the default parameters were used. Also, the effects of oversampling and feature selection methods on results were analyzed.

The *external validity* was assessed in the current experiments by putting to test the best classifiers trained on the *Reddit* data on two other external and heterogeneous corpora. *SWBD* contained synchronous conversations; a classification instance could be a part of a sentence in *SWBD* or composed of multiple sentences in *Bhatia*, the topics discussed in these conversations were different. However, the threats to external validity were not entirely mitigated as the experiments were conducted only for some speech intentions, for which an alignment to the classes of the external corpora was possible. Moreover, some results, such as the importance of the feature groups in prediction, appeared to be inconsistent across these corpora.

For the concept modeling part of *construct validity*, an experiment to assess if the perception of the speech intentions was consistent among multiple human annotators was conducted and presented in Chapter 4. Widely spread statistics for inter-rater agreement were used: the Cohen's and Fleiss' Kappas. The ground-truth corpus was derived then from the data used in the manual annotations by taking into account the validation results. However, not all the taxonomy classes were included in the validation—*require* and *declare* were excluded because of their absence in the selected corpus, or passed validation (e.g. *sustain*). Nonetheless, the final speech intentions selected to be used in the machine learning experiments were validated, either experimentally or through additional expert interventions.

The *conclusion validity* was supported by statistically comparing the results obtained by various classifiers with t-test (significance level set to 0.05). Moreover, as mentioned in the paragraph on internal validity, multiple measures were taken to ensure that valid conclusions were reached including: the evaluation of the classifiers using cross-validation and train-test split, the analysis of the effects of feature selection and balancing techniques and of the use of various parameters with the classifiers on the results and the scaling of features.

5.6.2 Directions for Further Research

The first proposed direction for further research is to explore and assess different ways to improve the identification of all speech intentions from utterances of varied types of asynchronous corpora. The improvement could stem from multiple actions:

- In the experiments, a relatively small dataset was used, in which not all the classes were sufficiently represented. The data imbalance proved to affect the performance of the classifiers and, in general, it represents one of the main challenges in automatic classification. Despite the best efforts, manually labeling the data is labor intensive. Other

options for this task include using crowd-sourced annotations [10, 150] or integrating instances from other datasets when class alignment is possible [106]. The second strategy was already taken into account in the current work, but the first is a promising future work direction to annotate, in particular, other subreddits and other conversations with the proposed speech intentions.

- The experiments revealed that the current features were not enough discriminative for some classes. To address this issue, more and especially more diverse instances for each speech intention should be collected. Moreover, based on manual and automatic corpus analysis and theoretical literature, new features can be introduced and their effectiveness evaluated. Better results may be expected by using pre-trained word embeddings in the classification and deep learning techniques, relying on the networks to independently learn features. However, a requirement of these approaches is the existence of large, labeled dataset. Additionally, these claims need to be substantiated via further experimentation.

- Regarding the overall solution, past works showed that semi-supervised classification could be promising for speech act discovery, being able to learn from fewer labeled examples [106]. Also, using multiple classifiers in cascade, starting with one for the least represented class in the training corpus, led to improved results [149]. These strategies and the use of structural learners should be explored in the future works too.

Then, the current approach should be integrated in practical solutions for health information seeking and for health compliance-gaining technologies. The solution in the current state can be included in health information seeking tools to allow for content filtering by speech intentions, apart from the topic. Additionally, approaches analyzing health social media with sentiment analysis [231] could be augmented to include richer information through speech intentions. Thus, the solution usefulness to and usability in medicine can be thoroughly evaluated in practice.

The impact of classifier performance on these practical applications varies. When searching for posts with suggestions or positive attitudes, results containing also false positives may not be disrupting for a health information seeker. However, for studying persuasion discourses and draw knowledge on strategies of persuasion, highly accurate results are needed. Nevertheless, there is also the case when the knowledge is used directly in persuasion and compliance-gaining solutions. Then, the derived persuasion strategies could be tested in practice and re-adjusted according to the real-time impact on or feedback from interlocutors. Moreover, if only the most common verbal behavior generated from crowd conversations is exploited, then the chance of repetitive false positive intentions is rather low with the current state of the classifiers.

Finally, with regard to the established goals and properties of an envisioned solution to be created through the current thesis, the process view of conversations is still not addressed. The current circumstances, as in the ground-truth corpus that contains conversations and the

satisfactory results of the classifiers, enable now the pursuit of this objective, presented in the chapter to follow.

CHAPTER 6

MODELING CONVERSATIONS AS PROCESSES OVER SPEECH INTENTIONS

Epure, E.V., Zitnik, S., Compagno, D., Deneckere, R., Salinesi, C. (2017). Automatic analysis of online conversations as processes, presented at Journées Analyse de Données Textuelles en Conjonction avec EDA 2017, Lyon, France.

Epure, E.V., Compagno, D., Salinesi, C., Deneckere, R., Bajec., M., Zitnik, S. (2018). Process Models of Interrelated Speech Intentions from Online Health-related Conversations. *Artificial Intelligence in Medicine*. doi = https://doi.org/10.1016/j.artmed.2018.06.007

Contributions: E.E.V. designed the method, designed and conducted the experiments and wrote the article, C.D. participated in the observational study, S.C, D.R., R.B., Z.S. and C.D. provided feedback on the article.

The contribution of this chapter is:

- The first time when business process mining is put to test on mining conversations annotated with multiple speech intentions per sentence (and the second time generally known). To enable the use of process mining techniques in this context, an algorithm to generate event logs from annotated conversations is designed. Additionally, an extensive rationale design that discusses how to map conversational concepts on event log concepts and the impact of the corresponding decisions on the output process models is provided.

The third objective of the current work is to discover processes of interrelated speech intentions from asynchronous conversations. Specifically, this chapter tackles the third proposed

research question: *How to automatically discover processes of interrelated speech intentions from asynchronous conversations independently of the domain and corpus characteristics?* (**RQ3**). To answer this research question, the first research question is revisited with a focus on the relations among speech intentions: *How to formalize conversations with comprehensive and corpus-independent speech intentions and process relations?* (**RQ1**).

Only a small part of the related works addresses this problem and, in practice, two strategies are observed. The first strategy is to discover frequent sequences and transition diagrams over speech acts from already annotated corpora [28, 178, 189]. The second strategy, employed in unsupervised machine learning approaches, is to jointly identify the speech act classes and their relations with Hidden Markov Models [154, 158, 171]. One disadvantage of the second approach is that it leads to poorer results for class identification compared to supervised learning. Although, as an advantage, there is no effort put in the ground-truth corpus creation. However, as outlined at the end of Chapter 2, more complex relations could also exist to represent processes and capturing just the immediate precedence is a simplification rather valid for synchronous conversations: the asynchronous ones are threaded and have much longer turns—thus, no single turn sequence per conversation or speech act class per turn.

Additionally, one question was raised at the end of Chapter 2 regarding the debate that exists in conversation analysis on the existence of relations among speech intentions. Instead of assuming true the existence of such relations and use this knowledge in machine learning solutions as contextual features, what if the existence is proven in the first place with the help of computer science methods?

Driven by the limitations of the existing works and the raised question, the current chapter proposes to discover relations from already annotated corpora with *process mining*—thus, separating the speech act identification from relation identification. Process mining offers techniques to discover and model human behavior from digital traces, generated during the interaction with information systems. The idea grounding this research is that process mining can be suitable to model conversations too. Communication represents human behavior and corpora of written or spoken conversations can be seen as traces of this behavior.

Thus, regular patterns may be inductively extracted from such conversation corpora, revealing relations among speech intentions if they exist. The advantages of process mining compared to other automatic analytic approaches are that the obtained models are visual and the techniques and tools are highly interactive [1]. This could enable easier exploration of conversations, in particular in multi-disciplinary research settings, as in-depth technical knowledge is not required. However, the process mining techniques rely on well-defined, representative and structured event logs as input in order to discover process models. Consequently, an approach to map conversations on such event logs that capture the relevant aspects of conversational behavior must be proposed.

To better understand this requirement, Section 6.1 presents in more detail what process mining is, how process mining techniques work and what are the expected event logs. Then, the

overall approach to obtain processes over speech intentions and the specific decisions made on how to semantically map the process mining concepts on conversational data and generate the event logs are described in Section 6.2. Evaluation methods that focus on demonstrating the relevance of the approach are presented and motivated in Section 6.3. Finally, the results of the evaluation are summarized in Section 6.4. The results show that the obtained process models and the proposed approach can reveal valuable insights for behavior and conversation analysis and could support further investigations in medicine and linguistics.

6.1 Introduction to Process Mining

Many organizations structure their core missions as business processes. A process is a collection of related tasks, also named activities, which, when executed in a certain order, lead to the fulfillment of various business goals such as the delivery of services or products. The representation of business processes as graphical models has been long recognized as beneficial, giving organizations the opportunity to analyze, improve and ultimately automate their business processes. Moreover, these representations used to be largely defined by business analysts or by domain experts manually until the emergence of *process mining* [1].

Process mining proposes a suite of techniques, methods and tools to exploit the digital traces generated by users during the interaction with information systems in order to extract, visualize and analyze *processes*. In human practices supported by information systems, user tasks are often traced for different reasons, such as for monitoring software (e.g. logging exceptions raised during software use) or for analyzing and monitoring transactions (e.g. computing key performance indicators in an ERP system). Researchers and practitioners identified the opportunity to use the already existing tracing mechanism to automatically build representations of organizational processes from data, leading to the creation of the new domain of process mining [1, 46, 114].

Let's further describe in more details what a process is [1]. A process consists in a collection of activities and their relations. The relations define the order in which various activities can be performed to reach a goal. Formally, a process model can be represented by a quadruple, $P = (S_{\text{start}}, S_{\text{stop}}, A, T)$, where $S_{\text{start}} \neq \emptyset$ is the set of initial process states—marks the process beginning; $S_{\text{stop}} \neq \emptyset$ is the set of final process states—marks the process end; $A = \{a_n\}_{n=1}^{N}$ is the finite set of process activities; $T \subseteq ((S_{\text{start}} \times A) \bigcup (A \times A) \bigcup (A \times S_{\text{stop}}))$ is the finite set of transitions between activities or activities and states. Process models capture several kinds of relations among activities: *sequence* (activity a_i follows activity a_j), *concurrency* (both activity a_i and activity a_j occur, possibly in overlapping time), *decision* (activity a_i or a_j is performed at a certain point in time) and *loop* (activity a_i is enacted multiple times in a row). Using the transition notation previously introduced, these activity relations can be defined as [1]:

- sequence: $t \in T : t = a_i \prec a_j$ or $t = s_{\text{start}} \prec a_k$ or $t = a_l \prec s_{\text{stop}}$ with $s_{\text{start}} \in S_{\text{start}}$, $s_{\text{stop}} \in S_{\text{stop}}$, $a_i, a_j, a_k, a_l \in A$ and $i \neq j$. The interpretation of $a_i \prec a_j$ is that once the activity a_i is finished,

the activity a_j follows. The transitions $s_{\text{start}} \prec a_k$ and $a_l \prec s_{\text{stop}}$ are implicit transitions marking the start and the end of the process.

- concurrency: $t_m, t_n \in T, t_m = a_i \prec a_j, t_n = a_i \prec a_k$ and $t_m \| t_n$ where $a_i, a_j, a_k \in A$ and $j \neq k$. The interpretation is that both t_m and t_n take place; once the activity a_i is finished, the two activities a_j and a_k follow in any order.

- decision: $t_m, t_n \in T, t_m = a_i \prec a_j, t_n = a_i \prec a_k$ and $t_m \oplus t_n$ where $a_i, a_j, a_k \in A$ and $j \neq k$. The interpretation is that either t_m or t_n takes place, meaning that a decision must be made; once the activity a_i is finished, either the activity a_j or a_k follows.

- loop: $t \in T : t = a_i \prec a_i$ where $a_i \in A$. The interpretation is that once the activity a_i is finished, it can be enacted again.

Depending on the process model formalism (e.g. BPMN [169], Petri Nets [87], Yawl [210]), more detailed concepts exist too. However, the introduced relations are general and frequently identified by the process mining techniques [1].

A process execution $E = s_{\text{start}} \prec \ldots \prec a_i \prec \ldots \prec s_{\text{stop}}$ represents a sequence in the process from a start state to an end state, with at least one activity in between. A process can contain multiple execution paths given the presence of loop, decision and concurrency relations. These different process executions are called process instances, cases or traces. The terminology selected in the current work to refer to an individual process execution is *process trace*. When a process is executed through the use of information systems, the enacted activities trigger events in the systems, which are logged—assuming that a logging mechanism is in place. Hence, a process trace is a succession of logged events generated during the enactment of a sequence of activities.

Let e be such an event captured during the user interaction with a software application for the enactment of a certain activity. The standard information about an event required by a process mining technique as input is the event unique identifier, the event timestamp, the event associated activity and the process trace to which the event belongs, identified by an unique identifier too. Often, the timestamps can miss whether the events are chronologically ordered within a trace, according to their activity occurrence in real-life.

To be noticed that events associated with an activity can be raised multiple times, by multiple users, for different goals. A specific process execution E is logged as a process trace $\tau_E = \{e_1, e_2, \ldots\}$. Let $L = \{\tau_v\}_{v=1}^{V}$ be a log containing a finite set of process traces. The process mining technique is then a function γ that maps the log L on a process model P [1]. For one specific log, multiple definitions of the function γ could be proposed as different algorithms. Also, variations of the same function γ could exist by parameter manipulation [1]. Considering this, different process representations can be discovered, the variations being in particular with respect to the discovered relations. Nonetheless, the most representative relations, captured by frequent patterns observed in the logged process traces, are consistently discovered by any process mining algorithm.

Apart from discovering process models, an effective visualization is central to process mining. Compared to other data mining techniques, process mining aims to be a bridge for the production of explicit, high-level and easy to interpret models. For this, several techniques have been aimed, not only at defining the function γ, but also at proposing ways to enable a better visualization of process models, while maintaining a fit to the observed behavior in the log L.

Heuristic Miner [220] and *Fuzzy Miner* [85] are two popular process mining techniques. These are presented in detail below. Nonetheless, many more algorithms exist [1].

6.1.1 Heuristic Miner Algorithm

The underlying assumption of Heuristic Miner is that not all information contained in an event log is always correct or complete. The goal of this technique is to derive the main observed behavior while excluding the noisy data [1, 220]. The input required by Heuristic Miner is an event log L, containing multiple process traces. In a process trace, the events are assumed chronologically ordered and, in this case, the timestamps are not needed. However, if the events are not ordered, the timestamps are required for sorting. The output of Heuristic Miner is a *dependency graph*, which is a graph with activities as nodes and dependency relations among activities as edges.

Let $a, b \in A$ be two random activities found in the event log L and $act(e)$ a function returning the activity associated with an event e. Six types of relations can be defined between two activities:

1. Direct sequence: $a >_L b$ iff $\exists \tau = \{e_1, e_2, ..., e_n\}$[1], a process trace with chronologically ordered events, and $1 \leq i \leq n-1$ such that $\tau \in L$ and $act(e_i) = a$ and $act(e_{i+1}) = b$.

2. Dependency: $a \rightarrow_L b$ iff $a >_L b$ and $a \not<_L b$.

3. Concurrency: $a \parallel b$ iff $a >_L b$ and $b >_L a$.

4. Neither dependency nor parallelism: $a \#_L b$ iff $a \not<_L b$ and $b \not<_L a$

5. Repeating sequence: $a >>_L b$ iff $\exists \tau = \{e_1, e_2, ..., e_n\}$, a process trace with chronologically ordered events, and $1 \leq i \leq n-2$ such that $\tau \in L$ and $act(e_i) = a$ and $act(e_{i+1}) = b$ and $act(e_{i+2}) = a$.

6. Indirect sequence: $a >>>_L b$ iff $\exists \tau = \{e_1, e_2, ..., e_n\}$, a process trace with chronologically ordered events, and $1 \leq i < j \leq n$ such that $\tau \in L$ and $act(e_i) = a$ and $act(e_j) = b$.

The first step for deriving the dependency graph is to compute the *certainty* of the existence of a dependency relation between each two activities, a and b, identified in the log L. For each pair of activities, the certainty $(a \Rightarrow_L b)$ is computed by taking into consideration the frequencies of the direct sequences between a and b and b and a, as follows:

[1] In a slight abuse of notation, curly braces are used to denote a sequence in the current work.

$$a \Rightarrow_L b = \frac{count(a >_L b) - count(b >_L a)}{count(a >_L b) + count(b >_L a) + 1} \tag{6.1}$$

Then the construction of the dependency graph is based on an *all-activities-connected heuristic*. The underlying principle is that each non-starting activity must have at least another activity as a cause and each non-ending activity must have at least another activity as a dependency. Then, for each activity, its best candidates for causes and dependencies are selected based on the highest dependency scores, computed for the pairs that include this target activity. Thus, a first version of the dependency graph is constructed.

Then, for deriving the next version of the dependency graph, three threshold parameters are introduced: the *dependency* threshold, the *positive observation* threshold and the *relative-to-best* threshold. These thresholds support the decision about which dependency relations are indeed accepted as significant—hence, they should be maintained in the graph, or non-significant—hence, they should be removed from the graph. The dependency threshold is the lower limit of certainty for maintaining a dependency relation. The positive observation threshold marks what is a minimum accepted frequency of a dependency relation ($count(a \rightarrow_L b)$). The relative-to-best threshold enables the selection of dependency relations for an activity relative to its strongest dependency relation. Specifically, if the difference between the certainty of the strongest dependency relation and the certainty of another dependency relation is higher than this threshold, then this latter dependency relation is maintained in the graph. In fact, the graph is updated by keeping only the dependency relations between activities, which comply with all the above conditions and thresholds. Through highly restrictive threshold parameters, Heuristic Miner can discover the most representative behavior. However, low values of these parameters could also lead to the discovery of less frequent and exceptional behavior.

So far, the strategy to compute the dependency measure does not identify the presence of short loops such as loops of length one (e.g. b in the trace $\{a,b,b,b,c\}$) or of length two (e.g. the sequence $a > b$ in the trace $\{c,a,b,a,b,d\}$. For handling this, Heuristic Miner uses two other dependency measures, the first for length-1 loops, the second for length-2 loops:

$$a \Rightarrow_L a = \frac{count(a >_L a)}{count(a >_L a) + 1} \tag{6.2}$$

$$a \Rightarrow_{2L} b = \frac{count(a >>_L b) - count(b >>_L a)}{count(a >>_L b) + count(b >>_L a) + 1} \tag{6.3}$$

The equation to identify the certainty of loops of length 1 is applied first. Then, each dependency relation to itself is represented in the graph only if it complies with the conditions introduced earlier regarding the thresholds. Loops of length two are searched for an activity unless self loops were not identified. Finally, in order to reveal if concurrency relations are present

in the graph, the following dependency measure is computed:

$$a \Rightarrow_L (b \parallel c) = \frac{count(b >>_L c) + count(c >>_L b)}{count(a >>_L b) + count(a >>_L c) + 1} \tag{6.4}$$

Additionally, a *concurrency* threshold is introduced to set the lower limit of certainty for a concurrency to be represented in the graph. Graphically, the concurrency can be shown in the graph by annotating the split from one activity to two or more other activities with the symbol &. Other process meta-models such as Petri Nets [87] or BPMN diagrams [169] would inherently allow the graphical representations of concurrencies and decisions. Finally, for the discovery of indirect dependency relations ($a >>>_L b$), a more complex algorithm is presented in [220] together with a detailed description of Heuristic Miner with examples.

6.1.2 Fuzzy Miner Algorithm

Heuristic Miner handles very well noise and is a very popular process mining technique. However, the results obtained for highly unstructured processes with a large number of activities are hardly readable, despite the manipulation of the process representation through the proposed thresholds [85]. This obtained representation of processes can still expose too many edges, which combined with a large number of classes, results in "spaghetti models" [85].

Fuzzy Miner is a process mining technique designed to discover simplified, but relevant and visual models from traces of highly unstructured processes [85]. Inspired from cartography, the algorithm leverages the level of representative behavior through abstraction—expose higher-level concepts on a map by aggregating lower-level details (e.g. merge areas of a city); emphasis—the more significant a concept is, the more is emphasized through stronger colors and bigger shapes (e.h. highway representation); customization—depending on the context, a map can expose different concepts at different levels of details (e.g. the map for a city versus a political map); and abstraction—low-level details are omitted (e.g. pedestrian paths). The process level of details is interactively set by the user through the interface of the process mining tool which implements Fuzzy Miner. Compared to other process mining techniques, the focus is less on a very complete and accurate representation of the process. Instead, Fuzzy Miner is aimed to be a technique for practical applications, which are highly exploratory, enabling the analysts to interact with the data through the tool and have more detailed or more abstract process views.

Similar to Heuristic Miner, the significant behavior is always represented in the process model—a dependency graph, while less significant behavior is abstracted from the representation. In order to make decisions about the significance, multiple metrics are defined and applied to simplify the graph. Contrary to Heuristic Miner, the visualization is customized and more expressive, by showing the most significant activities with stronger colors and the most significant relations with larger edges. The metrics to quantify significance are introduced further.

Unary significance defines how important an activity is for the behavior observed in the log. This is measured by counting the activity frequency (frequency significance). Additionally, the

routing significance is also computed for each activity as the difference between the number and significance of its preceding activities—incoming edges, and the number and significance of its subsequent activities—outgoing edges.

Binary significance defines how important a sequence relation is based on its frequency and distance significance. The distance significance metric is computed as the difference between the frequency significance of the relation and the significance metrics of the related activities. The intuition is that a sequence is globally significant if it is the most or among the most significant edges for both related activities. Additionally, indirect sequences are also assessed for their significance. The search can start with sequences of length two, which are increased until the values of the significance metrics decrease.

The first version of the dependency graph is constructed by representing activities as nodes and by adding an edge for each sequence observed in the log. Then, transformation methods are applied to simplify the graph. The first step is *conflict resolution* for each pair of activities, a and b, with direct sequences observed in both directions, $a >_L b$ and $b >_L a$. This behavior could show the presence of concurrency, of length-2 loop or of an exception to a dependency relation. In order to decide which scenario to select, the relative significance metric (rel) is defined [85]:

$$rel(a,b) = \frac{1}{2} \cdot \frac{sig(a,b)}{\sum_{c \in A} sig(a,c)} + \frac{1}{2} \cdot \frac{sig(a,b)}{\sum_{c \in A} sig(c,b)} \tag{6.5}$$

where $sig : AxA \rightarrow \mathbb{R}_{\geq 0}$ is the binary significance of a direct sequence.

The relative significance metric is applied for each pair, in both directions ($rel(a,b)$ and $rel(b,a)$). Moreover, a *preserve threshold* is established to mark what is the lowest acceptable value of the relative significance. If both $rel(a,b)$ and $rel(b,a)$ are higher than the threshold, it is assumed that the activities form a length-2 loop. If one direct sequence among the two has a relative significance lower than the threshold, then the absolute difference between the relative significance values is computed as follows (offset):

$$off(a,b) = |rel(a,b) - rel(b,a)| \tag{6.6}$$

Then, a *ratio threshold* is introduced to mark the lower boundary of an acceptable offset value. If the offset value is higher than this threshold, the activities are in a dependency relation, the direction being given by the highest relative significance value. Otherwise, a and b are concurrent, thus both edges are removed from the graph.

The second transformation method, after conflict resolution, is *edge filtering*. Depending on an *edge cutoff* parameter—set interactively by the user, only the incoming and outgoing edges of each activity with binary significance values higher than this parameter are maintained in the graph. Also, the binary significance values are normalized beforehand. Then, the third transformation is *activity filtering*, which occurs in a similar manner as the edge filtering, by considering only

activities with significance values higher than a *node cutoff* parameter, set by the user from the tool user interface.

Finally, the tools implementing FuzzyMiner—ProM [211] and Disco [84], propose a default parameter configuration customized for the given input data. These parameters leverage the precision and completeness of the behavior observed in the log, on the one hand, and the understandability and high-level representativeness of the processes to be used in practice by users, on the other hand.

6.1.3 Discussion

Heuristic Miner appears more suitable for structured processes or for processes with a lower number of activities. Fuzzy Miner is interactive and aimed at exploratory analysis, in particular, for highly unstructured processes. Both algorithms discover *dependency graphs*, which are directed graphs with activities as nodes and process relations as edges[2]. Their main differences consist of the metric framework used to compute the significance / certainty of graph edges and nodes, a more dynamic representation being available with Fuzzy Miner.

Compared to transitions diagrams, dependency graphs discovered with process mining techniques are able to capture more complex relations, such as concurrency, loops of length two (length-2 loops) and indirect sequences. Moreover, to filter out the log noise, multiple dependency measures are defined, customized for specific types of relations. In this way, the process representation varies from showing all observed patterns in the log to showing only the most significant patterns. Ultimately, if the algorithm thresholds are very relaxed or not applied at all and the length-2 loops and indirect sequences are excluded, a dependency graph could likely resemble a transition diagram.

This comparison is purely structural as both dependency graphs and transition diagrams share a graph-like layout with activities as nodes. However, with regard to the semantics, these diagrams are less comparable. An edge in the transition diagram refers to the probability of the source activity to be directly followed by the target activity. An edge in the dependency graph exposes an ordering, which also quantifies the certainty of the transition but with other metrics, and does not necessarily imply a direct sequence.

Consequently, it could be claimed that the process models obtained with process mining techniques are general and more comprehensive than transition diagrams / Markov models. Thus, an answer to the first research question (**RQ1**), addressed with a focus on relations among speech intentions, is provided.

[2]To differentiate between concurrency and decision relations, it is necessary to introduce new graphical elements, besides nodes and edges, in the network representation (e.g. see C-nets [1]).

6.2 Process Mining to Model Asynchronous Conversations

The approach taken in the current work to discover interrelated speech intentions from asynchronous conversations is to use standard process mining. Several reasons for putting process mining to the test exist. First, given logs of representative behavior, process mining techniques are proven to discover useful behavioral models and their correctness have been also assessed (e.g. through benchmarking) [1, 114, 220]. Second, the obtained models are visual and interactive [85]. This can enable an easier exploration of conversations, in particular in multi-disciplinary research settings, as in-depth technical knowledge is not required.

However, given that the process mining algorithms and tools rely on structured, well-defined event logs, a solution to transform conversations in the required input must be proposed. Moreover, the semantics to map conversations on event logs concepts should be defined by taking into account the relevance of the output generated by the process mining techniques to conversation analysis. Process mining techniques applied directly to digital textual traces without log generation exist [57]. Nevertheless, these techniques are unsuitable for the current goal because they imply that the text contains reported behavior [57].

The overall approach is presented in Figure 6.1. According to the theoretical literature on conversation analysis [187], conversations represented as processes are expected to show more often irregularities, hence to be unstructured. Considering this, Fuzzy Miner is chosen as the main technique to be applied in the current work. Additionally, as conversation analysis on real corpora is exploratory, the interactivity provided by the tool Disco [84] implementing a version of Fuzzy Miner is considered an advantage too. Nonetheless, once conversations are transformed in events logs, any other process mining technique or tool can be used too. The following subsection introduces the method designed to transform conversations in events logs, the types of process models that are expected to be discovered and how the mined knowledge should be interpreted.

6.2.1 Generating Event Logs from Annotated Asynchronous Conversations

A first important choice for applying process mining to conversations is to determine what conversational unit (e.g. turn, sentence) is considered as event. The choice of events impacts the output and the model interpretation. In the current work, the objective is to reveal relations among speech intentions. Consequently, the speech intentions are the most suitable choice for activities in applying process mining for the aim of linguistic analysis. In this work, speech intentions are used to characterize individual utterances. Considering that an event is interpreted as an occurrence of an activity, a similar parallel can be drawn between utterances and speech acts: an utterance brings out a speech intention (or more).

Therefore, speech intention classes or speech act types are mapped on process *activities* and utterances are associated with *events*. Applying process mining on this data is equivalent with identifying recurring patterns of speech intentions in posts or conversations. These discourse

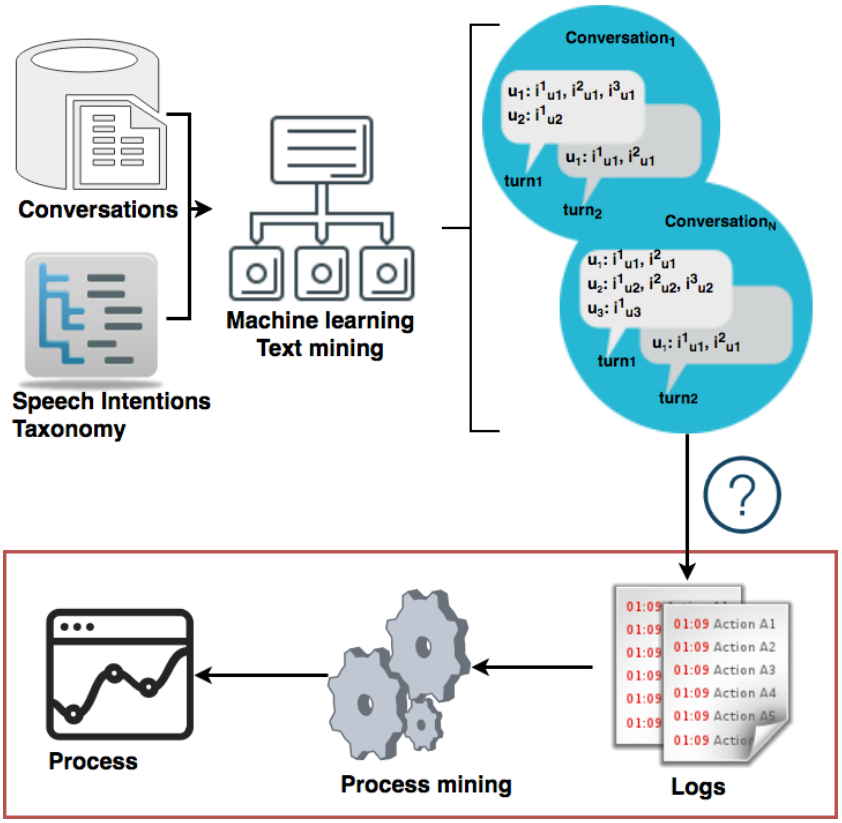


Figure 6.1: The overall approach to model conversations as processes over speech intentions. The current chapter is focused on applying process mining, emphasized in the bottom part of the figure. Existing process mining techniques are reused. The challenge then is how to transform conversations annotated at utterance-level in well-formed logs of representative verbal behavior in order to discover relevant processes for conversation analysis.

regularities are conceptualized through the proposed speech intentions and their relations expressed as process relations. For instance, this approach could enable the automatic identification of frequent adjacency pairs in online dialogues (e.g. questions and answers). However, several important aspects of log construction still need to be tackled: what a process trace should be and how to ensure an one-to-one event-activity mapping considering that utterances may contain multiple labels—hence an event is linked to multiple activities in the same time.

Asynchronous conversations are often composed of concurrent threads. Different users can reply to the same initial post. Each of these replies can further develop in an individual thread, giving a tree-like structure to the whole conversation. A conversation cannot be mapped on a trace because there is no linear exchange of turns, which could be transformed in a sequential flow of events. However, conversation threads are linear. Therefore, let's suppose instead that each thread is mapped on a process trace.

Each individual turn is composed of a sequence of utterances and for simplicity, let's assume that each utterance is labeled with a single speech act. For instance, considering that

$Conversation_1$ and $Conversation_N$ in Figure 6.1 have a single thread each and that each utterance is characterized by the first speech intention only, then two process traces are generated $\tau_1 = \{e_{C_1-t_1-u_1}, e_{C_1-t_1-u_2}, e_{C_1-t_2-u_1}\}$ and $\tau_N = \{e_{C_N-t_1-u_1}, e_{C_N-t_1-u_2}, e_{C_N-t_1-u_3}, e_{C_N-t_2-u_1}\}$, where $e_{C_j-t_k-u_l}$ is an event associated with the first speech intention of utterance u_l of turn t_k in conversation C_j. Both process mining algorithms presented in Section 6.1 derive the models by looking at direct sequences. While all sequences presented in these two traces are correct, the sequences between two different turns are not complete. Between two consecutive turns, only the sequence from the last utterance of the former to the first utterance of the latter is captured (e.g. $e_{C_1-t_1-u_2}, e_{C_1-t_2-u_1}$ in τ_1). However, linguistically, direct sequences should be defined between the speech intentions of all utterances of the former turn and the speech intentions of all utterances of the next turn, as the entire turns form a sequence. To some extent, these relations may still be captured as repeating indirect sequences. Nonetheless, these are less prioritized compared to direct sequences, leading to unequal treatment of utterances in the process mining algorithms. Additionally, the process models generated in this way would capture in reality mixed relations among speech intentions: within turns (e.g. $e_{C_1-t_1-u_1}, e_{C_1-t_1-u_2}$) and across turns (e.g. $e_{C_1-t_1-u_2}, e_{C_1-t_2-u_1}$), which appear as equivalent, being a potential bias for studying conversations.

One could think in this situation to generate instead a trace for each unique combination of utterance speech intentions (e.g. for C_1, there would be two process traces $\tau_{C_1-1} = \{e_{C_1-t_1-u_1}, e_{C_1-t_2-u_1}\}$ and $\tau_{C_1-2} = \{e_{C_1-t_1-u_2}, e_{C_1-t_2-u_1}\}$). However, the complexity becomes high: assuming m turns with an average of l utterances per turn then, just for one conversation, l^m process traces are generated. Of course, in reality, a conversation may not have more than 50 turns and a turn may not have more than 30 utterances[3], which means that the previous complexity has a higher bound. However, the value is still very high. Additionally, the sequence frequencies are also important for the process mining techniques as previously shown in the presentation of their metric frameworks. Hence, using this strategy for trace generation inflates the frequencies related to turns closer to the root in the conversation tree. These turns are common to multiple concurrent threads: for instance, each utterance of a turn being the root in the conversation tree would be repeated l^{m-1} times. Consequently, by taking in consideration all these arguments, mapping process traces on conversation threads does not seem to be an appropriate decision.

Therefore, the selected approach is to "cut" conversation trees into levels lying at equal distance from roots (the initiating posts) and consider a *process trace* to be a turn. As a consequence, the process mining technique first generates models of conversational processes within-turns. Second, to obtain an overall perspective on conversations, these process models are placed in a sequence, first being the model obtained from the initiating posts, second the model generated from the first-level comments and so on (see Figure 6.2 for an example). This approach is exten-

[3]Values randomly picked.

sive because it reveals both process models of turn building and knowledge about conversations overall. However, limitations are also identified: the relations among speech intentions across turns are only assumed to be potential dependencies and no information about the relation significance or about specific interrelated speech intention across turns (e.g. *direct* and *engage*) is available. Nonetheless, this limitation is accepted as a compromise in order to *correctly* discover relations among speech intentions, but future work is meant to complete the solution for addressing across-turn relations among speech intentions too.

So far, the method to obtain event logs is the following: speech intentions are process *activities*; each utterance, automatically annotated with a speech intention, represents an *event*; the sequence of utterances in a turn is transformed in a sequence of events and represents a *process trace*; all process traces generated for each turn in the corpus, belonging to the same level in the conversation tree, compose an *event log*. Finally, process mining techniques are applied on such event logs and visual *process models* of strategies to construct turns are obtained. However, an additional challenge exists because of the multi-label annotation, advocated in the current work. If a single label had been allowed per utterance, then this scenario would have been feasible.

To better expose the challenges raised by the multi-label annotation in the log generation, let's start with considering the following example. An utterance is annotated with *assert* and *suggest*. This could be interpreted as: a) the utterance is both *assert* and *suggest* (concurrency), or b) the utterance is either *assert* or *suggest* (decision), or c) the utterance contains first an *assert* then a *suggest* or vice-versa (sequence, possible in complex utterances). One approach to discriminate between these interpretations would have been to ask the annotators to explicitly state which of the aforementioned situations applied. Nevertheless, more effort would have been required in the manual annotation, while the goal was to design a procedure as simple as possible (see Chapter 4). Additionally, dedicated machine learning algorithms would have been required to automatically identify these relations. Further, as utterances correspond to multiple speech intentions, then an one-to-one mapping between utterances and events is questioned. In fact, it appears more correct to transform an utterance in multiple events, one for each associated speech intention. However, assumptions are made, first of all, about the nature of relations—that they are always direct sequences, and second, about the order of the activities in a relation—which speech intention of the three labels should be considered as the first, second or third event?

Let's further consider the following post with the associated speech intentions per utterance in parentheses: u_1: "Hello, I have some great news!" (*rejoice*, *greet*) $u2$: "You can now subscribe to my Youtube channel for new videos on quiting smoking" (*assert*, *request*). If the relations between the speech intentions of each utterance are manually inferred, then *greet* and *rejoice* are in direct sequence, while *assert* and *request* are concurrent. However, an event log could contain just sequences of events grouped in process traces; hence, more complex relations cannot be injected. At this point, there are two options: either to create a new process mining technique from scratch, capable of discovering process models directly from threaded conversations with

multi-labeled utterances—and also capable of revealing relations among speech intentions across turns. However, this idea is left for future work and currently a line of inquiry focusing on the generation of event logs is followed. The main reason for this course of action is that the existing process mining tools, which are rather mature in functionality, can be directly used.

Given the arguments so far, a heuristic solution is created in order to ensure an one-to-one mapping between utterances and events. The strategy is to choose randomly one of the utterance speech intentions, by giving lower priority to *assert*. This strategy aims to obtain the least biased mapping, while reducing the impact of the highly frequent *assert* class. Thus, it maintains a high precision of what could be observed if conversations were manually interpreted. Additionally, by following this strategy, the direct sequences in the event log are correct because any speech intention of an utterance is followed by any speech intention of the next utterance (e.g. *greet* and *rejoice* are followed by *assert* and *request* in the previous example). However, a limitation exists too: some known information is discarded, resulting in a sub-maximal representation. For the current work, exhaustiveness was not a criterion in the manual annotation process either—but accuracy, as the human annotators were not required to find all possible speech intentions realized in an utterance, but the most evident ones (see Chapter 4). The algorithm to generate event logs from annotated asynchronous conversations is presented in Algorithm 3.

For N conversations, with each conversation with a maximum of T turns and each turn with a maximum of U utterances, the complexity of the algorithm is $O(NTU)$. Additionally, it is assumed that the conversation tree data structure maintains lists of turns per level. Hence, getting the maximum depth of all trees is $O(N)$ and getting the turns at a certain level is $O(1)$.

6.2.2 Alternative Heuristics to Map Utterances on Events

Other strategies to create event logs are further proposed. The first strategy leads to a precise sequence representation, similar to the one presented in the previous section. The utterance is mapped on a single event, but the activity is the concatenation of its speech intentions in alphabetical order. The advantage of this approach is that all known information is exploited. Nonetheless, even if the tags are ordered alphabetically, the number of possible activities in the process is very high. For the proposed intention taxonomy with 18 classes, the number of possible activities reaches hundreds of possibilities making the process model unfeasible for visual exploration. If the goal of the analysis is to capture as much information as possible and the number of possible intentions is rather low (or trimmed by using a high node cutoff parameter in activity filtering), then this strategy can be used.

However, another aspect should be taken into account in the analysis of the models. Two activities such as *"assert greet"* and *"assert greet rejoice"* would be separate activities; pragmatically though, they overlap. For this case, applying clustering of activities based on their labels before model analysis or using a process mining technique which already supports node clustering (e.g. Fuzzy Miner implementation in ProM) may be considered.

Algorithm 3 Generate event logs from a list of conversation trees, one file per tree level.

Require:

convTrees—a list of conversation trees; each tree node stores a linked list of *utterances* and its *level* in the conversation; each utterance object has a property to get its *labels*.
outputFolder—the folder where the event logs are serialized to files.

procedure GENERATEEVENTLOGS(*convTrees*, *outFolder*)
    maxTreeDepth ← GETMAXDEPTH(*convTrees*) ▷ get the maximum depth of the trees.
    for each *level* = 1 to *maxTreeDepth* **do**
        eventLog ← GENERATEEVENTLOG(*convTrees*, *level*)
        SAVEEVENTLOG(*eventLog*, *outFolder*) ▷ serialize event log object to csv file.

function GENERATEEVENTLOG(*convTrees*, *level*)
    eventLog ← []
    for each *convTree* ∈ *convTrees* **do**
        turns ← GETTURNS(*convTree*, *level*) ▷ return the turns found at the specified level in the tree.
        for each *turn* ∈ *turns* **do**
            pTrace ← GENERATEPROCESSTRACE(*turn*)
            add *pTrace* to *eventLog*
    return *eventLog*

function GENERATEPROCESSTRACE(*turn*)
    pTrace ← []
    traceId = GUIDGENERATOR( ) ▷ generate a global unique identifier.
    for each *utterance* ∈ *turn.utterances* **do**
        event ← GENERATEEVENT(*utterance.labels*, *traceId*)
        add *event* to *pTrace*
    return *pTrace*

function GENERATEEVENT(*labels*, *traceId*)
    activity ← GETACTIVITY(*labels*)
    eventId = GUIDGENERATOR( ) ▷ generate a global unique identifier.
    return (*traceId*, *activity*, *eventId*)

function GETACTIVITY(*labels*)
    if *labels.length* = 1 **then**
        activity ← labels[1]
    else
        if "assert" ∈ *labels* **then**
            remove "assert" from *labels*
        activity ← RANDOM(*labels*) ▷ select a random element from the list.
    return *activity*

Next, strategies that map an utterance on multiple events in the log are described. As all introduce some assumptions regarding the relations among the events per utterance, care should be taken to interpret the process models. One is to simply transform utterance labels in consecutive events considering the order of the labels as in the annotation or alphabetically. Thus, relations are captured between speech intentions, but the direction of the relation is not necessarily correct. A second strategy is derived from the previous, but with one change. A possible "correct" ordering of events could be created by using context and pragmatic knowledge. For instance, an utterance tagged with *(assert, greet)* in the beginning of a post could be transformed in $greet >_L assert$. However, creating these rules is a bias and formulating hypotheses about speech intention relations from process models derived with these strategies are less advised.

6.3 Evaluation Methods

The solution presented in Figure 6.1 meets now the goals established in Chapter 1: it extracts behavioral knowledge as processes over speech intentions from asynchronous conversations. In order to achieve these goals, multiple artifacts were designed: a speech intention taxonomy, the automatic annotation of asynchronous conversations with speech intentions by using supervised machine learning and the generation of event logs from annotated asynchronous conversations. The two first artifacts were created in two iterations.

Several desired properties of the solution were defined in Chapter 1 too: the solution should be automatic, corpus-independent and effective—as in, comprehensive (it covers different perspectives of the studied phenomenon in more detail than the existing works), relevant (it is useful for at least one application domain), correct and complete (the results obtained by applying the solution are correct and complete).

The properties that apply for each artifact of the solution were evaluated through various methods. Table 6.1 presents an overview of the evaluation methods used throughout the current work, methods selected from those described by Hevner et al. [93] to evaluate artifacts in design science.

Table 6.1: Evaluation methods used in the current work, selected from [93].

Artifact	**Chapter**	**Properties Evaluated**	**Evaluation Method**
Intention taxonomy	Chapter 3 Chapter 4	correct, complete comprehensive, relevant corpus-independent	Analytical: static analysis, Descriptive: informed argument, Experimental
Automatic annotation	Chapter 3 Chapter 5	automatic, correct complete, comprehensive, relevant, corpus-independent	Analytical: static & dynamic analyses, Descriptive: informed argument, Experimental, Testing
Event log generation	Chapter 6	automatic, correct complete, comprehensive, relevant, corpus-independent	Analytical: static & dynamic analyses, Descriptive: scenario, informed argument, Observational: field study, Testing

By analyzing its static properties, by referring to the knowledge base—the related works in computer science and the theoretical linguistic works (informed argument), and experimentally, it was shown that the speech intention taxonomy was corpus-independent, correct, complete—to some extent and dependent on the context (e.g. only for public tweets), more comprehensive than the existing works, while also revealing the feasibility of its application by non-experts.

Then, the automatic approach to annotate asynchronous conversations with speech intentions was statically analyzed by focusing on various parts: the features, the machine learning algorithms and the experimental setups. Thus, it was shown that the solution was automatic and corpus-independent. Through experiments, testing and dynamic analysis, the correctness, completeness and generalization of the approach were assessed. The arguments for its relevance and comprehensiveness were built with reference to the literature in the knowledge base.

As for the evaluation of the currently created artifact—the algorithm for event log generation, and of the overall approach to obtain processes of interrelated speech intentions, multiple evaluation methods are chosen. Testing and static and dynamic analyses are used to assess if the algorithm to generate event logs from annotated asynchronous communication is automatic, correct and complete. Then, informed argument is also added to the previous methods, to evaluate if the overall approach is automatic, corpus-independent, correct, complete and comprehensive. Finally, the potential relevance of the solution to medicine is shown through the presentation of a scenario. For that, process models were generated from the Reddit corpus, introduced in Chapters 4 and 5, and discussed. In addition to this scenario-based evaluation, a field study was conducted with a Linguistics researcher interested in conversation analysis. The details regarding the observational research are further provided.

The field study consisted in both direct and indirect observations. Initially, a day meeting was scheduled with the researcher. The day started with an introduction of the annotated input data, of how to use the process mining tool Disco and how to interpret process models in general. The event logs for the first three levels in conversations were generated in advance from the same Reddit corpus and loaded into Disco. Additionally, events logs were also generated by considering as activities the Searle's speech act types.

Then, the researcher was asked to analyze conversations, to extract interesting patterns from the process models and to compare the models when using speech intentions or speech act types as activities. At the end of this day, the researcher expressed his desire to have access to two additional event logs: one corresponding to all turns existing in the corpus and another one corresponding to all comments of the corpus. These were provided and the researcher continued the analysis remotely throughout the week. Any additional questions regarding the tool features and the interpretation of the models were answered by the author through emails and brief online video calls. As the second part of the field study consisted in indirect observations, the conversation analysis results were mainly analyzed to assess relevance. Moreover, an open discussion was organized with the researcher at the end to find out more about his perceived

relevance and utility of the approach for conversation analysis.

6.4 Results

The algorithm to generate event logs from annotated conversations was implemented as a Python script. The required input is a comma-separated values file (csv) with each line containing an utterance, the utterance label set (up to three) and the unique identifier of the turn to which the utterance belongs. Thus, the complete approach is automatic.

As previously discussed in Section 6.1, this representation of processes is general and more comprehensive than transition diagrams—the most common alternative representation used in the related works [28, 178, 189]. The correctness and completeness of the event log generation algorithm was thoroughly argued in Section 6.2. However, the correctness is also influenced by the quality of the automatic annotation.

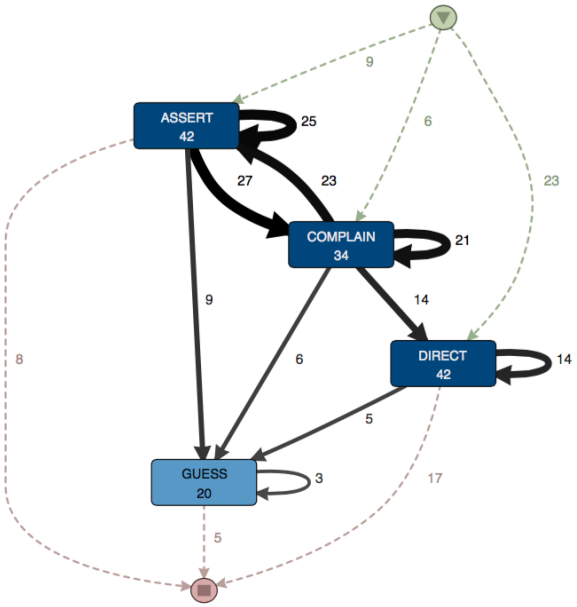
Finally, Fuzzy Miner, the process mining technique used to create process models, correctly identifies common behavior from event logs [85], assuming that the event logs contain much more representative observations than noise. The completeness of the model is obtained by relaxing the parameters to capture all observed behavior. Nonetheless, this approach suffer from poorer visualization, resulting in "spaghetti models". Ultimately, it is the decision of the process analyst to leverage the completeness and usability.

The relevance of the global solution is discussed in the remaining subsections.

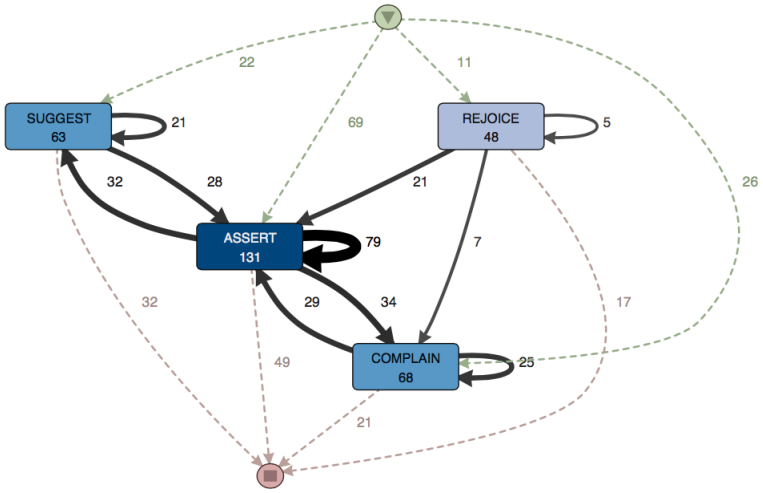
6.4.1 Scenario-Based Evaluation Results in Medicine

Figure 6.2 presents the process models discovered automatically from the annotated Reddit corpus: the top model is generated from the initial posts only, the middle model from the immediate comments to the initial posts and the bottom model from the comments replying to the first-level comments. The process models are simplified to expose representative behavior through the process mining interface. Specifically, lower values are associated with the activity and path cutoff parameters. The process mining tool also allows to visualize and analyze accurate representations of the models or, for instance, edge cases, if desired.

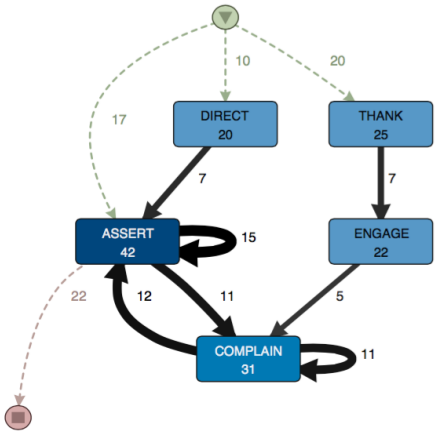
Further, it is illustrated how to interpret such process models. The activities, in this case, the speech intentions are represented through the blue rectangles and their relations through the arrows. The darker blue a rectangle is, the more frequent that speech intention is observed in the logged events. Similarly, the size of the links is correlated with the frequency. Then, there are two types of links: with continuous and with interrupted lines. An interrupted line relates a speech intention to the *Start* and *Stop* default process states — represented with circles. The numbers annotating these lines show how many times an intention is first or last in a turn. By contrary, an arrow represented with a continuous line should be interpreted as: the source intention is *eventually followed* by the target intention; hence it does not necessarily imply direct sequence,



(a) Model–first posts only.



(b) Model–1st-level comments.



(c) Model–2nd-level comments.

Figure 6.2: Within-turn processes over speech intentions, mined from a *Reddit* corpus with Disco.

as also explained in Section 6.1. For instance in the top process model of Figure 6.2, a *complain* utterance is eventually followed by a *direct* utterance in 14 turns.

In the process model from posts only (the top part of Figure 6.2), it is observed that *assert*, *complain* and *direct* dominate and *assert* and *complain* very often co-occur. It appears that stories are shared and questions or requests are communicated. What is interesting is that some posts do not contain any question at all, considering the presence of arrows from *assert* and *guess* directly to the *Stop* state. Hence, it appears that these kinds of posts are aimed to share a story—rather with negative feelings, given the representative *complain*—and, possibly and implicitly, asking for witness and support. It can also be observed that *direct* is sometimes the unique speech intention between *Start* and *Stop* or it could be reached through an initial story—the loops involving *assert* and *complain*. Consequently, various strategies to ask a question or make a request emerge from these conversations.

Then, in the process model from the first-level comments (the middle part of Figure 6.2), *suggest* utterances formulated for giving advice and encouragements appear highly popular. This is expected because, normally, these comments are aimed to reply to the previous questions and requests. It can also be noticed that the expressive utterances diversify as now *rejoice* appears too. In fact, it looks like some replies stick to a positive note from the beginning until the end, while others either use mixed opinionated statements or only negative ones. However, *complain* stands also for grieving, apart from being associated with negative feelings. Hence, these utterances could represent signs of empathy in comments.

Finally, in the process model from the second-level comments (the bottom part of Figure 6.2), three tendencies can be observed: either appreciations are shown followed by a commitment probably towards the received advice (*thank* and *engage*); or more questions are asked followed by extra narration (*direct* followed by the length-2 loop between *assert* and *complain*); or only an extension of the initial story is shared (the *Start* state directly preceding the length-2 loop between *assert* and *complain*).

Already expected knowledge about the use of health-related online forums was discovered, which could be considered as a validation evidence for the correctness of the approach to generate events logs and to discover process models of interrelated speech intentions from these logs. The results are also very encouraging with regard to the knowledge they represent and how this knowledge could be further used.

The processes over speech intentions extracted from turns could be used to analyze narratives and conversations including persuasion and to formulate new hypotheses relevant to narrative medicine studies and persuasion and compliance gaining technologies. For instance, questions such as "Are there specific strategies to offer an advice, which result more often in a commitment?", "Is the communication different in various health-related communities—for instance, regarding different diseases?", "What are the most frequent communication strategies potentially leading to new beliefs?" or "Does the communication in a community evolve in time?" can now be answered.

Another interesting path of inquiry would be to compare the processes of conversations from different communities (e.g. for different diseases, chronic vs. non-chronic conditions, Reddit vs. non-Reddit data, conversations from the medical domain vs. from other domains) in order to analyze if they differ and what would be the implications of such results for practical solutions. These type of insights are potentially very valuable for getting more knowledge about the online health information seeking and dissemination and for supporting the creation of innovative AI artifacts such as virtual assistants.

6.4.2 Observational Evaluation Results in Conversation Analysis

The linguistic researcher appreciated the possibility to visualize common patterns of speech intentions or of speech act types as process models, the user interface and the interactivity of the process mining tool. Although the complexity of the process mining technique was hidden, one challenge the researcher faced in the beginning was to get acquainted with the concepts (e.g. what is an event, a case, a variant) and to understand the relation between event logs and process models. Additionally, the interpretation of the process models was initially challenging too because of the learning curve regarding the graphical representation of the dependency graphs: understanding what the boxes, edges, numbers attached to the edges represent and how different views or filters impact the process model. However, after the support received during the first day, the researcher used the tool independently.

Further, the discovered knowledge is presented. The linguistic researcher outlined the first insights with regard to the Searle's speech acts types, namely about their relations as extracted from the full corpus. Figure 6.3 shows the process model containing all speech act types, except for *other*, which was filtered out. The model was simplified by setting the edge cutoff parameter to 75%. Several observations were made:

- Assertive speech acts appear to connect all other types of speech acts; in contrast, directive, expressive and commissive speech acts appear usually to not succeed one another directly.
- One in two expressive utterances is followed by another expressive utterance; a similar observation can be made for one in three directive utterances; however, commissive utterances do not appear in loops.
- The most common speech act type in the beginning of a turn is assertive, while the most common class at the end of turns is expressive; normally, commissive speech acts maintain a middle position.

Considering these observations, the linguistic researcher formulated two hypotheses:

1. The position of commissive speech acts in turns, surrounded by assertive utterances which are factual, may suggest that a commissive utterance formulated outside this pattern may be perceived as "improper".

2. The loops regarding the assertive and expressive utterances appear to show that, compared to commissive, these speech act types may need more utterances conveying them in order to produce the illocutionary effect. In this situation, for instance, expressive utterances may be seen as inappropriate if found isolated in conversations.

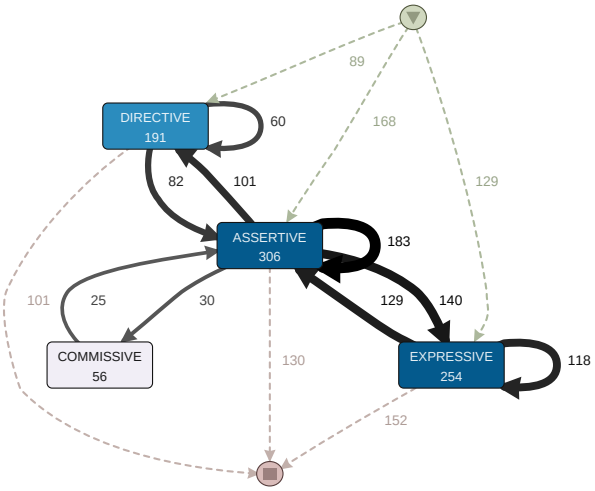


Figure 6.3: Process model mined from all corpus turns with Searle's speech act types as activities.

The linguistic researcher concluded that this type of empirical analysis could lead to the extension of the speech act theory to expose typical and proper sequences for each speech act type.

Then, the linguistic researcher chose to analyze a process model featuring the extended speech intentions as activities, mined again from the full corpus (presented in Figure 6.4). The activity cutoff parameter was set to 70% and the edge cutoff parameter to 30%. In this process model, it was observed that:

- *Complain* and *rejoice* appear in very different positions in turns. *Complain* utterances are often at the beginning or end of a turn, while *rejoice* utterances are less common and have a more embedded position. However, the researcher also emphasized that these patterns of expressive use may be due to the corpora that contains only health-related conversations.
- *Assert* and *complain* utterances form a loop.
- With regard to the peripheral position of directive speech acts in turns, *direct* utterances tend to be more often at the beginning than at the end and the opposite is true for *suggest*.

The researcher found that the more detailed speech intentions gave a different view on conversation turns. Additionally, aligned with the previous insights, the fact that positive and

negative expressive appeared in very different sequences suggested that the speech act definitions may be extended with proper sequences and these sequences should rather be formulated at a more detailed level. The latter aspect is especially relevant as, although *complain* and *rejoice* belong to expressive, the same speech act type, they are not similarly situated in conversations.

Further, the researcher investigated the per-level processes (Figure 6.2). In this case too, he found more convenient to use speech intentions as activities instead of speech act types. The reason for this was that the process models with speech intentions revealed expected knowledge about the use of online forums (ground-truth), thus helping the researcher to gain trust in the approach. Specifically, he looked into what were the starting activities for each process model. Utterances containing *direct* are mainly first in the process model mined from initial posts, as these posts often start with questions, standing for post title. Then, *assert*, *complain* and *suggest* utterances follow in the first-level comments, which the researcher interpreted as answers to the previous questions or requests in the form of plain statements, recommendations or expressions of empathy. Finally, the second-level comments start with *thank*, *assert* or *agree*. These comments are frequently written by the users initiating the threads; thus, they may express gratitude regarding the received recommendations or agreement with the previous posts.

Finally, the researcher returned his focus on exploring the event log generated from the full corpus, with speech intentions as activities. This time he looked for significant, unexpected regularities. He noticed that a preference to use *complain* utterances eventually followed by *direct* existed in conversations, while the opposite was less frequent. Additionally, when *direct* was eventually followed by *complain*, intermediary *assert* utterances appeared between the two. For the researcher, this was a possible indication that *direct* follows *complain* is the norm and going against the norm, with *complain* follows *direct*, may require a mediation through *assert*. Consequently, a complaint could open the possibility for questions, while a question frequently requires other types of discourse continuation. Finally, the researcher emphasized that finding such preferences between speech act pairs could reveal new scenarios for linking speech acts than what was already known and thus new investigation paths in conversation analysis.

At the end of the analysis, the researcher expressed his wish to also find processes across turns, as this knowledge could be extremely valuable. In this way, knowledge about speech act sequences such as if they are followed by other speech acts in the same turn or followed instead by new turns could be outlined. The researcher referred to this as a differentiation between "turn-ending sequences" and "turn-keeping sequences". The sequences ending a turn could be identified then as some sort of turn-taking punctuation, which could lead to the discovery of new norms in the conversational grammar.

To sum up with, the observations showed that the approach was relevant and could be used independently. At the end of the trial, the linguistic researcher acknowledged that the insights from the discovered process models were very interesting and even unexpected sometimes. He expressed his strong support to continue using process mining techniques for investigations in

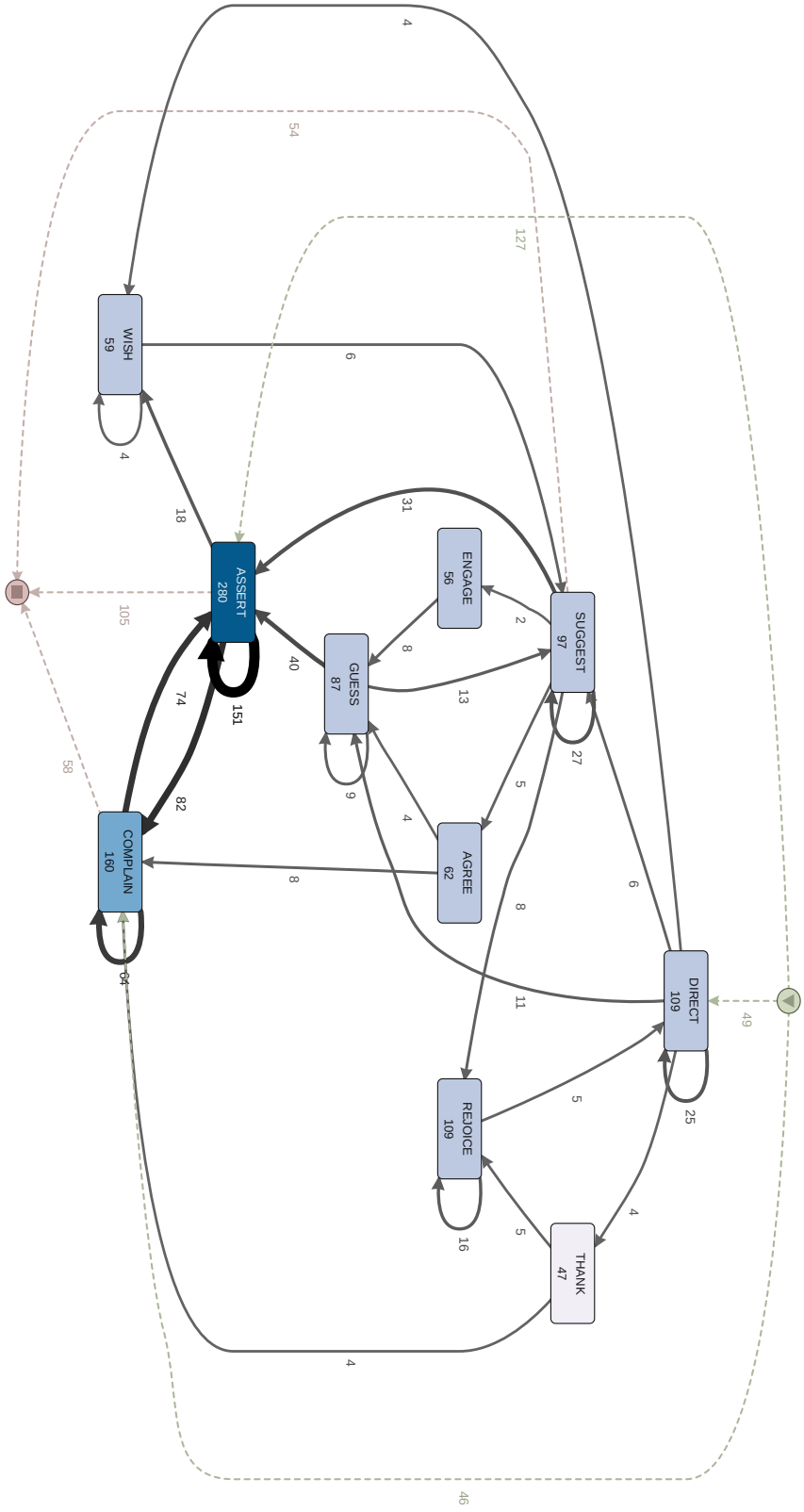


Figure 6.4: Process model mined from all corpus turns with speech intentions as activities.

conversation analysis. The type of knowledge the researcher outlined was diverse.

First, he noticed that preferences may exist in the ordering of speech intentions ("typicality effects") and he put forward the hypothesis that, if these effects were confirmed, corpus pragmatics combined with computational approaches, like process mining, could expand the theoretical understanding of speech acts. Specifically, the speech act definitions could also include rules of proper use in conversations derived from aggregated verbal behavior analysis. Second, he formulated another hypothesis by analyzing the process models on the existence of sequences of utterances associated with complex speech acts and these sequences acting as "a sort of punctuation for turn-taking" [41]. In general, he found that exploring real-world corpora in this manner could formulate new lines of inquiry and create evidence for phenomena already investigated in the conversation analysis and speech act theory.

Limitations were identified too. The separation of event logs per conversation levels appeared of little interest for the researcher, apart from checking ground-truth knowledge and enabling him to trust the models. Additionally, he also mentioned the preference to rather work on synchronous conversations or non-threaded asynchronous conversation corpora, arguing that sequences of speech acts across-turns could lead to even more interesting results for conversation analysis.

6.5 Conclusions

A solution centered on process mining to derive process models of interrelated speech intentions from asynchronous conversations was proposed (**RQ3**). In particular, existing process mining techniques and tools were put to test and the relevance of the obtained results was assessed. The discovery of the relations among speech intentions is realized in two steps. First, asynchronous conversations are annotated with speech intentions. Second, event logs capturing relevant aspects of conversational behavior are generated from the annotated conversations. An extensive argumentation was provided for the choice to map speech intentions on activities, utterances on events and turns on process traces. Also, different heuristics to map multi-labeled utterances on single or multiple events were discussed.

Compared to the related works, where the relations among speech intentions are revealed as transition diagrams capturing just direct sequence, the process models obtained with process mining can expose other types of relations too: length-2 loop, significant indirect sequence and concurrency. Additionally, the analysis is interactive, allowing different views of the same behavior, from a very representative one to a very detailed one.

Although the evaluation was mainly qualitative, the results were encouraging. Already expected knowledge about the use of health forums was discovered, which validated the approach to generate events logs and to discover process models of speech intentions with process mining. Moreover, interesting insights, potentially valuable for further investigations in medicine or integration in innovative artificial intelligence solutions, were identified.

The observational study with a linguistic researcher validated the relevance of the results for conversation analysis, but also exposed limitations of the approach that should be considered in the future work. Contrary to the initial expectations, the generation of process model per conversational levels was less interesting for the linguistic researcher—instead, the full corpus was preferred. As expected, the detailed speech intentions led to richer and more accurate observations to be made, which could have been otherwise incomplete or misleading by using only the speech act types as activities. Finally, the linguistic researcher expressed his wish to analyze across-turns processes too, in order to have even more relevant information for conversation analysis.

There is only one other known work that uses process mining for analyzing conversations [216]. The first difference from the current solution is that the classes used for annotating their corpus are adapted from [118], being inspired from dialogue acts and customized for question and answer forums. Moreover, the annotation unit is the turn and the annotation strategy is single-label. In this situation, the mapping of process traces on threads is the natural choice [214]. The discussion provided in this chapter on how to use process mining for conversation analysis is very comprehensive, including this scenario but also others. Consequently, the contribution of the current work is a new way of modeling conversations with process models—as sequences of within-turns processes over speech intentions, but also an extensive argumentation of the chosen design and other design alternatives and their rationales.

6.5.1 Study Limitations

As previously proven in Section 6.2, by randomly selecting one speech intention per utterance, the event log generation is correct. Moreover, the process models are obtained with existing process mining algorithms and tools, which have been already extensively validated in the past [1]. Thus, this aspect of the *construct validity* is satisfied. Nonetheless, with regard to the *internal validity*—to prove the relevance of the generated models from tagged conversations, and the measurement aspect of *construct validity*—to correctly measure the relevance, no quantitative evaluation has been designed. Qualitative methods, a scenario-based evaluation and an observational study, were chosen for now considering that the automatic discovery of processes from conversations with process mining is an emerging topic and the nature of research is exploratory. Consequently, multiple threats may stem from this. However, a thorough motivation on why such process models could be relevant for medicine have been provided in Chapter 5, supported also by the theoretical knowledge on behavior. Additionally, a standard protocol for observational study was followed, although the involvement of just one subject is a limitation. Also, the subject was already accustomed with the speech intention taxonomy as he was involved in its creation. Nonetheless, the subject did not have any knowledge of process mining before the study. Regarding the *external validity*, it can be claimed that there are no threats regarding the generation of process models from other annotated conversation turns, while the degree of relevance depends on other subjects,

application domains and specific applications, and still requires future validation.

6.5.2 Directions for Further Research

First, the solution should be extended to capture relations among speech intentions across turns too. With the current strategy of annotating conversations—per utterance and multiple labels, across-turns processes are challenging to be discovered with existing process mining techniques without increasing the complexity exponentially or altering the results, by omitting the integration of all direct sequences between two consecutive posts in the event log. If very large corpora of annotated conversations are available, randomly selecting one speech intention for each turn and mapping threads on process traces is feasible to potentially reveal significant relations with the already existing process mining techniques and tools. Otherwise, the development of new techniques, which address the particularities of conversations as behavioral data and still provide the same degree of interactivity and visualization as the existing process mining tools, should be researched.

Second, further research should be geared towards a more extensive validation of the proposed approach, particularly as regards its relevance. This is also a way to gather new requirements and to identify limitations, as shown in the preliminary observational study. Additionally, other corpora including from other domains should be considered in the evaluation. If the number of subjects in the future observational studies is high enough to derive statistically significant conclusions, the relevance could be also assessed through structured surveys and statistics.

Third, the relevance of the proposed approach is ultimately proven if the knowledge extracted from the mined process models can be useful for future research and applications. If the hypotheses formulated from these models are proven true or if the knowledge from the process models leads to designing useful artifacts, then the relevance is fully supported, as new theories or applications are created. For instance, past work conducted by the author, showed that devising news recommendation strategies from processes over news reading intentions resulted into more effective and diverse recommendations and process mining could be used directly by news organization stakeholders with limited data science background [52, 56]. Similar future research endeavors should be disseminated and their impact assessed.

CHAPTER 7

CONCLUSION AND PERSPECTIVES

The objective of the current work was to propose a corpus-independent method to automatically reveal comprehensive behavioral models from text. The automatic extraction of behavioral knowledge from digital traces is central to many domains, either to study phenomena revolving around behavior or to enable the creation of various technologies (e.g. for fraud detection, for ensuring cyber-security, for recommender systems, for evolving organizational information systems and so on). Most of the existing approaches rely on structured traces, such as event or web logs, which are passively generated during the interaction of people with computer applications.

However, the largest amount of digital traces are actually unstructured and represent content pro-actively generated by people mostly as text. Text, as a product of communication, is a very rich behavioral data as shown by many domains in humanities and social sciences. However, it is still underexploited in computer science—the most popular approach to analyze text in order to extract behavioral knowledge being sentiment analysis.

Multiple challenges are identified regarding the automatic analysis of text to extract behavioral knowledge. Text as input, apart from being unstructured, appears also in very diverse forms and the basic technologies to process text are still imperfect. Further, when representations of behavioral knowledge from text are adopted by technology originators from theoretical domains, these representations are adapted to particular domains or applications; when domain-independent, these representations often do no provide a sufficient level of detail; and when detailed representations are adopted or designed, their interpretation and use in practice by non-experts in discourse analysis are challenging.

Therefore, to achieve the objective and tackle these limitations, first, a better understanding of the relationship between text and behavior, from a theoretical standpoint, was provided. Communication, either through writing or speaking, is behaving through words. Linked to verbal behavior, speech acts or speech intentions reveal the linguistic intentions brought about

by the wording. One of the most famous frameworks to conceptualize speech intentions is Searle's taxonomy, consisting of five types of speech acts [184]: assertive, expressive, commissive, directive and declarative. Additionally, apart from analyzing individual utterances or messages for their meaning, conversations as a whole could be modeled as processes, with each exchange in conversation, conceptualized through speech intentions, opening the possibility to other speech intentions to be further conveyed. Thus, in order to model behavior from text, one can look into speech intentions associated with individual text units and into processes of interrelated speech intentions which characterize conversations overall.

The related literature showed that, when this framework based on speech intentions was adopted to analyze text, synchronous conversations used to be the most frequent input. Research has shifted much later to asynchronous conversations. Their exploitation is arguably more difficult and many of their characteristics generalize the characteristics of the text-based synchronous conversations. Consequently, the current objective was scoped to address only asynchronous conversations and was achieved through multiple contributions. These are further summarized, grouped by the area to which they belong: *tangible solutions*, *innovative methods* and *theory development*.

The core of this PhD was to propose a *tangible solution* to a practical problem, which was shown to be in fact an opportunity for many domains. An end-to-end, corpus-independent approach to automatically transform conversations in comprehensive visual models of interrelated speech intentions was created. Multiple individual artifacts resulted from this:

- A framework to analyze and compare the literature on modeling asynchronous communication with speech acts was developed. This framework emerged throughout a systematic literature study, including all the existing publications on the topic since 2000. The framework has multiple facets regarding the varied application domains, the knowledge representation, the corpus creation and validation, the automatic approaches to model conversations, the diverse groups of features relevant to conversation modeling and the presence of relations among speech intentions in the modeling (Chapter 2).

- An automatic and corpus-independent method to analyze speech intentions in the Twitter public communication was proposed. Specifically, a speech intention taxonomy for public tweets was created through theory study and empirical corpus analysis and further validated. Then, discourse features incorporating linguistics means interpreted as potential cues to express speech intentions were defined. These features were further used with supervised machine learning algorithms to automatically annotate tweets with speech intentions. Compared to the existing works that annotate tweets with speech acts, this solution is generic, designed to discover multiple speech intentions per tweet, capable of identifying finer-grained speech intentions characterizing the public communication on Twitter and achieved satisfactory results by relying only on discourse features, making it applicable to any context or domain (Chapter 3).

- A fine-grained corpus-independent taxonomy of speech intentions, suitable for any type of communication was defined, by revising the theoretical body of knowledge and by empirically reconsidering a new type of corpus: forum conversations. A further layer of organization was added to the proposed 18 classes of speech intentions through conceptual analysis. Specifically, oppositional traits among the speech intentions were identified, in order to create a more structured and applicable classification. The final taxonomy consisting in speech intentions and oppositional traits was validated for reproducibility and interpretation consistency with non-experts too. The taxonomy was proven comprehensive, while allowing facile re-use and interpretation (Chapter 4).

- A machine learning-based method to annotate forum conversations with speech intentions was proposed and validated. Although it relies on standard classifiers, the originality emerges from multiple aspects of the experimental setup. Similar to the previous iteration and contrary to all existing works on forum and email conversations, sentences are associated with multiple speech intentions, in order to account for indirect speech acts and to enable fuzzy interpretations, being closer to how human perception occurs in reality. Moreover, extensive experimentation was conducted to select the best setup for multi-label classification, to explore strategies to improve the results and to assess the external validity of the solution (Chapter 5).

- A method to transform annotated conversations in structured, well-defined logs required by process mining techniques and tools was developed. The design of this method took into account the relevance of the discovered knowledge, which was also validated in practice in two domains: linguistics and medicine. Moreover, other strategies to generate these logs were discussed in terms of design decisions and their impact on the output knowledge. Regarding the originality, this is the first work to integrate process mining for the general modeling of conversation turns as processes of interrelated speech intentions. The majority of the related works resort to probabilistic graphical models. Only another work applied process mining to threads from a question and answer forum [214]. However, their adopted speech intentions are not suited for general linguistic research on the structure of conversations, being domain-specific. Moreover, the strategy to obtain the event logs relied on simpler input representation—each turn had associated one label [214] (Chapter 6).

- Two corpora were tagged with speech intentions, validated and released to be used by the community for further research and experiments. The first corpus[1] contains 1100 health-related public tweets, 600 being collected by the keyword "autoimmune" and 500 by commonly used medical terms and jargon. The second corpus[2] consists of 51 conversations (2280 utterances) on an autoimmune disease from Reddit, a popular online forum.

[1] http://tinyurl.com/hk9t83y
[2] http://tinyurl.com/yc8mjtb2

Further, the second area of contribution was the novel use of an existing method—process mining, for conversation analysis (*innovative methods*):

- Several reasons for putting process mining to the test in this thesis existed. Given representative logs of behavior, process mining techniques have been proven to correctly discover behavioral models [1]. Moreover, the obtained models are visual and interactive, enabling easier exploration of conversations, in particular, in multi-disciplinary research settings, as in-depth technical knowledge is not required. Process mining applied directly to text is very rare and most of these existing techniques specify clear restrictions on the expected input (e.g. the text must contain reported behavior [57, 58]). Moreover, for automatic conversation analysis, only one other work used process mining [214], but on a simplified scenario compared to the current one which is based on threaded conversations annotated with multiple speech intentions per utterance.

Last but not least, this thesis brings contributions in the area of *theory development* by revising and creating new knowledge in several fields:

- In the field of automatic conversation modeling, new knowledge was created by systematically analyzing the existing works. To date, this is the first systematic literature study on modeling asynchronous conversations with speech acts.

- In the machine learning and text mining domains, knowledge about the performance of various supervised machine learning algorithms on a text classification task was obtained through experiments. The best results obtained by Logistic Regression and Linear SVM in similar setups has been already shown by other works; thus, the current experimental results support these conclusions too.

- In the field of medicine, by analyzing specific sub-domains (narrative medicine, health information seeking and dissemination, persuasion and compliance-gaining technologies) and the theoretical literature on behavior, future directions to support the needs of these sub-domains via the automatically identified behavioral knowledge as processes of speech intentions were formulated. Moreover, empirical knowledge was gained about how conversations developed and turns were constructed in a health-related forum and about the perceived speech intentions of the public health-related communication on Twitter.

- In the field of conversation analysis, several hypotheses on the typical development of conversations as sequences of speech intentions were formulated, by empirically analyzing corpora with the proposed solution. The results obtained in the validation of the proposed taxonomy raised multiple questions on whether there is a continuity between assertive and expressive classes or instead they are ontologically divided as Searle claims [184], and on whether the definitions of speech acts should include rules of proper use in conversations.

Although the objective of the current work was achieved, multiple limitations were also identified and should be addressed in the future work. The validation of the speech intention taxonomy in the first design iteration showed that *advise* could not be consistently identified by the two human annotators. In the second set of experiments with multiple pairs of human annotators, *advise*, *warn* and *propose* were grouped under *suggest* and appeared to reach satisfactory inter-rater agreement. However, is this result influenced by the dataset, meaning that if the second version of the taxonomy had been applied on a Twitter corpora, would it have led to similar results as when applied on Reddit? Would it be more useful to try to separate *suggest* back in *advise*, *warn* and *propose*? Additionally, the experiments at the end of the second design iteration also revealed problems with some other speech intentions. The nature of the problems was different. For instance, *sustain*, which was very frequently used in annotations, appeared to have a problematic definition and labeling instructions. However, the high disagreement for *accept*, *refuse* and *disagree* seemed to be due to their low frequency. Also, *require* and *declare* were excluded because of their absence in the selected Reddit corpus. All these aspects show that there is a need to further improve some taxonomy classes and that more experiments to manually annotate corpora are required, including more diverse corpora of asynchronous communication which cover all speech intentions.

A consequence of the previous limitations is that not all the proposed speech intentions were targeted by the machine learning experiments. Thus, the feasibility to discover all of them automatically with the proposed approach was not evaluated. Also, for an extensive validation of the results obtained by the classifiers, more annotated conversations are needed. This could also overcome the challenges posed by the class-imbalance. Alternatively, semi-supervised learning appears promising in the literature to learn models from fewer instances and should be explored. Also, many related works seem to obtain better results with structural learners and the state of the art in text mining is to use deep neural networks, which should be other future directions of research. Finally, a much larger ground-truth corpus, could also help with understanding the effectiveness of the current features. Specifically, are the current features sufficient or is more feature engineering required? Would automatic feature learning be a better alternative to manual feature engineering? Also, the most predictive content features appeared generic, but only by extracting and using them in classifying other corpora, their effectiveness and generalization would be indeed assessed.

Next in the pipeline came the generation of event logs from annotated conversations. The transformation algorithm is correct but the correctness of the information in the event log ultimately depends on the output of the classifiers. Consequently, a question that should be explored is: what is the impact of the classification on the relevance of the obtained process models for practical applications? One could imagine that for some practical solutions, a lower recall may be acceptable, while a lower precision may be detrimental. Then, a very important aspect in the agenda is how to reveal processes of interrelated speech intentions across turns?

Further, the relevance was assessed qualitatively and rather preliminarily. Therefore, a more extensive evaluation is required too. Also, more research needs to be done to analyze the bias in question because of language and choice of data. Finally, the most effective way to prove that this type of behavioral knowledge is useful and relevant to the domains that study and create solutions related to behavior is to disseminate the current research and attend its adoption.

While further research is still needed to improve and validate this work, the results are promising and open up new perspectives for automated behavioral analyses and for automatic interpretation of human language. Building behavioral maps, correlating speech intentions and processes of interrelated speech intentions to human traits such as personality or to human mental states such as boredom, happiness, anger, or identifying public narratives that could adversely affect individuals such as deceptive and hate speeches are other expected applications of the current work.

APPENDIX A

INTRODUCTION TO DESIGN SCIENCE

Design Science is a pragmatic research paradigm emerging from engineering, frequently employed in the Information Systems (IS) research [93, 156, 224]. Its main goal is to create innovative artifacts as solutions to real-world problems or opportunities. Consequently, the creation of an Information Technology (IT) artifact is central. Nonetheless, demonstrating the artifact relevance to an application domain is as important as the solution itself. In practice, design science research consists of three research cycles: the *relevance cycle*, the *design cycle* and the *rigor cycle* [93].

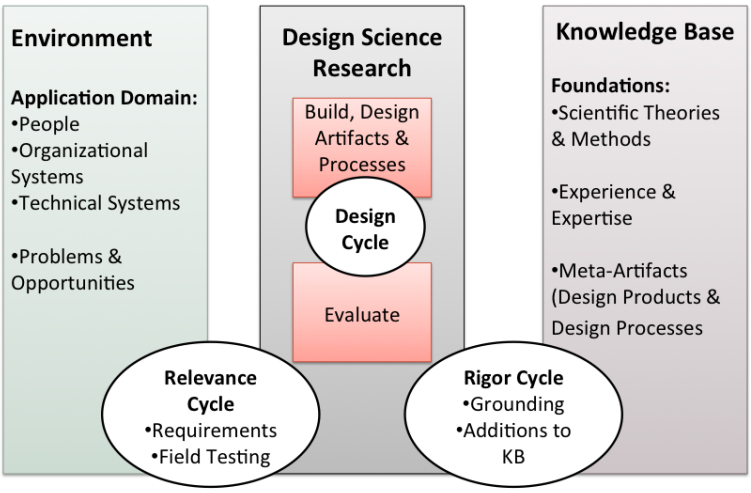


FIGURE A.1. Design science research cycles [93].

The relevance cycle (on the left side of Figure A.1) motivates the design activities in relation

to the environment of the research project. The main objective of a design science research project is to improve the environment. The improvement consists either in solving problems or exploiting opportunities. People and systems—also referred to as the application domain– are the specific concepts of the environment impacted by the research project and that could benefit from the potential improvement. Therefore, the first step in a design science project is to analyze the application domain in order to formulate the requirements of the new artifact. These requirements will also play the role of acceptance criteria in the validation of the artifacts [93, 156, 224].

The rigor cycle (on the right side of Figure A.1) relates the design activities to the knowledge base. The knowledge base consists of scientific knowledge–theories and engineering methods, and of the application domain knowledge–expertise and experiences and existing artifacts and processes. The existing knowledge drives the innovation by revealing what is already known and what could be used to create the new artifact. Furthermore, the knowledge base does not only seed the design activities with knowledge. The reverse relation is also present: in the process of creating and evaluating innovative artifacts, new knowledge is learned and could be added to the existing base [93, 156, 224]. In fact, the created knowledge can be considered as a key contribution to the academic community, as much as the created artifacts are contributing to the improvement of the environment.

The design cycle is primary to the design science paradigm (in the center of Figure A.1). It implies an iterative creation and evaluation of solutions, being interwoven with the other research processes as previously explained: the requirements of the potential solution are defined in the relevance cycle and the creation and evaluation methods in the rigor cycle. The evaluation should initially take place in laboratory and experimental setup, followed by testing the artifacts in the application domain against the acceptance criteria [93, 156, 224].

Two types of research questions are addressed in design science [223, 224]. One the one hand, there are Design Questions (DQ), also named Practical Questions, which aim at improving the real-world environment through the changes triggered by the creation and adoption of the new artifact. On the other hand, there are Knowledge Questions (KQ), which lead to the discovery of new knowledge without producing a change in the world. These two types of research questions are nested in design science: answering a design question might require to first answer some knowledge questions or answering a knowledge question might require to first design a solution [223]. To differentiate them is nonetheless important as their nature calls for different research methods to be used. Moreover, with regard to the cycles, design questions are under the relevance cycle while knowledge questions under the rigor cycle.

APPENDIX B

RUNNING EXAMPLE TO ILLUSTRATE THE CONTRIBUTION

The running example is extracted from the Reddit discussion board. It can be also found in the released corpus (see Chapter 4). The names of the Reddit users are anonymized. Let's assume the post with its corresponding comments in Figure B.1.

In the first step, each sentence is automatically annotated with speech intentions. The annotated sentences are presented in Table B.1 ("+" before a sentence is used to show the imbrication level of the comment to which the sentence belongs).

Table B.1: Sentences annotated with speech intentions.

Intention_1	**Intention_2**	**Sentence**
ASSERT	GUESS	No diagnosis yet, but.
ASSERT		So for about the past 8 months I've been experiencing what most of you have but never knew it was related to lupus or other AI diseases.
ASSERT		It started with these weird nodules popping up on my arms, legs, and body, they are raised, red, itchy, hot, about the size of a quarter or a little bigger, and after about 4-6 weeks they go down and form what looks like bruises that stay anywhere between 4-8 weeks.
COMPLAIN		Then I started getting server brain fog and as a college student (mid life career change) this killed me this semester.
ASSERT	COMPLAIN	I cannot focus long enough to write a paper, I get so overwhelmed with the work that I just submit subpar work.
ASSERT		I had straight A's going into this semester, my 3rd.
COMPLAIN	GUESS	Now I think I failed 3 classes.

APOLOGIZE		I haven't looked at my grades, I'm afraid.
COMPLAIN		It takes so much for me to concentrate and get out my thoughts.
ASSERT		I finally went to my primary care physician, he ran some blood work.
ASSERT		ANA came back positive, sediment rate was abnormal, and I forgot the other tests he did.
COMPLAIN		(My memory is gone now, I have to write everything down or I forget).
ASSERT		He called me back and told me my test results and symptoms are consistent with Lupus, but I need to see a rheumatologist for diagnosis.
COMPLAIN		So I'm waiting now, unsure what to do, feeling overwhelmed and for the first time in my life depressed.
GUESS		I'm not sure why I posted, just felt like I needed to vent to others that understand.
SUGGEST		+ Look into your college's disability association.
ASSERT		+ Mine allow for accommodations, such as extra time to make up work, if I'm flaring.
ASSERT		+ Makes all the difference.
ASSERT	GREET	+ (I'm changing careers too - hang in there!!)
THANK		++ Thank you for the advice.
ENGAGE		++ I will look into it.
AGREE	SUGGEST	+++ Yes, definitely do this!
ASSERT		+++ Some schools even give you the option to defer your grades (so you can finish your work later - may be helpful until you get a treatment plan squared around).
SUGGEST		+++ I'd also recommend talking to your professors directly, too.
ASSERT		+++ When I was in undergrad, some of mine were willing to do more than what they had to to help me.
DIRECT		++++ Will I still be able to do that even though I have not been "diagnosed" yet?
ASSERT		+++++ It depends on what your school's disability policy is.
ASSERT	SUGGEST	+++++ You should definitely be able to at least talk to your professors, though.
SUGGEST		+++++ It might not hurt to get a note from your doctor that says you're in the process of being tested for x, y & z in case anyone wants proof

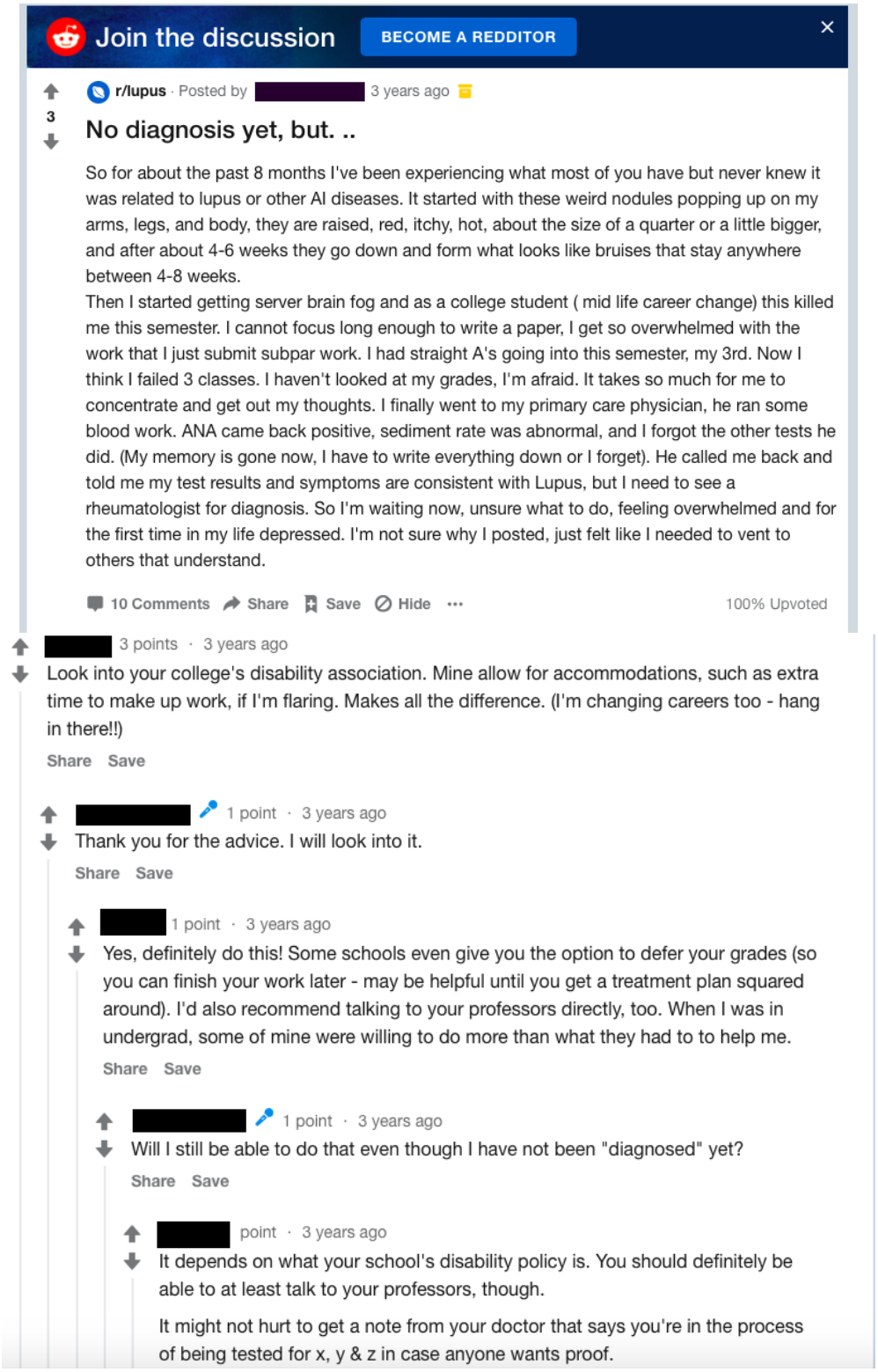


Figure B.1: Example of a Reddit conversation.

In the second step, the conversation tree is formed. In Figure B.2, blue marks the post and green the comments belonging to the thread presented in Figure B.1. This tree will be further used to generate the event log.

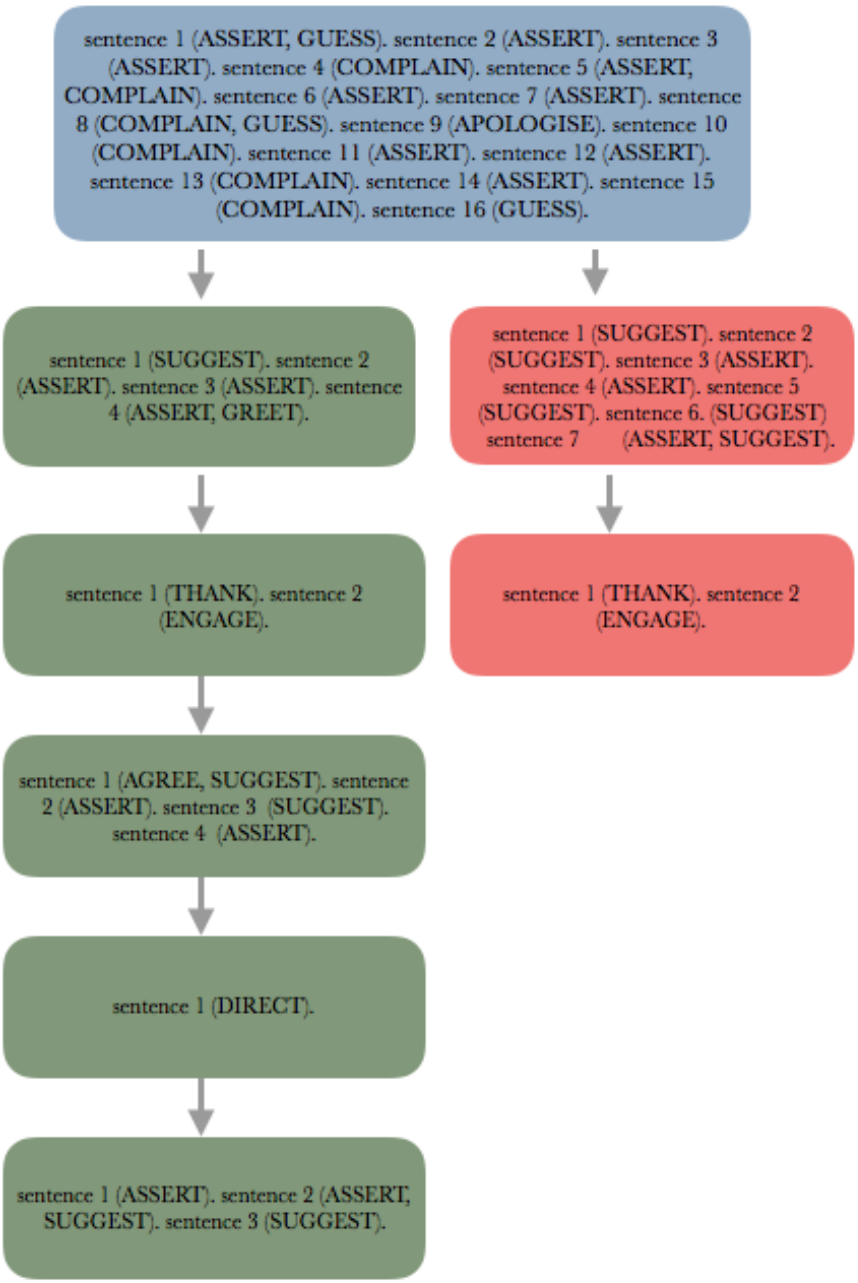


Figure B.2: The conversation tree obtained from the annotated conversation.

What is presented in Figure B.2 with green shapes form a thread in the conversation. In

reality, this conversation had multiple threads. Another thread was added to the tree using red shapes, corresponding to the following comments:

Figure B.3: Another thread of the conversation introduced as a running example.

In the third step, the logs to be used with the process mining tool are generated. Let's say that there are N trees corresponding to N conversations. An event log will be generated for each level in the tree:

- level 0: an event log from all the starting posts (blue shapes);
- level 1: an event log from all the first-level comments (those directly replying to the post);
- level 2: an event log from all the second-level comments (those directly replying to the first-level comments);
- and so on, for all the remaining levels.

For illustration, let's assume that an event log is generated from the conversation above using the first-level comments. The strategy is the following:

- each unique turn becomes a process trace;
- each utterance becomes an event;
- speech intentions are mapped on activities in the following way: if the sentence has multiple speech intentions, pick one randomly by giving lower priority to ASSERT.

The obtained event log is a cvs file, as follows:

```
process id, event id, activity
1,1,SUGGEST
1,2,ASSERT
1,3,ASSERT
1,4,GREET
2,1,SUGGEST
2,2,SUGGEST
2,3,ASSERT
2,4,ASSERT
2,5,SUGGEST
2,6,SUGGEST
2,7,SUGGEST
```

Figure B.4: The resulting event log as a csv file.

If this file is loaded in the process mining tool, the following process model is obtained:

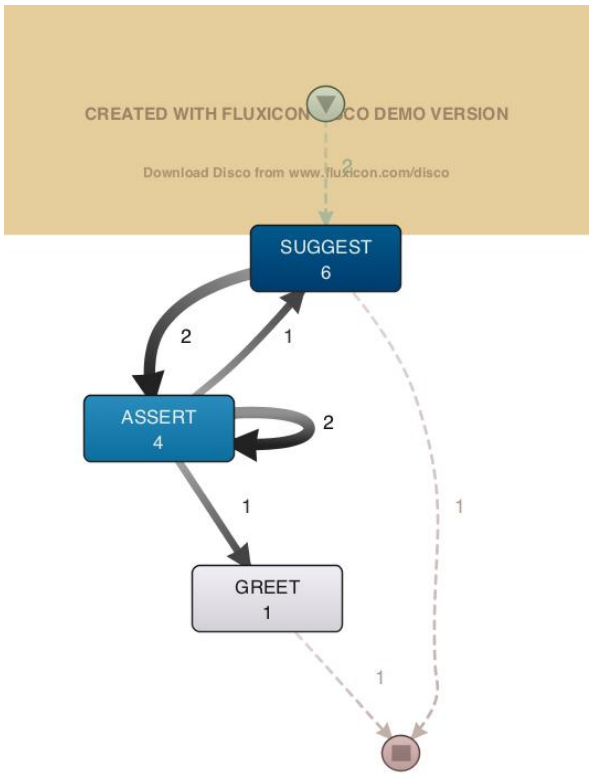


Figure B.5: The process model obtained from the event log.

Another view of the same process is the following:

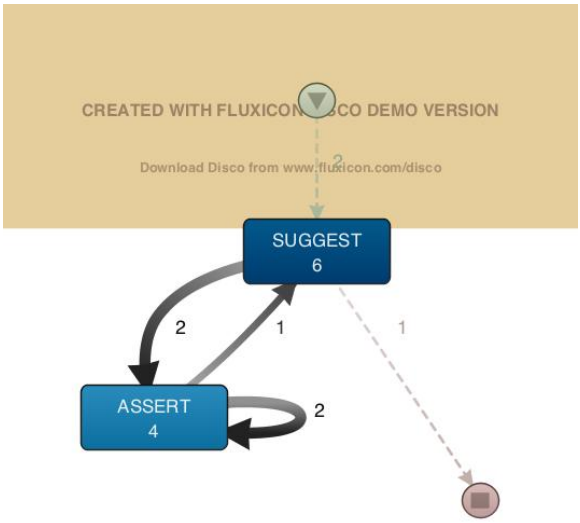


Figure B.6: Another view of the process model obtained from the event log.

These models become more interesting with a lot more data. Let's imagine 10000 comments. Then, more variation is observed in the models showing strategies to build turns and relations among speech intentions. Also, the capacity of the process mining techniques to abstract from data becomes more evident then, compared to transition diagrams which represent every single observation.

Finally, in order to get a global understanding of the corpus, the process models for each level are ordered sequentially considering the position in the conversation trees:

Process Model 1—posts only

|

Process Model 2—1st-level comments only

|

Process Model 3—2nd-level comments only

|

...

Figure B.7: Global overview of a conversation corpus based on the process mining models obtained at each level in the conversation trees.

APPENDIX C

RESEARCH PROCESS FOR THE SYSTEMATIC LITERATURE REVIEW

Problem formulation. In the problem formulation step, the research questions guiding the review are stated. Moreover, a detailed description of the focus and goal of the literature review is provided, while emphasizing on the link to the research questions [42]. Finally, criteria for including or excluding a piece of research are established. Several requirements with regard to the definition of the inclusion/exclusion criteria exist. It should be comprehensive and explicit in order to ensure that any article that might appear in the studied research area could be selected or excluded only by referring to the defined criteria. Also, the definition should be enough detailed so that the article inclusion or exclusion lead to the same result if replicated by other researchers [164]. The inclusion/exclusion criteria are often reviewed during the data collection step.

Data collection. The data collection step determines which sources could lead to relevant publications and how these relevant publications should be identified. Specifically, it implies multiple actions or decisions. Frequently, electronic databases containing academic works within the domain of interest are identified. Then, a search query for the identified databases is defined driven by the research goals, foci and questions. When possible, the manual exploration of a publication corpus can be also chosen instead of relying on a search engine. At this point, the type of search strategy must be decided. Four types of search strategies are identified in the literature [42]: *exhaustive*—when all the existing resources on a certain topic are located and analyzed, *exhaustive* with selective citation—when an exhaustive process is bounded by clearly defined decisions that are motivated (e.g. consider only journal articles), *representative*—when a selection of publications considered representative for that domain is made, and *pivotal*—when only the publications central to the domain of study are considered.

Usually, the collection of possibly relevant publications does not end with searching the targeted databases. References mentioned in the already collected resources that appear as highly

relevant are also analyzed. In case of exhaustive or exhaustive with selective citation strategies, this process continues until saturation is reached, meaning that no new relevant publication is found. In the search process, records of the search queries for each academic database and the number of resources returned by the search engines should be carefully maintained.

Data evaluation. In the data evaluation step, the relevance and the quality of the publications collected during the previous step are assessed. The result is a final corpus of publications that is included in the literature review. For the relevance, the inclusion and exclusion criteria defined in the problem statement are used. For the quality assessment, a checklist should be defined. This step is finished by deciding how the selection of the relevant and qualitative publications should be conducted. One could exclude an article by just reading the title, the title and the abstract or the entire text. Moreover, there are multiple levels of understanding of the studied content: *preliminary*—when quickly reading is used for getting oneself familiar with the paper, *comprehensive*—when the concepts and research terms are grasped, and *analysis*—when one seeks to thoroughly understand each part of the study [19].

Analysis and interpretation. The analysis and interpretation step consists in two activities. On the one hand, the reviewer should decide what data must be extracted from each publication. The result of this decision is a conceptual framework that enables the literature analysis and interpretation. Specifically, it could be a classification schema with multiple facets [159] or a coding book describing essential themes and their variables [164]. In the process of data extraction, the reviewer could realize that the conceptual framework is incomplete and an evolution is necessary. Thus, an iterative process for defining the conceptual framework is often the case.

On the other hand, the reviewer should decide how the data is analyzed and interpreted. There are multiple means of analysis: *meta-analysis*—use of statistical approaches to integrate the outcomes of multiple studies, *comparative analysis*—comparisons based on logical simplification, *thematic analysis*—papers related to the same theme are counted (similar to mapping studies [159]) and discussed, and *narrative summaries*—a qualitative review focused on narrative explanations [164]. In the interpretation of the extracted data, relations between various facets of the conceptual framework are analyzed and hypotheses can be formulated [164]. Overall, this step leads to a deeper understanding of the research area.

Public presentation. In the public presentation step, a report with the findings and the process followed to reach them is created. The report may be organized on historical, conceptual or methodological grounds. The perspective is expected to be neutral when reporting quantitative research findings, but more subjective stances may also exist when the analysis and interpretation are performed qualitatively.

APPENDIX D

LITERATURE SEARCH RESULTS

The following table presents how the established search query was used exactly in each digital library that was queried and how many publications were returned by the search engine of the target digital library.

Digital library	**Search query**	**Hit count**
ACM	("speech act" OR "dialogue act") AND ("automatic classification" OR "automatic discovery" OR "automatic annotation" OR "machine learning" OR "text mining" OR "natural language processing")	166
IEEE Xplore	(("speech act") AND "automatic classification")	51
	("speech act") AND "automatic discovery")	3
	(("speech act") AND "automatic annotation")	4
	(("speech act") AND "machine learning")	50
	(("speech act") AND "text mining")	10
	(("speech act") AND "natural language processing")	124
	(("dialogue act") AND "automatic classification")	29
	(("dialogue act") AND "automatic discovery")	1
	(("dialogue act") AND "automatic annotation")	5
	(("dialogue act") AND "machine learning")	30
	(("dialogue act") AND "text mining")	3
	(("dialogue act") AND "natural language processing")	74
SpringerLink	"speech act" AND "automatic classification"	56
	"speech act" AND "automatic discovery"	12
	"speech act" AND "automatic annotation"	50

	"speech act" AND "machine learning"	573
	"speech act" AND "text mining"	77
	"speech act" AND "natural language processing"	634
	"dialogue act" AND "automatic classification"	45
	"dialogue act" AND "automatic discovery"	4
	"dialogue act" AND "automatic annotation"	4
	"dialogue act" AND "machine learning"	375
	"dialogue act" AND "text mining"	27
	"dialogue act" AND "natural language processing"	343
ScienceDirect	"speech act" AND "automatic classification"	6
	"speech act" AND "automatic discovery"	4
	"speech act" AND "automatic annotation"	5
	"speech act" AND "machine learning"	105
	"speech act" AND "text mining"	21
	"speech act" AND "natural language processing"	110
	"dialogue act" AND "automatic classification"	8
	"dialogue act" AND "automatic discovery"	0
	"dialogue act" AND "automatic annotation"	7
	"dialogue act" AND "machine learning"	64
	"dialogue act" AND "text mining"	5
	"dialogue act" AND "natural language processing"	56
Microsoft Research	("speech act" or "dialogue act") and ("automatic classification" OR "automatic discovery" OR "automatic annotation" OR "machine learning" OR "text mining" OR "natural language processing")	29
AAAI	"speech act"	504
	"dialogue act"	102
ACL Anthology	("speech act" or "dialogue act") and ("automatic classification" OR "automatic discovery" OR "automatic annotation" OR "machine learning" OR "text mining" OR "natural language processing")	126
JSTOR	(("speech act#" OR "dialogue act#") AND ("automatic classification" OR "automatic discovery" OR "automatic annotation" OR "machine learning" OR "text mining" OR "natural language processing")) AND la:(eng OR en)	70

ArXiv	((speech AND act) OR (dialogue AND act)) AND ((automatic AND classification) OR ((automatic AND discovery) OR ((automatic AND annotation) OR ((machine AND learning) OR ((text AND mining) OR ((natural AND language) AND processing))))))	15
DBLP	"speech act" AND "automatic classification"	4
	"speech act" AND "automatic discovery"	1
	"speech act" AND "automatic annotation"	1
	"speech act" AND "machine learning"	4
	"speech act" AND "text mining"	0
	"speech act" AND "natural language processing"	0
	"dialogue act" AND "automatic classification"	1
	"dialogue act" AND "automatic discovery"	0
	"dialogue act" AND "automatic annotation"	2
	"dialogue act" AND "machine learning"	1
	"dialogue act" AND "text mining"	0
	"dialogue act" AND "natural language processing"	0

APPENDIX E

SPEECH ACT VERBS

The trees of speech act verbs presented in [212] and used in the current work to define the speech intention taxonomy are further presented.

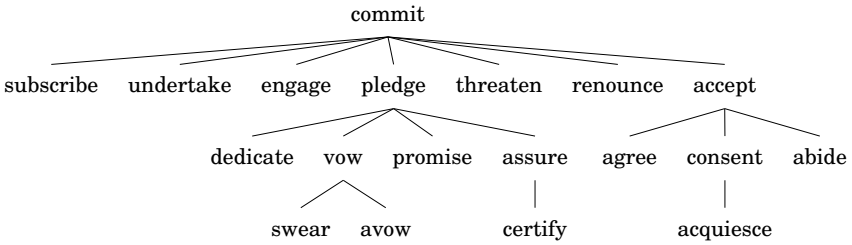


Figure E.1: Semantic tree model for commissives [212].

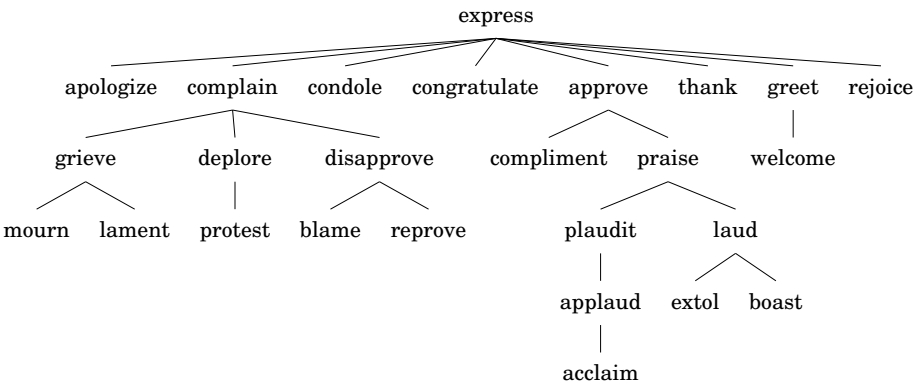


Figure E.2: Semantic tree model for expressives [212].

Figure E.3: Semantic tree model for assertives [212].

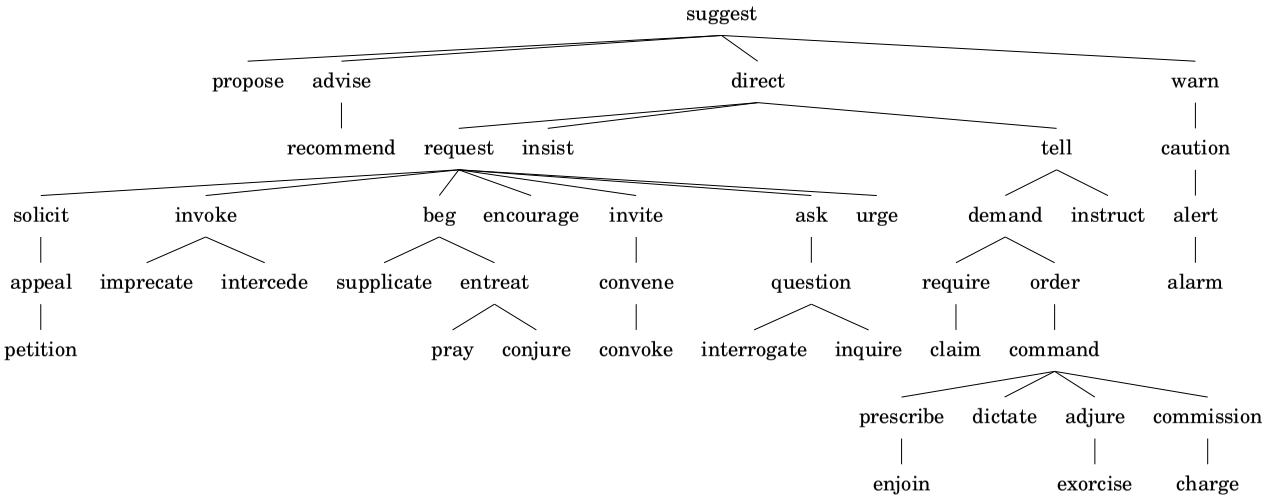


Figure E.4: Semantic tree model for directives [212].

APPENDIX F

EXPERIMENTAL RESOURCES FOR MANUAL ANNOTATION

The email presented below was sent to each participant. In the rest of Appendix F, the instruction document and the document containing the description of the conversations spreadsheet referred to by the first and second links are presented in this order.

Dear Participant,

Thank you for agreeing to take part in our experiment. Your help is extremely appreciated.

The goal is to classify a corpus extracted from Reddit conversations according to the provided instructions. You are given 4 links: One for the Instruction document (URL). One for the document describing the Conversation spreadsheet (URL). One for the Conversation spreadsheet itself (URL). One for the Feedback form (URL).

Start by reading carefully the Instruction document (1st URL). Then, read the description of the conversations we provide (2nd URL). Perform the classification in the spreadsheet (3rd URL) while consulting the Instruction document. Finally, please complete the feedback questionnaire (4th URL). We want to emphasize that there is no good or bad answer. Your judgment is what matters to us.

Thank you once again,

Elena and Dario

Online Conversations' Classification

Every time we speak or write we do it *for an aim*, we want to do something with our words. For example: sometimes we want to give an advice, some other times to share some information, to thank somebody or to express our feelings about something or someone.

We would like to ask you to help us recognize *the aim* of some sentences used in online conversations. We are going to propose you some broad classes among which you will choose, following some instructions regarding how to discriminate between them.

For the names of the classes we chose common English verbs such as REJOICE or THANK, but be careful not to take these verbs as they are used in everyday situations and instead **refer to the provided instructions**.
Please read these instructions very carefully, with all the examples provided, and really try to understand the differences between the classes. This is not an easy task, consider these instructions a form of training.

Presentation of the classes

There are 16 classes, divided into 4 groups. We begin by presenting the 4 groups.

1st group - **Assertives**: they are used for stating information (as being true or false). For example, *I went to buy some bread yesterday*.
Classes: ASSERT, SUSTAIN, GUESS, AGREE, DISAGREE

2nd group - **Expressives**: they are similar to Assertives, but they express the speaker's attitude and feelings towards some events or people. For example, *It's great that you passed your exam*.
Classes: REJOICE, COMPLAIN, WISH, APOLOGIZE, THANK, GREET

3rd group - **Directives**: they are used to influence the actions of the hearer. For example by asking something (*Can you pass me the salt?*) or by giving an advice (*I think you should visit Paris in July*).
Classes: DIRECT, SUGGEST

4th group - **Commissives**: they are used to state that the speaker is going to do something, engaging in a future action. For example: *I will stop smoking starting tomorrow*.
Classes: ENGAGE, ACCEPT, REFUSE

The class OTHER can be used for statements that do not fit any of the proposed classes.

These are straightforward examples, but sometimes it is harder to understand which is the right class for a given sentence. That's why we are giving you a table with further instructions and examples for each class.

Group	Classes	Description	Examples
Assertives	ASSERT	**Plain statement.**	*- The labs billed my insurance for the initial battery of tests.* *- It is extremely photosensitive, and at times a discoid rash will appear on it especially during summers.*
	SUSTAIN	**Statement that increases the certainty of the claim.** It gives explicit reasons or examples to support the claim; it shows confidence.	*- We have a $3,500 deductible plan, so I'd rather use the money for medications.* *- I strongly believe this is the good thing to do.*
	GUESS	**Statement that weakens the certainty of the claim.** It implies the speaker's doubt, possibility or probability.	*- I don't know.* *- So I would say it's kind of both?* *- I say that as someone who does not get this every day!*
	AGREE	**Agreement with a previous statement or positive answer.**	*- He's absolutely right.* *- Yes.*
	DISAGREE	**Disagreement with a previous statement or negative answer**.	*- No you aren't being needy.* *- Doesn't sound like nearly enough going on to be lupus.*
Expressives	REJOICE	**Positive attitude or feeling about something or someone.** It can imply positive attitudes towards negative situations.	*- Great!* *- He always believes me and it's so nice because everyone else doesn't.* *- Oh well, glad it's not just me I guess!*
	COMPLAIN	**Negative attitude or feeling about something or someone.** It can imply dissatisfaction, blame, condolences or grieving.	*- My doctor should have told me earlier about the secondary effects.* *- I feel sorry for your loss.* *- I'm tired of this medication.*
	WISH	**Desire for some future or possible event.** It can imply a desire for negative situations.	*- Hopefully, you have a net of support (family, friends, doctors, nurses).* *- Good luck to you both.*
	APOLOGIZE	**Excuse for some fault.**	*- I apologize in advance if any of these questions are convoluted or redundant.* *- I am sorry I was late this morning.*
	THANK	**Expression of gratitude, acknowledgement, appreciation.**	*- Thanks in advance.*
	GREET	**Salutations of all kinds.**	*- Hi there, from Colorado USA!*
Directives	DIRECT	**Statement that expects the reader to reply, accept or refuse.** It includes orders, requests, questions and invitations.	*- How can I be of help to her?* *- Any advice is appreciated.* *- Go home and decide by tomorrow.* *- Subscribe to this group for information.*
	SUGGEST	**Statement aiming to influence the reader's actions by implying what is good (or bad) for him or her.** It includes advices, warnings, suggestions and recommendations.	*- The best thing you can do is be her pillar of support.* *- Careful with the online advices.* *- Go and buy an aspirin, you'll feel better!* *- Always trust your sensations*
Commissives	ENGAGE	**Promise, commitment or intent of the speaker to do something in the future.**	*- I will follow a new treatment next summer.* *- I'm returning the book tomorrow*
	ACCEPT	**Acceptance to comply with some request, invitation, proposal.**	*- I'll definitely consider your suggestions.*
	REFUSE	**Refusal to comply with some request, invitation, proposal.**	*- I don't think this solution would be the best for me.*
	OTHER	**Statement not fitting any other class.**	

Procedure for classification

The sentences you will classify are presented in conversations.

Classify all the sentences once, **then read again from the beginning** and modify any case that you feel needs a reclassification. This second pass is important because you will have a much clearer idea after having performed the task once.

The **context of the conversation** should be taken into account for understanding the sentences' aims.

Sometimes, **a sentence clearly has multiple aims**, and belongs to two or more classes. In this case, please write down all of the classes. For example:

Example	Classes	Clues
Hi can you help me with this please?	GREET, DIRECT	Greet: "Hi" Direct: It's a question.
Good luck and remember back rubs are necessary for relaxation	WISH, SUGGEST	Wish: "Good luck" Suggest: It's an advice, a recommendation.
The doctor told me in this case one should rest	ASSERT, SUGGEST	Assert: The doctor told me something. Suggest: It's an indirect advice about resting.
Probably everyone wants me to feel better because I am such a bother	GUESS, SUSTAIN, COMPLAIN	Guess: "Probably" captures doubt. Sustain: "Because" in this case strengthens the claim. Complain: "I am such a bother" expresses a feeling of sadness.
This medication makes me dizzy	ASSERT, COMPLAIN	Assert: A statement about medication. Complain: It's used as a way to express some feelings of disappointment.
I am inboxing you because I wrote an absolute essay.	ENGAGE, SUSTAIN	Engage: States the intention to do something. Sustain: "Because" in this case strengthens the claim.

Rules of thumb

1. General rule
   Why is the speaker saying that sentence?
2. Sustain (VS Assert)
   Is the speaker trying to actively be more persuasive?
   Are there markers emphasizing the statement or introducing reasons?
3. Complain or Rejoice (VS Assertives)
   Does the speaker want you to empathise?

Thank you.

UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

Centre de Recherche en Informatique

UNIVERSITÉ SORBONNE NOUVELLE PARIS 3

Information regarding the Conversations spreadsheet

The spreadsheet contains 3 conversations separated by an empty row. Each conversation is presented chronologically.

The first 3 columns in the spreadsheet corresponds to the classes, which will be filled in by you according to the instructions. If you click on a cell corresponding to one of these 3 columns, you will notice that a list with all the possible classes appears.
We provided 3 columns because, in some cases, you might decide that one sentence could be associated with multiple classes. However, in many situations, you might decide that one class is enough. Consequently, complete just the 1st column and leave the other 2 empty. The same applies if you choose 2 classes out of 3.

The 4th column is the sentence to be evaluated and classified.
The 5th column is the id of the Reddit post or comment to which the sentence belongs. This column helps you to identify where a post or a comment starts and ends.

A sentence belonging to a post does not have any prefix while the comments to it are prefixed by a certain number of symbols “+” depending on their level in the conversation. This helps you to figure out the destinatary of the comment.

Example:

Post sentence1	*Hi there, my name is Dario.*
Post sentence2	*Can you tell me where to find a good bar in Paris?*
+ Comment1 sentence1	*Sure thing!*
+ Comment1 sentence2	*I lived in Paris for years, while I was a student.*
+ Comment1 sentence3	*I would definitely suggest to go to Oberkampf area.*
++ Comment2 sentence1	*I totally agree, plenty of bars there!*
+ Comment3 sentence1	*It depends on what kind of bars you like.*

In this example, the post has 2 sentences.
Comment1 is addressed to the author of the post and has 3 sentences.
Comment2 is addressed to the author of Comment1 and has 1 sentence.
Comment3 is a reply to the post again and has 1 sentence.

If you have any questions about the spreadsheet, do not hesitate to contact us before starting the classification.

Thank you!

APPENDIX G

RECOMMENDING PERSONALIZED NEWS IN SHORT USER SESSIONS

Epure E.V., Kille B., Ingvaldsen J.E., Deneckere R., Salinesi C., Albayrak S. (2017) Recommending Personalised News in Short User Sessions. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)* (pp. 121-129). Como, Italy. ACM.

Contributions: E.E.V. and K.B. designed the research and wrote the article. E.E.V. conducted the experiments regarding the news reading dynamics. K.B. conducted the experiments regarding the recommendation policy comparison. I.J.E., S.C., D.R. and A.S. provided feedback on the article.

News organizations employ personalized recommenders to target news articles to specific readers and thus foster engagement. Existing approaches rely on extensive user profiles. However, readers rarely authenticate themselves on the websites of the news publishers. This paper proposes an approach for such cases. It provides a basic degree of personalization while complying with the key characteristics of news recommendation including the news popularity and recency, and the dynamics of the reading behavior. We extend existing research on the dynamics of news reading behavior by focusing both on the progress of the reading interests over time and their relations. The reading interests are considered in three levels: short-, medium- and long-term. Combinations of these are evaluated in terms of value added to the recommendation performance and ensured news variety. Experiments with 17-month worth of logs from a German news publisher show that the most frequent relations between news reading interests are constant over time but their probabilities change. Recommendations based on combined short-term and long-term interests result in increased accuracy while recommendations based on combined short-term and medium-term interests yield higher news variety.

G.1 Introduction

The digital transformation has been a game changer for legacy news media. Keeping readers loyal has become highly competitive for publishers as their new business models rely on advertisement or related revenues [144]. The personalized recommendation has emerged as a popular way to tackle this challenge by automatically suggesting news to online readers. However, creating effective solutions faces particular constraints [83, 131]. Large publishers release hundreds of news daily, implying that they must deal with fast-growing numbers of items that get quickly outdated and irrelevant to most readers. News readers exhibit more unstable consumption behavior than users in other domains such as entertainment. External events, e.g. breaking news, affect reader interests [83]. In addition, the news domain experiences extreme levels of data sparsity.

Existing news recommender solutions address these challenges to some extent. Most of them suggest fresh and popular news. Some consider the dynamics of news reading behavior. However, they rely on the availability of rich user profiles for personalized recommendations. Still, many publishers lack this knowledge, unlike news aggregator services [4, 43] or company blogs [176], as readers tend to consume online news without authentication. Cross-device and cross-browser tracking are technicllly challenging endevaors. Thus, in practice, publishers observe relatively short sessions with fewer than ten clicks on average [48, 55, 131].

As per-user models are unsuitable in short sessions, we propose a new approach to recommendation which ensures basic personalization. The approach combines crowd reading behavior over time and the current user session. Thereby, we model crowd reading behavior for different time frames. However, when providing recommendations, should the models from the same day, from the last weeks, or from the last months be used? In the current work, we analyze the dynamics of the crowd news reading behavior for various time frames and the effects on recommendations. We show that such design choices can significantly affect the outcome.

This work includes four contributions. First, we extend the existing distinction between short- and long-term reading behavior as we establish medium-term reading behavior. The reading behavior concerns readers' interests, which are linked to news categories. Section G.5.1 conveys a detailed description. Second, we identify reading episodes and derive models specifically reflecting engaged reading. Third, we assess the dynamics of the crowd reading behavior, complementary to the related studies in news media. Specifically, we focus on the evolution of the relations between news categories rather than on standalone categories. Fourth, we compare three recommendation policies in terms of performance and news variety: (a) a policy based merely on short-term behavior; (b) a policy integrating short- and medium-term behaviors; (c) a policy integrating short- and long-term behaviors. Experiments are conducted with real data from a German news publisher, spanning 17 months. The long period is chosen to account for fluctuations caused by seasonal or other types of effects.

Our findings show that, in reading episodes, users are likely to read news within the same

category. The relations between news categories are stable in time as the most likely target categories from a given source stay the same. However, their priorities, represented by transition probabilities, change every 1 to 4 months. Augmenting the short-term behavior with knowledge about the long- and medium-term behavior improves the recommendations. The policy combining short- and long-term interests yields higher accuracy than combining short- and medium-term interests. Contrarily, combining short- and medium-term interests yields higher variety.

Further, the formalization of the proposed approach is presented. The experiments are described in Section G.3 and discussed in Section G.4, followed by the related work in Section G.5. Finally, conclusions are drawn and directions for future works are indicated.

G.2 Defining Reading Behavior in News

We consider an online environment in which a publisher provides a collection of news articles to interested readers. Let $\mathscr{U} = \{u_m\}_{m=1}^{M}$ represent the increasing set of visitors. Further, let $\mathscr{I} = \{i_n\}_{n=1}^{N}$ represent the increasing collection of news articles. The publisher observes how visitors act on the website. In particular, the publisher keeps track of events whenever a visitor reads an article. Let $\theta = (\theta_u, \theta_i, \theta_t, \theta_c)$ represent such an event where the individual variables refer to visitor, article, time, and context. The publisher observers a sequence of events $\Theta = \{\theta^{(\alpha)}\}_{\alpha=1}^{A}$ such that $\theta_t^{(\alpha)} < \theta_t^{(\alpha+1)}, \forall \alpha$. We represent the reading behavior of individual users as $\Theta_u = \{\theta : \theta_u = u\}$. We assume that visitors have limited time and desire to read articles. In particular, the number of news articles a visitor is willing to read is $X \ll N$.

The publisher tries to engage visitors by providing a small set of suggestions, every time an article is read. Formally, the publisher employs a policy π which takes a given event $\theta^{(\alpha)}$ and automatically produces a ranked list of suggestions $S^{(\alpha)} = \{s_k \mid s_k \in I\}_{k=1}^{K}$. The publisher monitors how visitors react upon the received suggestions in order to drive policy improvement. The policy that produced the suggestions gets credited with the reward R whenever a visitor reads any of the suggested articles:

$$R(\pi, \theta^{(\alpha)}, S^{(\alpha)}) = \begin{cases} 1 & \text{if} \quad \exists \beta > \alpha, \theta^{(\beta)} \in (\Theta_u \cap S^{(\alpha)}) \\ 0 & \text{otherwise} \end{cases} \tag{G.1}$$

For practical purposes, the publisher disregards future events if the visitor is inactive for more than a specified time τ_u. Publishers can employ a variety of policies $\pi \in \Pi$. Their objective is to find the policy that maximizes the cumulative rewards:

$$\pi^* = \arg\max_{\pi \in \Pi} \sum_{\theta \in \Theta} R(\pi, \theta, S). \tag{G.2}$$

We investigate three policies: (a) the short-term news reading; (b) the long- and short-term news reading; (c) the medium- and short-term news reading. These are further described.

Editors put the most recent and significant news at the start of newspapers. Analogously, our first policy suggests articles which have been popular recently. We refer to this policy as *baseline*

and it corresponds to the short-term news reading interests. The baseline has associated a list L of fixed size ϵ. As the system observes another event, θ_i is added to L. If adding θ_i exceeds ϵ, the oldest element is dismissed. Thereby, L continuously stores the most recently read articles. At the same time, the more popular an article is, the more often it will appear in L. The baseline policy suggests elements from L, most recently added and different from the article being currently read by the user u. Also, the recommendation has an implicit popularity bias as the probability of choosing a list of articles $S^{(\alpha)}$ depends on $f_{i|L}, i \in S^{\alpha}$, the frequency of article i in L.

The remaining two policies enrich the baseline with information about the news reading process—how readers transition between news categories. A stochastic process models a random system changing over time. Formally, if D is a subinterval of $[0,\infty)$, a continuous-time stochastic process is a set of random variables $\{X_d\}, d \in D$. If we restrict $D = \mathbb{N}_0$, we obtain a discrete-time stochastic process. Let $\Xi = \{\xi\}_{v=1}^{V}$ be the finite set of news categories, corresponding to the random variables of the news reading process. Let $\xi(\theta_i)$ refer to the category assigned to the article i. Hence, we can transform the reading behavior of an individual user u, $\Theta_u = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(\alpha)})$ into a sequence of categories: $\Xi_u = (\xi(\theta_i^{(1)}), \xi(\theta_i^{(2)}), \ldots, \xi(\theta_i^{(\alpha)}))$. The Markov property characterizes a stochastic process whose current state captures all information necessary to compute the probability of the next state [113]. For the news reading behavior, this is formally defined as:

$$\Pr[X_{d+1} = \xi_v | X_d = \xi_w, X_{d-1} = \xi_{w-1}, \ldots, X_0 = \xi_{w_0}] = \Pr[X_{d+1} = \xi_v | X_d = \xi_w] = p_{ij}, \forall d \in \mathbb{N}_0, \xi_{w_0}, \ldots, \xi_w, \xi_v \in \Xi \tag{G.3}$$

A transition matrix over all permutations of two categories ($|\Xi| = V$) represents the dynamics of such as system:

$$T = \begin{bmatrix} p_{11} & \cdots & p_{1V} \\ p_{21} & \cdots & p_{2V} \\ \vdots & \ddots & \vdots \\ p_{V1} & \cdots & p_{VV} \end{bmatrix} \quad p_{vw} \geq 0, \sum_{w}^{\xi_w \in \Xi} p_{vw} = 1, \quad \forall v, \xi_v \in \Xi \tag{G.4}$$

The transition probabilities p_{vw} can be estimated from observations as the ratio between the frequency of transitions from category ξ_v to category ξ_w and the total number of transitions from ξ_v. If the transition matrix T stays constant as the system evolves, the stochastic process is referred to as *First-Order Markov Process*.

We use the defined reading behavior process to create a new recommendation policy on top of the baseline. For each category ξ, there is a separate list L_ξ used to model short-term reading interests. For each event θ, we determine the category $\xi(\theta_i) = \xi_v$. Subsequently, we select the most likely category for the next article: $\xi^* = \arg\max_{p_{vw}} T$ and return suggestions from L_{ξ^*}. We distinguish the medium-term news reading interests from the long-term ones based on the observations used for estimating T. The matrix corresponding to the long-term behavior contains all the existing observations. The matrix corresponding to the medium-term behavior contains observations from a limited period only.

G.3 Experiments

The research questions we address in this paper are:

Q_1 How stable in time is the news reading behavior, modeled as Markov processes over news categories?

Q_2 How do different strategies of combining long-, medium- and short-term news reading interests in a recommender compare in terms of performance and news variety?

G.3.1 Data Description

The data set is provided by a German news publisher that accommodates on average 3 million visits per week. It consists of ≈ 196.6 million events, which span 475 days from August 2014 to December 2015. Data from June 16 to June 27, 2015 is missing due to technical problems with the logging system. An event is generated and logged when an article is accessed on the news publisher's website. The events are logged with the following information: the user session identifier, the event time-stamp, and the article link and meta-data which includes the associated categories. The categories are manually assigned by editors, each time an article is uploaded on the website, and are a set of pre-established keywords (see Table G.1). We rely on editors' experiences to assign categories in a consistent manner. Most of the user sessions have less than 10 events as shown in Figure G.1.

Table G.1: News categories and their associated codes

Cars, Motor, Traffic	1
Science, Communication	2
Games, Virtual World, Toys	3
Politics, Business, Economics	4
Travel, Tourism, Navigation	5
TV, Radio, Video, Photo	6
General News	7
Professional, Career	8
Computers, Technology	9
Family, Education, Leisure	10
Banking, Finance, Insurance	11
Health, Sports, Nutrition	12
Real Estate, Home, Gardening	13
People, Relationships	14
Fashion, Lifestyle, Culture	15

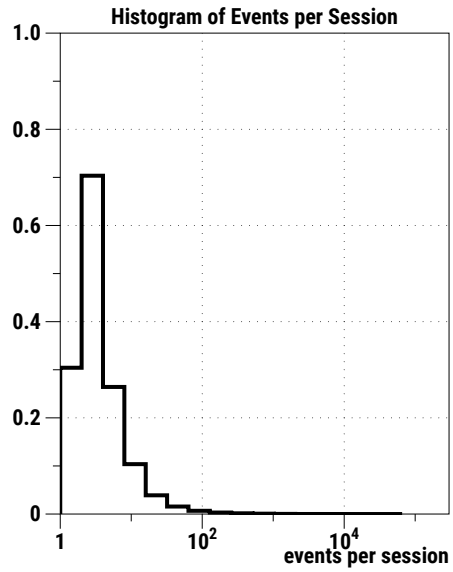


Figure G.1: Histogram of events per session (log scale)

G.3.2 Data Transformation

Before the experiments, the data has been pre-processed. First, we eliminated events associated to the default session. This session captures all events with cookies disabled such that individual users cannot be tracked. Second, we chronologically sorted the events within sessions. Third, we segmented sessions into reading episodes. Therein, we cut a session if events occurred in less than 60 seconds—meaning that the user did not engage in reading, or longer than 3600 seconds apart—meaning that the user started a new reading episode. These thresholds were set after observing and questioning our personal circle about the time spent on article reading. Other sources also suggest an average of 20 minutes spent per article[1].

The transition matrices over news categories have been computed as follows. Consecutive events in a reading episode become a transition, the first being the source, the second the target. Multiple transitions emerge if either the source or target events have several categories assigned. In this case, the Cartesian product of the source and target category sets is calculated, each obtained element becoming a transition.

G.3.3 News Reading Dynamics

The first research question aims to assess the dynamics of the news reading behavior modeled as a Markov process over news categories. We have analyzed the stationarity of the Markov process using a method based on the χ^2 statistical test [15]. The method starts with computing the transition matrix for the complete period τ and the transition matrices for all consecutive

[1] http://contentmarketinginstitute.com/2016/01/visitors-read-article/

sub-intervals $t \in \tau$. Then, these local transition matrices are tested against the overall transition matrix for statistical difference. We formulate the null hypothesis and the alternative hypothesis:

$H_0 : \forall t \in \tau : p_{vw|t} = p_{vw} \leftrightarrow H_a : \exists t \in \tau : p_{vw|t} \neq p_{vw}$

- p_{vw} denotes the estimated transition probability from state ξ_v to state ξ_w for the entire period τ;
- $p_{vw|t}$ denotes the estimated transition probability from state ξ_v to state ξ_w only for the sub-periods $t \in \tau$;

Given that there are at least two positive values in each row v of the overall transition matrix T, the χ^2 test is computed as follows:

$$Q = \sum_{t \in \tau} \sum_{\xi_v \in \Xi} \sum_{\xi_w \in \Xi_v} z_{v|t} \frac{(p_{vw|t} - p_{vw})^2}{p_{vw}} \approx \text{asy}\chi^2 \left(\sum_{\xi_v \in \Xi} (a_v - 1)(b_v - 1) \right) \tag{G.5}$$

- $z_{v|t}$ denotes the number of observed transitions from state ξ_v in sub-period t, it could be 0;
- $\Xi_v = \{w : p_{vw} > 0, \xi_w \in \Xi\}, \xi_v \in \Xi$; contains all target states observed from state ξ_v for the entire period τ;
- Q has an asymptotic chi-squared distribution (asyχ^2) with the number of degrees of freedom computed as sum over all states $\xi_v \in \Xi$ by considering two terms: a_v is the number of sub-periods t for which transitions from state ξ_v are observed; $b_v = |\Xi_v|$ is the number of positive values in the row v of the transition matrix for the entire period τ.

The test could be also adjusted to assess how much a certain sub-period t differs from the complete period τ. In this case, the outer sum in (Eq. G.5) has 2 terms ($|\tau| = 2$): the period t and the transition matrix computed for all the other periods from τ except t.

The first prerequisite to conduct the stationarity analysis on our data set is to decide the magnitude of the sub-periods t: days, weeks or months. Periods spanning days or weeks were excluded. In a descriptive analysis, we observed that breaking news changes the local reading pattern. Also, we randomly selected consecutive days and weeks, and ran the stationarity test. The results showed a significant matrix variance (p-value $p < 0.001$). Consistent with this, Bickenbach and Bode [15] claim that while more granular sub-samples are preferred, they should not be too small. Otherwise, non-stationarity could emerge from the test, even though the process is Markov [15]. Eventually, the sub-periods t were set to 1 month.

When working with matrices of millions of events, a difference caused by few thousands transitions is not necessarily significant. However, as the Chi-Square test is very sensitive to the number of observations, it could yield non-stationarity [67]. For this reason, we decided to correct the transition frequencies by a factor of 0.001. The values were brought from big-data (millions) to a magnitude equivalent to what has been observed in similar studies (thousands) [15]. After correction, the frequencies still reached tens of thousands, which we consider a representative sample size.

G.3.4 Recommendation Policy Comparison

The second research question aims to compare the three policies introduced in Section G.2 in terms of recommendation performance and ensured news variety. Publishers focus on maintaining the engagement of their readers. Thus, one way to measure performance is through the cumulative rewards $R(\pi, \Theta)$. Frequently, they normalize the rewards by the number of requests to obtain the click-through rate:

$$(CTR) = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} R(\pi, \theta, S) \tag{G.6}$$

More sophisticated evaluation metrics such as the ranking-based ones—normalized discounted cumulative gain or mean reciprocal rank— exist but they cannot be applied in conditions with insufficient user feedback.

Besides monitoring how well readers' preferences are met, we look at the overall variety of suggestions. Having a more diverse set of suggestions can lead to readers experiencing serendipity. Let $S^{(A)}$ refer to all recommendations produced by applying the policy π to the upcoming events Θ. Further, we define the number of times an article i is recommended: $\nu_i = \sum_{s \in S^{(A)}} \mathbb{I}(s = i)$, where $\mathbb{I}$ refers to the indicator function which returns 1 if the condition applies and 0 otherwise. Finally, let $I^{(K)}$ refer to the K items with highest ν_i. We quantify the overall news variety as,

$$\delta = 1 - \frac{\sum_{i \in I^{(K)}} \nu_i}{|S^{(A)}|}. \tag{G.7}$$

In other words, the number of recommendations subsumed by elements of $I^{(K)}$ is normalized by the total number of recommendations and subtracted from 1. Thus, $\delta \in [0, 1]$ with $\delta = 0$ referring to the case that all readers received the identical K suggestions and $\delta = 1$ signaling that all readers received disjoint sets of suggestions.

We use the same data for this evaluation as has been used for the stationarity analysis in the previous section. This allows us to have a sound and transparent base for comparing the recommender outcomes, especially for changing periods or strongly deviating months. The evaluation follows a sliding-window approach. First, the three news recommendation algorithms are initialized with 10000 events from September 2014 and the initial transition matrix is computed for August 2014. Subsequently, all events are processed in chronological order. Each recommendation algorithm computes a list of $K = 4$ suggestions. The choice of 4 suggestions corresponds to the number of suggestions displayed on the website of the news publisher. Then, for each user associated with the suggestions, it is checked whether they accessed one of the recommended articles within an hour. Having exceeded the 1-hour limit, the suggestions are discarded. Then, the cumulative rewards and news variety are hourly computed. This yields a total of 11355 measurements.

G.4 Results and Discussion

This section presents the results of our experimentation. Section G.4.1 is devoted to the stationarity analysis. Section G.3.4 presents the observations regarding the recommendation policies.

G.4.1 News Reading Dynamics

We assessed the stationarity of the news reading behavior, modeled as a Markov process over news categories, for the entire period of August, 2014–December, 2015. Using the Chi-Square test (Equation G.5), we obtained $Q = 8709.744$ with $\text{df} = 1778$ degrees of freedom, leading to the rejection of H_0 ($p < 0.01$). Thus, the transition probabilities fluctuate during the 17-month period. Further, we have checked whether longer periods exhibit stationary patterns. For getting better insights into how to find these periods, and whether they existed, we used the Chi-Square test (Equation G.5) adjusted for computing individual period difference: $|\tau| = 2$; the first sum term is t, the assessed month; the second sum term is the rest of the period τ except this month. Then, the resulting Q values per month were plotted. Thereby, we observed a major structural break from March 2015 to April 2015. Then, the test was re-run for the period August 2014 and March 2015 still rejecting H_0 with $Q = 1124.85, \text{df} = 938, p < 0.01$. However, when Q values for each month of this period were plotted, a potential strong similarity has been observed between September, 2014 and February, 2015. Indeed, the Chi-Square test confirmed this similarity, accepting the null hypothesis with $Q = 700.51, \text{df} = 645, p > 0.05$. We repeated this procedure for the period April, 2015–December, 2015. Finally, the following homogeneous periods have been discovered:

- Sep. 2014–Feb. 2015: $Q = 700.5, \text{df} = 645, p > 0.05$,
- Apr. 2015–Jul. 2015: $Q = 6.56, \text{df} = 48, p > 0.99$,
- Sep. 2015–Oct. 2015: $Q = 14.95, \text{df} = 55, p > 0.99$,
- Nov. 2015–Dec. 2015: $Q = 62.16, \text{df} = 47, p > 0.05$.

Figure G.2 is consistent with these findings. The y-axis represents the transition probabilities greater than 0.05 from category 7—General News, to all news categories; the x-axis reflects the month. The discovered homogeneous periods become visible: the circles for the months within this period line up almost horizontally. Further, we noticed a tendency to read within the same category. The highest transition probability corresponds to category 7, same as the source. Finally, it appears that few target categories are preferred to be read next, after the source category, over the entire period: Science, Communication—code 2; Politics, Business Economics—code 4; Travel, Tourism, Navigation—code 5. However, their priority changes in time. For instance, travel and tourism news are read most frequently together with the general news in the beginning of the year (January–March). We created visualizations for other source categories as well observing similar results (figures omitted for space reason).

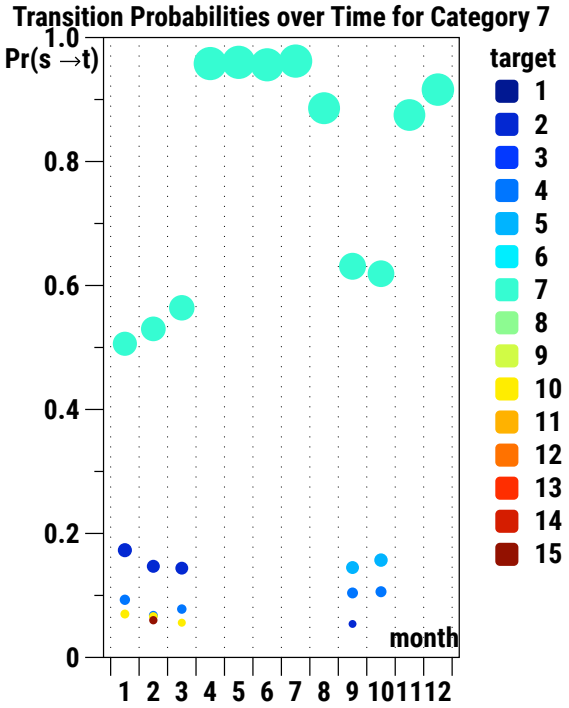


Figure G.2: Probabilities per month for transitions from General News—code 7, to all the other categories in 2015.

Often, in solutions, the model stationarity is taken for granted. However, these results show that news reading behavior has both stable and dynamic parts. The most preferred news categories read after a given category seem to be stable. By contrary, the transitions between these news categories are stationary for limited periods of time, with some singular months with strongly deviating patterns (e.g. August, 2014 and 2015). Consequently, the preceding month appears to be often a suitable base for recommendation, except for the months with strongly deviating reading patterns or starting a homogeneous period. Being able to identify the cause of these changes and predict them could be very useful for adjusting dynamically the recommendation. We contemplate that various factors could influence news reading behavior: changes in human habits induced by seasons, structural changes of news publisher websites or very important news events spanning longer periods.

G.4.2 Recommendation Policy Comparison

Figure G.3 and Figure G.4 present the performance results of the proposed strategies for recommendation: short-term interests only (baseline (B)), short- and long-term interests (transition complete (T_c)), short- and medium-term interests (transition 1-month (T_1)). In Figure G.3, each curve reflects the proportion of per-hour measurements, y-value, exhibiting a maximum response rate of x-value. For instance, baseline intersects the ordinate at ≈ 0.27, meaning that about 27%

of per-hour measurements have associated a 0.0% response rate. Alternatively, T_c achieves in about 80% of the cases a response rate up to 10%. Thus, a more distant curve from the top left corner is preferred. Figure G.3 shows the pair-wise comparison of the recommendation strategies: transition 1 month vs. baseline, transition complete vs. baseline, and transition 1 month vs. transition complete. The x-axis conveys the per-hour measurements in chronological order, the left-most point being the first hour of September 1, 2014, the right-most point the last hour of December 31, 2015. The y-axis values illustrate the difference between the response rates of each pair of measurements. Positive values are shown in blue and negative ones in red.

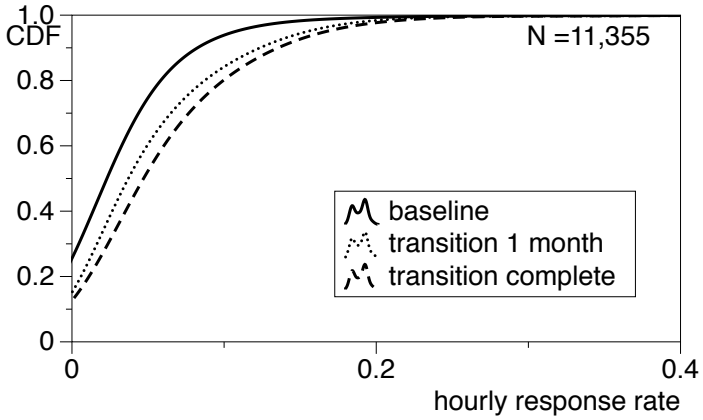


Figure G.3: The cumulative distribution function (CDF) of the hourly response rates for the compared recommenders. The inclusion of long and medium-term interests in the recommendation policy improves the results over using the short-term interests only.

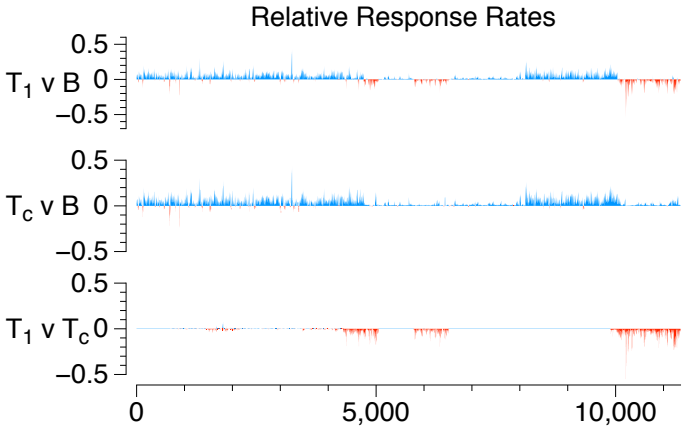


Figure G.4: The pair-wise comparison of the three recommendation policies: T_1 v. B, T_c v. B and T_1 v. T_c. The strategies with long and short-term interests yield consistently the best results.

Both T_1 and T_c achieve higher response rates on average than the baseline. T_c performs similarly to or even better than T_1. Analyzing the granular measurements, we observe that

T_1 falls short of T_c in particular during the months: March, May, November and December, 2015. The sub-performance of T_1 could be explained for March—strongly deviating month from February 2015, its base of prediction, and November—the first month of a new homogeneous period so different than October 2015. Nonetheless, May and December, 2015 have associated low response rates even though they were very similar to their preceding months. This outcome could be explained by what happens in the second part of the recommendation, when the short-term interests strategy is used. The dynamic data structures of articles are sensitive to the latest crowd reading behavior. It appears that there is a division among the crowd reading interests, hypothetically caused by the co-existence of multiple strongly influential news, competing in popularity. Another interesting aspect is that sometimes strongly deviating months or the last months of homogeneous periods are a good base for prediction such as March, July and August, 2015. This indicates that even if the overall reading behavior changes, the most likely transition from a given source category stays consistent. Therefore, even for time-variant periods, the target category associated to a source appears stable.

Further, Figure G.5 plots the histogram with news variety measurements for T_1 and T_c, hourly computed with Equation G.7 and K set to 4. We observe that T_1 varies suggestions better than T_c as shown by the right-most peak in Figure G.5. The long-term interests represent the transition probabilities in the long-run between news categories. Aligned with the literature [43, 202], this reading behavior converges to few most preferred categories chosen in recommendations. However, such behavior is not sensitive to local strong trends such as season-induced changes that could temporally modify the most likely transitions. Contrarily, the medium-term interests strategy is able to overcome this, leading to a higher news variety.

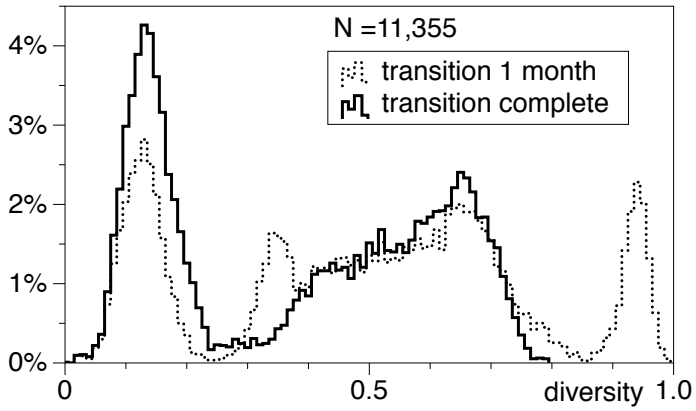


Figure G.5: Analysis of the news variety ensured by T_1 and T_c. The two histograms show the distribution across the spectrum. T_1 has a peak of very high variety, which T_c lacks.

G.4.3 Reflections on News Recommendation

The proposed solution shows that categories provide noticeable improvements despite their simplistic nature and a fair level of personalization when rich user data is missing. Categorical information is readily available, thus it could be implemented as-is in other news recommenders.

Its simplicity ensures scalability and maintainability; the crowd-based models for recommendations and the lack of authentication ensures high privacy. Data sparsity is not an issue as the proposed hybrid solution uses categories, which are limited and mostly stable in time, and the dynamic data structures are pruned to store a fixed number of articles. The proposed analysis of news reading dynamics could support decisions about which data to use for building models.

Our experiments show that by varying the time horizon publishers can trade some accuracy for news variety. Accuracy reflects how well the news recommender anticipates readers' preferences. Variety indicates to what degree readers are exposed to different articles. Which should publishers prioritize? Mark Thompson (CEO of the *New York Times*) [200] lists major challenges that news publishers face as they compete online. Revenue from printed newspapers declines. Publishers have to counteract the downward trend by attracting users online. Hereby, they cannot negotiate as comfortably as they used to for the printed news. Digital social media sites constitute the major players in digital advertising. Thus, successfully harnessing users' attention is crucial to generate online revenues. This suggests focusing on accuracy. Give readers what they want to read to keep their sessions alive.

At the same time, news consumption affects society, economy, and politics. Rasmus Kleis Nielsen (Director of Research, *Reuters*) [146] finds that world-wide people prefer personalized news recommendations. Still, he points out that this may create "filter bubbles". These are spaces in which people consume similar news, reinforcing their existing believes and opinions. They represent a serious danger to social, economic, and political discourse as they prevent exposure to deviating opinions. This favors news variety over accuracy. Finding the right balance between both aspects remains a major challenge for publishers to address in the future. Maximizing profits may lead to social divides. Optimizing news variety may yield economic damage to the publishers. Publishers may have to find additional sources of revenue to compensate reduced accuracy. Readers may have to spend money to have access to high-quality journalism.

G.5 Literature Review

Traditionally, recommender solutions are grouped into two types [2, 73, 105]. A content-based recommender suggests items similar to previously liked ones. A collaborative recommender suggests items by comparing user preferences. Hybrid solutions are frequently reported to perform best in the news domain [4, 16, 43, 108, 127, 128, 130].

G.5.1 News Reading Interests

News readers' interests are seldom expressed explicitly. Common approaches discover readers' interests from click behavior and article categories [126]. In some cases, the categories of the articles are already defined in advance and represented as contextual meta-data [43, 48, 55]. In other cases, categories are discovered automatically and represented either as more granular vocabularies associated to the categories [4, 16, 128], or keywords defining general, well-known topics such as sport or politics [127, 130, 229]. In most of these works, once known, recommenders use the categories as explicit knowledge. Other approaches do not infer users' interests but incorporate them implicitly as built-in features. For instance, Billsus et al. [16] and Li et al. [127] address short-term interests through standard content-based components by recommending articles similar to what users recently read.

Multiple authors distinguish short-term and long-term interests [16, 127, 130]. The long-term interests are considered by most of related works as the users' genuine interests which are less likely to change over time. Contrarily, the short-term interests are discovered from the most recent reading behavior and could represent deviations from the long-term interests triggered by momentary events such as breaking news or interesting reading discoveries. Billsus et al. [16] prioritize the short-term interests model in the formulation of recommendations in Google News as they claim it is more sensitive to changes. In a follow-up work on the Google News system, Liu et al. [130] showed that the short-term interests of an individual user follow closely the short-term interests of the general public. Consequently, in this updated solution, they use the unpersonalized model to address the short-term interests, and the individual user's past reading behavior for predicting categories within the long-term interests.

Another mechanism for manipulating the changes in users' preferences is decaying older interests in favor of the newest ones [43, 127].

We distinguish between short- and long-term news reading interests as well, but we implement them differently. The short-term interests are implicitly captured by providing recommendations according to the real-time popularity distribution of the news most recently read by the crowd. The long-term behavior is explicitly captured with transitions over news categories, dynamically maintained. We also introduce a new type of behavior corresponding to the medium-term interests. Also, our models are created per-crowd and from data reflecting engaged reading.

G.5.2 News Recency, Popularity, and Variety

While some news articles could be relevant even weeks or months after their publication, others get quickly outdated for the majority of users. Thus, recency and popularity are often considered in news recommendation [4, 16, 43, 48, 108, 128].

Billsus et al. [16] report that their recommender takes a list of articles as input, which have been selected in advance by several criteria including recency. Likewise, in the work

of Li et al. [127] a probabilistic graphical model is built with recent articles. Furthermore, Das et al. [43] choose to re-build the recommender models every hour in order to present the freshest information to the users. They use a covisitation metric in recommendation that captures the relative popularity (what is the most popular article visited together with the current read article). Similarly, in Yahoo's news aggregator, articles are selected to represent both new and popular events [4]. For this, they consider article timestamps and key events identified through the event count distribution within the selected article set [4]. Liang et al. [128] set a higher weight to the most recent articles from those discovered by using models for short and long-term interests. External knowledge from Twitter can also be used to determine popularity [44, 108, 160].

While popularity and recency are common news features included in recommender systems, variety is mentioned as a future work in some papers [16, 43] or less relevant in some others [4, 48]. The users' long-term interests model used in [16] does not vary the recommendations for users with similar profiles but per user. Das et al. [43] limit the news variety to the cluster of similar users. Li et al. [127] handle news variety explicitly through random walks in the user-item affinity graph created in advance. Likewise, a probabilistic model is proposed in [105] to address this aspect. In contrast, having analyzed the transition matrices extracted from the CLEF NewsREEL 2014 competition data set, Doychev et al. [48] show that users tend to read news from the same category. Moreover, few dominant categories and clicks between these categories account for the majority of items that are read [48]. Driven by similar reasons, Ahmed et al. [4] provide recommendations within the same category—story, but they also focus to some extent on other tangential categories as they claim users are interested in different aspects of an event (e.g. political, economic).

In the short-term reading policy, we consider popularity and recency, without any external sources. As for news variety, we do not prioritize it for categories, but we ensure that two readers are likely to be recommended different articles in a short time window.

G.5.3 Session-based Recommendation

Session-based recommender systems focus on sessions rather than on complete user profiles. Shani et al. [188] highlight the session-based character of recommender systems deployed in business setups. They move the underlying model from a matrix completion task toward a *Markov Decision Process* (MDP). Businesses encounter session-based recommender systems in domains including music [230], products [98], and news [126]. Deep learning architectures have been applied to session-based recommendation (cf. Hidasi et al. [94], and Tan et al. [116]), achieving promising results. Still, they involve a multitude of parameters to optimize. Our method circumvents the efforts to tune as many parameters albeit sacrificing accuracy to some degree.

In our solution, sessions are analyzed and split in reading episodes in case of too short or long durations between consecutive events. Related works that approached the session splitting to some degree are [128, 232]. Nonetheless, the former considers fast browsing only, setting a time

within 3-250 seconds, while the later splits in sub-sessions of 30 minutes, without reasoning on clicks.

G.5.4 Process Models for Recommendation

Several works model news recommender solutions as probabilistic graphical problems [4, 127, 176]. Li et al. [127] consider the states in the process being both users and articles and the possible transitions are user to article, article to article and article to user. They populate the transition matrix with similarity scores between articles and between users and articles [127]. Ahmed et al. [4] use a transition graph with three types of states: views, clicks, and documents where the views are considered latent variables. The probability of a click to happen is conditioned on the current document, current view, and the previous click [4]. Sahoo et al. [176] choose a Hidden Markov Model where users are the observed variables and the latent classes represent globally preferred consumption patterns per month. Yang et al. [232] propose a topic-aware Markov model for recommending web pages. Segments of sessions—consecutive web page visits, are transformed to temporal states while sequences of articles of the same topic become topical states. The prediction of a page to a user considers the similarity to other users and the probability of observing the user session containing the predicted page [232].

Though effective, the presented models are difficult to maintain with an increasing number of clicks, users, and items. Categories in our approach are a more stable choice. Also, these solutions rely on much longer sessions. Yang et al. [232] report an average of 3700 clicks per user and others in the news domain [127, 128, 130] select only authenticated sessions of minimum 10 clicks. As per-user models are likely unsuitable in short sessions, we aggregate knowledge on news consumption over time and introduce personalization in sessions. Compared with the other works, Sahoo et al. [176] also uses successfully a global model with personalization based on user cases, but with reported high computational costs.

G.5.5 News Reading Behavior Analysis

In order to learn more about user reading behavior and lead the design of the news recommenders, certain works performed high-scale user log analysis over time [59, 127, 130].

Li et al. [127] compared the reading behavior regarding long- and short-term interests. They considered short-term interests being the categories preferred by users within a time window of three days and similarly, the long-term interests were associated with a time period of 15 days. Moreover, three groups of users were created based on the click frequency and separately analyzed. The approach they used for comparison was to plot the Kullback-Leibler (KL) divergence scores, averaged per group, and computed between two consecutive periods for each user. Results showed that long-term goals are quite stable while the short-term ones vary significantly. Liu et al. [130] conducted their analysis by visually projecting the differences between category distribution for each user, for each two consecutive months, and also the click distribution for each category per

month. The news categories are general topics such as world news, sport, entertainment, and the distribution is represented by the averaged number of clicks per category. They discovered that the current preferences of users change, that the general public interests follow the big events trend (e.g. national news are read more during elections) and, to certain extent, that the individual users' interests follow those of the general public. Esiyok et al. [59] analyzed immediate transitions between categories.

Compared to these works, we assess the dynamics of the news reading behavior through the stationarity of the Markov process across news categories. To our knowledge, we are the first to investigate the evolution of the relations among news categories over time.

G.6 Conclusion and Future Works

With the digital transformation, news has started being created and delivered by many entities besides news publishers. News consumers are exposed nowadays to increased sources of information and setups. In this highly competitive ecosystem, news publishers have sought ways to engage existing and attract new readers. Hence, personalized news recommendation has become a key element in their strategy. The proposed solutions aimed to ensure the delivery of fresh and interesting news, taking into consideration the dynamics of reading interests.

Still, limited research has tackled news personalization in hybrid recommenders when extensive user profile are unavailable. Most of the news publishers face this issue as the website authentication is a rare habit among online readers and the cookie-identification has its limits. In this paper, we prove that a recommender based on the dynamics of reading behavior, news popularity, and recency can provide a basic level of personalization and comply with the main domain constraints revealed by the related solutions. Specifically, we design and compare three variants of news recommendation centered on short-, medium-, and long-term reading interests. The short-term interests are captured at the article level, by recommending news recently read by the crowd, biased by popularity. The medium- and long-term interests are captured at the news category level, by predicting the next category to be read from the current user session and the Markov process over all categories. Moreover, the dynamics of the news reading behavior modeled as transitions between news categories is assessed over an extensive period.

The proposed research questions are addressed through experimentation with real data from a German news publisher. First, news readers appear to stay loyal to certain categories, which are frequently read together within a reading episode. Nonetheless, the priority of these relations, captured by transitions, changes in months possibly because of seasons, popular events lasting longer time, or structural changes in publishers' websites. Second, augmenting a short-term interests recommendation policy with the long- or medium-term behavior leads to higher response rates. Considering the transitions between news categories from the previous month leads to a higher variety than the long-term transitions.

Several limits could be identified nonetheless. The proposed solution handles cold start well on the user side. Still, it cannot immediately recommend new items until they have been read at least once. Moreover, the dynamic data structures storing articles can have the popularity manipulated artificially, propagating thus a fake filter bubble. Currently, the recommendation of news could be sometimes redundant as the solution checks only the user's most recent read. Parameters are present in the solution: ϵ the size of article buffers per category, the thresholds involved in the identification of reading episodes. Future work should explore other values of these parameters and their effect on recommendations. Online evaluation will show how well this approach performs under real-life conditions. Also, evaluation with additional data sets and with different session-based recommenders is required.

More content and context features could be considered in the future extending the proposed approach. In the current solution, news categories are manually associated to articles by editors or journalists. An alternative approach is to use more of the news content data directly. If humans are biased or inconsistent in their tagging practices, algorithms for category assignment could assist them or augment in the back-end the manually attached category set. Solutions for classifying news articles into categories or representing the data items in terms of semantic entities with links to conceptual abstractions have already been proposed [4, 26, 45, 128].

External sources of knowledge such as social media could be explored for identifying the most recent trends and injecting news variety in recommendations by external drivers [71, 90, 209]. Suggestions at the category level in the form of a time-line is another option to be explored [205]. For this, the variety of the recommended categories should be ensured.

The literature revealed that there is a lack of agreement about how short- and long-term interests are defined. For instance, long-term is limited to 15 days in [127] and to a month in [130]. However, in order to more easily integrate conclusions emerging from various studies, a sound, well-documented framework should be proposed.

Finally, our approach could be applicable to other domains where the item consumption is sequential within the same session, and items have a relevant feature associated or inferred. This could be the case for music or job recommendation.

APPENDIX H

AUTOMATIC PROCESS DISCOVERY FROM TEXTUAL METHODOLOGIES

Epure E.V., Martin-Rodilla P., Hug C., Deneckere R., Salinesi C. (2015). Automatic process model discovery from textual methodologies. In *2014 IEEE Ninth International Conference on Research Challenges in Information Science (RCIS)* (pp. 19-30). Athens, Greece. IEEE.

Contributions: E.E.V. designed and implemented the method. E.E.V, MR.P. and H.C. designed and conducted the empirical study. E.E.V. analyzed the data. E.E.V and MR.P. wrote the article. H.C., S.C. and D.R. provided feedback on the article.

Process mining has been successfully used in automatic knowledge discovery and in providing guidance or support. The known process mining approaches rely on processes being executed with the help of information systems, thus enabling the automatic capture of process traces as event logs. However, there are many other fields such as Humanities, Social Sciences and Medicine where workers follow processes and log their execution manually in textual forms instead. The problem we tackle in this paper is mining process instance models from unstructured, text-based process traces. Using natural language processing with a focus on the verb semantics, we created a novel unsupervised technique *TextProcessMiner* that discovers process instance models in two steps: 1.*ActivityMiner* mines the process activities; 2.*ActivityRelationshipMiner* mines the sequence, parallelism and mutual exclusion relationships between activities. We employed technical action research through which we validated and preliminarily evaluated our proposed technique in an Archaeology case. The results are very satisfactory with 88% correctly discovered activities in the log and a process instance model that adequately reflected the original process. Moreover, the technique we created emerged as domain-independent.

H.1 Introduction

Although Humanities have increasingly adopted digital research practices, there is still a resistance to changing the traditional research methods mostly based on the manual production of textual sources and qualitative information [78, 153]. This issue is also fueled by a lack of alignment of existing IT systems for the proper management, analysis and operation of the specific type of information resulting from Humanities research [81, 82, 101, 137].

Contrary, other areas such as business have fostered a plethora of research in this respect leading to the actual emergence of several standalone disciplines such as knowledge management [203], method engineering [22] and process mining [1], together with their varied solutions. In this work, we focus in particular on process mining driven by the fact that the Humanities workers keep a record of the processes they follow as textual sources. We believe that by exploiting these textual sources for mining process models we could facilitate and improve the teamwork and the knowledge sharing. Moreover, we could enable a common ground for the comparison, validation and centralization of the applied processes and methodologies. These are already proven benefits of process mining [1].

We then decided to focus on Archaeology in the beginning, given an established collaboration of the authors with this community. We consider the methodology section of the archaeological report a text-based process trace where the archaeologist describes the process the team followed during their work. This section might appear under slightly different titles—methodology, survey methodology, excavation methodology, evaluation methodology, but it has the same goal. Example of archaeological reports can be found at the British public repository[1]. Thus, the research question we aim to answer is: How to use textual methodologies for producing structured knowledge as process models?

Before presenting our solution, we want to clarify several theoretical concepts: activity, process, process instance, process model and process mining. An *activity* is a task, which once completed leads to a full or partial accomplishment of the goal it is related to. The activity is the building block of a process. A *process* consists of a collection of activities and their ordering [1]. The activity ordering is represented through several activity relationships: sequence, parallelism and mutual exclusion [1]. More complex relationships exist but we currently consider only these. The mutual exclusion implies in fact a decision in the process flow: an activity is chosen over others based on some decision parameters. Consequently, a process could be executed in multiple ways. The *process instances* are different executions of the same process. Further, a *process model* is the representation of a process. Finally, *process mining* is the discipline aiming at automatically discovering process models from process traces [1]. Most of the work in process mining so far has considered that the process traces are the logs extracted from the information systems where the processes were executed [1, 53, 115]. However, there are other areas such as Humanities where processes exist and are followed without a similar support from information systems.

[1] `http://www.archaeopress.com/ArchaeopressShop/Public/defaultAll.asp?intro=Home`

We envision a solution in two steps: (1) mine process instance models; (2) mine process models by aggregating process instance models. In this paper we focus on step (1) while step (2) is left as future work. The technique we created, *TextProcessMiner*, generates the log of activities from text and then discovers the process instance model. Compared to other related works [69, 76, 80, 193, 201], *TextProcessMiner* is fully unsupervised and uses natural language processing techniques with a focus on the verb semantics. The technique has emerged as a result of technical action research [225]. We validated it in a specific case in Archaeology though it can be applied to any other domain where processes are described in natural language. The results we obtained during validation are very satisfactory with 88% correctly discovered activities in the log and a process instance model that adequately reflected the original process.

The paper is organized as follows: Section H.2 describes the followed methodology, Section H.3 introduces a demonstrating example of the proposed technique, Section H.4 presents the technique in details, Section H.5 presents the evaluation, Section H.6 discusses the related work and Section H.7 highlights the conclusions and perspectives.

H.2 Methodology

The followed research methodology is Technical Action Research (TAR), an approach centered on technique creation [225]. The goals are to design a technique, use the technique to help the identified stakeholders—in our case humanities specialists—and reflect on the benefits and limitations of the technique in a practical case. Technical Action approach emerged naturally as the most appropriate research methodology for us, for two main reasons. First, all the authors of this research work are affiliated with humanities institutions. This situation allows us to address this research not only in terms of problem identification, but also in terms of actively seeking technological solutions for improving the situation in the real context. This study of artifacts in context, not only as diagnosis but also as intervention, is by definition TAR [225]. Second, Technical Action Research is technology-driven, being employed for learning more about the created technique which might after be used for solving various problems, similar in architecture with the case where it was initially validated [225]. This characteristic of TAR provides more flexibility to study the implications in the real context of the created technique than in other research methodologies as case study [233] or design science [156].

In the next section we discuss further the three phases of Technical Action Research: problem investigation and the design and validation of the technique.

H.2.1 Problem Investigation

In this first phase, we have identified the problem our technique might deal with: process-mining solutions are not currently adapted for exploiting the unstructured information created by the Humanities workers.

Archaeology is a suitable application domain for our approach given the following reasons:

- Archaeological practices, as a humanities discipline, generate a large amount of textual sources;
- There is a growing demand of textual analysis solutions to support humanities;
- The methodology sections from the archaeological reports describe processes in natural language.

H.2.2 Design of the Technique

We carried out an initial screening of the textual corpus in order to identify the objectives of the technique: (1) clean the methodology section, (2) discover the process activities and (3) discover the process instance.

This phase implied the creation of a potential design of the solution for the identified problem. The design was brainstormed separately for each objective of the technique and afterwards integrated to yield the final version.

The development of the technique followed the design we defined in the previous phase. The partial testing results outlined improvements areas, triggering changes in the design. This also reflected in multiple iterations in the development process.

H.2.3 Validation of the Technique

We validated the proposed technique in an archaeological case that allowed us to gather detailed observations. Moreover, the implied in-depth analysis helped us to identify the solution strengths and weaknesses together with potential perspectives.

The selected report consisted of the description of the archaeological works carried out in the site of Villa Magna[2] [62]. The site lies in the Valle del Sacco in Lazio, south of the town of Anagni (Italy). The final report with the methodology section we used was published in 2010.

In order to perform a rigorous validation, the archaeological report used in the case study was different than the corpus we used for testing the solution. Moreover, we defined a protocol with two questionnaires regarding the validation of the logs and the validation of the models. In addition, we created a third group of questions for a preliminary evaluation to discover the solution strengths and weaknesses from the point of view of the potential users.

H.3 Illustration

In this section, we demonstrate the technique using a fragment extracted from the methodology section of the report "Land at Station Road Honeybourne Worcestershire Archaeological

[2] http://www.villa-magna.org

Excavation"[3]. In this case, for example, the methodology section describes the excavation and all other related activities, such as the analysis of the archaeological finds, their recording and documentation. Consequently, a methodology section is a process trace capturing the activities completed during the process and their ordering.

We noticed while analyzing manually the archaeological reports that sometimes the methodology section contained a description of the same process (the excavation), but for different areas. Normally, the text referring to each process instance (the excavation of each separate area) should be identified. Each process instance should be mined separately and all the resulted process instance models should be aggregated for obtaining the final process model. While we recognize this situation and we identified the requirements of the final more general solution, we decided to select from the archaeological reports only those whose methodologies contain a single process instance, leaving the other cases for future work.

In Table H.1, a fragment of the archaeological methodology is presented before and after the cleaning.

Table H.1: Example of a methodology fragment before and after cleaning.

1.14 The archaeological works comprised the mechanical removal of non-archaeologically significant soils, under constant archaeological supervision, using a toothless ditching bucket. The machining ceased when the natural substrate was revealed. All archaeological features were recorded in plan using a Leica 1200 series SmartRover GPS and surveyed in accordance with CA Technical Manual 4 Survey Manual (2012). 1.15 (...) no deposits were identified that required sampling. (...)
The archaeological works comprised the mechanical removal of non_archaeologically significant soils , under constant archaeological supervision , using a toothless ditching bucket . The machining ceased when the natural substrate was revealed . All archaeological features were recorded in plan using a Leica 1200 series SmartRover GPS and surveyed in accordance with CA Technical Manual 4 Survey Manual .

Initially, the text is processed and cleaned. Specifically, the following actions are taken:

- Remove "1.14" and "1.15", the numbering preceding each paragraph;
- Replace "-" by "_" in "non-archaeologically";
- Write each sentence in a separate line;
- Prefix all the punctuation signs with space;
- Remove the text in the parentheses "(2012)" because it does not contain any verb;

[3] http://archaeologydataservice.ac.uk/archives/view/cotswold2_WSM49638/

- Remove the sentence "no deposits were identified that required sampling" because it contains the negation in the form of "no" + noun. However, keep the sentence with "non-archaeologically" because the negation in the form of 'non' + adverb does not induce a negation in the semantics of the whole sentence.

Further, the objective is to discover the log of activities. During the initial screening of the methodology sections, we noticed the activities were introduced most of the time by verbs and in some few cases by nouns derived from the corresponding verbs (e.g. "the removal of spoil"). An activity by definition implies someone doing something[4]. Consequently, two parts compose an activity: the verb and its object(s). The propriety of the verb taking objects is called transitivity[5].

Existing works in natural language processing support the identification of verbs and their objects in text: the parsers. A natural language parser is able to grammatically analyze the structure of a sentence and produce a *treebank*. A treebank is a representation of a sentence as a tree with annotations at different levels: clause level, phrase level and word level [135]. At the clause level, for example, the speech tag S marks a simple declarative clause while the tag SBAR marks a clause introduced by a subordinating conjunction. At the phrase level, the most interesting for our objective are NP—which marks a noun phrase, VP—which marks a verb phrase and ADJP—which marks an adjective phrase.

Table H.2: Treebank for the first sentence of the fragment.

Sentence 1

```
(ROOT
 (S
  (NP (DT The) (JJ archaeological) (NNS works))
  (VP (VBD comprised)
   (NP
    (NP (DT the) (JJ mechanical) (NN removal))
    (PP (IN of)
     (NP (ADJP (RB nonarchaeologically) (JJ significant)) (NNS soils))))
   (, ,)
   (PP (IN under)
    (NP (JJ archaeological) supervision)))
  (, ,)
  (S
   (VP (VBG using)
    (NP (DT a) (NN toothless) (VBG ditching) (NN bucket))))
 (. .)))
```

Finally, at the word level, each word is marked with a speech tag: NN for noun singular, NNS for noun plural, VB for verb, VBG for verb in "-ing" form, VBN for verb in past participle form etc.

[4] http://www.oxforddictionaries.com/definition/english/activity

[5] http://en.wikipedia.org/wiki/Transitive_verb

A complete list of all the tags can be found online[6]. The treebanks are presented in Tables H.2, H.3 and H.4.

All the VP and ADJP sub-trees are checked if they contain at least one transitive verb. To identify the verbs we extract the leaf values of the VB, VBD, VBN and VBG sub-trees. The sub-trees containing verbs are highlighted in Tables H.2, H.3 and H.4 too. The values "was" and "were" are not considered because the verb "be", though transitive, is not an action verb and, in this situation, it has an auxiliary role in the passive voice.

Table H.3: Treebank for the second sentence of the fragment.

Sentence 2

```
(ROOT
 (S
  (NP (DT The) (NN machining))
  (VP (VBD ceased)
   (SBAR
    (WHADVP (WRB when))
    (S
     (NP (DT the) (JJ natural) (NN substrate))
     (VP (VBD was)
      (VP (VBN revealed))))))
  (. .)))
```

Table H.4: Treebank for the third sentence of the fragment.

Sentence 3

```
(ROOT
 (S
  (NP (DT All) (JJ archaeological) (NNS features))
  (VP (VBD were)
   (VP
    (VP (VBN recorded)
     (PP (IN in) (NP (NN plan)))
     (S
      (VP (VBG using)
       (NNS (NP (DT a) (NNP Leica) (NNP 1200) (NN series) (NNP SmartRover)
(NNP GPS)))))
     (CC and)
     (VP (VBN surveyed)
      (PP (IN in) (NP (NN accordance)))
      (PP (IN with) (NP (NNP CA) (NNP Technical) (NNP Manual) (CD 4) (NNP
Survey) (NNP Manual))))))
   (. .)))
```

[6] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Next, we discover the objects for each verb by checking if the sub-tree introducing the verb has, as right sibling, a NP sub-tree. This is the case for: Sentence 1—"(VBD comprised)", "(VBG using)"; Sentence 3—"(VBG using)". For all the others, it is checked if their corresponding VP or ADJP sub-tree or any of their ancestors have a NP sub-tree as left sibling. The first NP clause found is taken. Additionally, if a node of type S is encountered the search stops. This is the case for: Sentence 2—"(VBD ceased)", "(VBN revealed)"; Sentence 3—"(VBN recorded)", "(VBN surveyed)". For "(VBG ditching)" in Sentence 1 no objects are found following the defined method, therefore it is not considered an activity. The verbs and the nouns in the NP elements are extracted and lemmatized, obtaining the activities in Table H.5.

Table H.5: Activities extracted from the fragment.

Sentence 1	**Sentence 2**	**Sentence 3**
1.comprise removal_of_soil	3.cease machining	5.record feature
2.use toothless_bucket	4.reveal substrate	6.use leica_series_smartover_gps
		7.survey feature

Finally, we discover the relationships between the activities by applying a set of rules from a knowledge base we define. The special symbols we use for representing these relationships and the final process are explained in Table H.6.

Table H.6: Symbols in the process model representation.

Symbol	**Usage**	**Example**
_	For activity names: when the object is composed of multiple nouns or when the verb has a particle	area_of_trench, carry_out
→	For activity relationship: when an activity follows another activity (*sequence*)	excavate trench → inspect soil
||	For activity relationship: when two or more activities are executed in the same or overlapping time (*parallelism*)	excavate trench || collect find
x	For activity relationship: when there is a decision between two or more activities (*mutual exclusion*)	take photograph x draw plan
()	For activity relationship: when influencing the precedence of the relationships	(excavate trench || collect find) → draw plan

Prior, the sentences are transformed:

1. by replacing the verbs from activities with their tags;

2. by keeping key structures as prepositions, conjunctions, punctuations; and

3. by replacing everything else with the placeholder "...".

The results of the pre-processing are presented in the first line of each sentence in Table H.7. The rules we apply are presented in the second line and the result after applying the rules are in the third line, for each sentence. The final process instance is presented in the last line.

When we find the pattern ", ... ," between two verbs in the text, we consider it as an explanation, addition or detail. This is the reason why we replace it with "..." in Sentence 1 before applying the actual rule.

Table H.7: Process instance discovered from the fragment.

Sentence 1
(S... 1.VBD ... , ... , (S 2.VBG ... .))
"1.VB/VBD/VBN ... 2.VBG " => 1||2
(1||2)
Sentence 2
(S... 3.VBD when (S ... 4.VBN .))
"1.VB/VBD/VBD ... when ... 2.VB/VBD/VBN" => 2→1
4→3
Sentence 3
(S ... 5.VBN ... (S 6.VBG ...) and 7.VBN ... .)
"1.VB/VBD/VBN ... 2.VBG" => 1||2
"1.VB/VBD/VBN/VBG ... and ... 2.VB/VBD/VBN/VBG" => 1→2
(5||6) → 7
(1.comprise removal_of_soil || 2.use toothless_bucket) → 4.reveal substrate → 3.cease machining → (5.record feature || 6.use leica_series_smartover_gps) → 7.survey feature

The parentheses and the tag S are ignored when applying the rules. We, nonetheless, keep them as they influence the grouping of activities. In Sentence 1 and Sentence 3 the activities which are extracted from dependent clauses will be grouped: (1||2) and (5||6). Moreover, the parentheses and the tag S allow us to decide the pair of verbs for which rules are checked. The default pairing takes two consecutive verbs as they appear in the sentence. In Sentence 3, the second rule is applied to 5.VBN and 7.VBN instead because of the reasoning over the clauses. For this illustration, we do not have any verb with multiple objects. In that case, the activity would have been fragmented in multiple parallel or mutual exclusive activities, which are also grouped.

H.4 Solution

In this section, we introduce *TextProcessMiner* (Figure H.1) the technique we created for mining processes from text. Although there are similar works discussed in more details in Section H.6, our approach is fully unsupervised and uses natural language processing techniques with a focus on the verb semantics. By understanding the verb semantics: we handle the passive/active sentences implicitly and we minimize the number of false positive activities enforced by the discovery of only transitive verbs.

We implemented our technique, following the technical action research principles, focusing on delivering fast results and thus enabling early experiments. The results of these experiments were further used for improving the tool output during multiple iterations. In the final experiments, the authors, as process modeling experts, were also involved to evaluate the results and provide feedback. The tool was developed in Python using different natural language processing libraries: NLTK [132], Stanford [121] and Enchant[7].

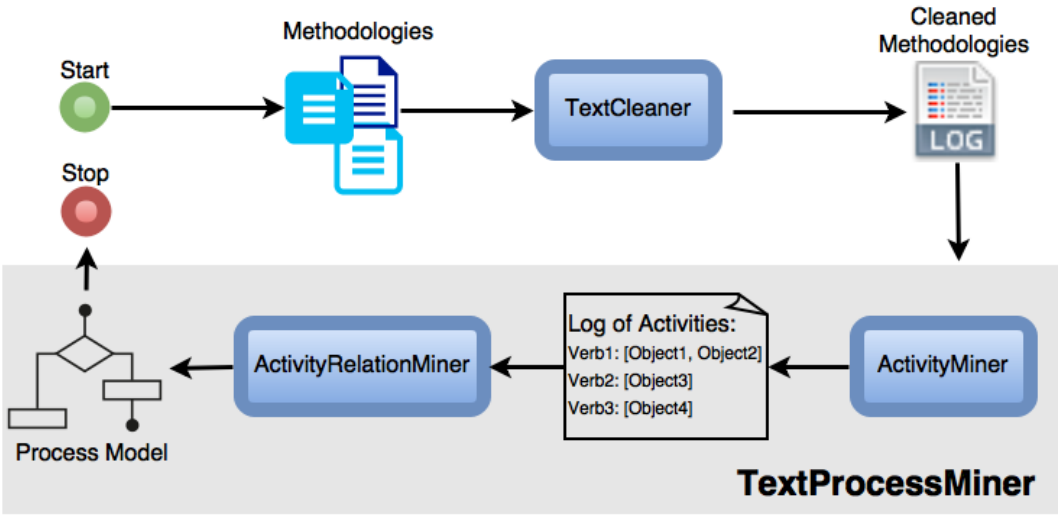


Figure H.1: Proposed solution for automatically discovering process models from text: an overview.

H.4.1 *TextCleaner*

The input of *TextCleaner* is the report methodology. Initially, the processing of the text consisted in the section extraction and sentence segmentation. However, after multiple experiments with different reports and the analysis of the results we decided to introduce cleaning actions to improve the output. These actions are summarized further:

- Replace "-" with "_";
- Add a space before and after each punctuation sign;
- Remove ' from the possessive nouns;
- Remove the comments introduced by parentheses if no verb was found inside them.

We also removed the sentences that contained negations because we assumed the activities were introduced by positive structures: they represent something that was done. After analyzing our corpus, we noticed negations can be linked to a noun: "no" + noun, "not" + noun, "none of" + noun, or can appear in relationship to a verb: "not" + verb, "-n't" + verb, "cannot" + verb. We also

[7] https://pypi.org/project/pyenchant/

identified several exceptions, when a negation keyword might appear in an expression that is not a negation (e.g. "not only ... but also"). In this situation, we do not discard the sentence.

The output of *TextCleaner* is a text file containing one sentence per line. This text file is further fed into *TextProcessMiner* as input and used by its sub-components: *ActivityMiner* and *ActivityRelationshipMiner*.

H.4.2 *TextProcessMiner*

TextProcessMiner is a process mining technique because it discovers process instance models automatically. However, it distinguishes from other process mining techniques by having as input textual process traces and not system-captured logs. Discovering the process in this situation becomes a matter of text mining.

The partial output of *TextProcessMiner* is a log text file with all the discovered activities. The final output is a text file with the process instance model (see Figure H.1).

ActivityMiner. The goal of *ActivityMiner* is to mine activities. We formulate the following requirements of *ActivityMiner*:

- The algorithm must identify the transitive verbs, which are most likely linked to activities;
- For each discovered transitive verb, the algorithm must identify its object(s).

First, *ActivityMiner* produces the treebank for each sentence using the Stanford and NLTK taggers, and the Stanford parser. We opted for the Stanford parser for its very high accuracy [121] and its online version used to test our results. Though the Stanford parser comes as a Java library, NLTK contains an interface[8], which allowed us to continue working in Python.

The Stanford parser is able to produce a treebank either directly from a raw text or from a tagged sentence—a list of tuples containing the sentence words and their part of speech tags. Initially, the parser was used in the first manner, creating treebanks from raw texts. However, later we decided to explore both possibilities as relevant words were wrongly tagged sometimes (nouns being marked as verbs, verbs being marked as adjectives etc.). We also noticed the speech tagging could be performed either by using the Stanford POS Tagger [204] or NLTK POS Tagger [132]. An experiment was set up to measure which possibility of the following yielded the most accurate and complete results when parsing five archaeological reports:

- Using the Stanford parser with raw text;
- Using the Stanford parser with the text priorly tagged with the Stanford tagger;
- Using the Stanford parser with the text priorly tagged with the NLTK tagger.

[8] `http://www.nltk.org/_modules/nltk/parse/stanford.html`

The best results were obtained with the second option. However, sometimes the NLTK tagger seemed to be able to correct some of the errors of the Stanford tagger. Therefore, we decided to use the result of the Stanford tagger as the base input of the parser but to artificially and automatically modify it by favoring the NLTK tags in the following situations:

- If the NLTK tagger tags the word as noun (NN, NNS) and if the Stanford tagger tags the word differently but not as verb in the "-ing" form (VBG) followed by a preposition (IN), and not as verb in the past form (VBD), and not as verb in the past participle form (VBN);

- If the NLTK tagger tags the word as verb in the base form (VB) and if the Stanford tagger tags the word differently but not as verb in any other form (VBG, VBD, VBN) and not as noun (NN, NNS).

Having found the method to obtain the most accurate treebank, we focus then on identifying the transitive verbs. Identifying the verbs in a sentence is a straightforward task: we look in the treebank for the VP or ADJP sub-trees to identify its children who start with VB, if any. However, for the automatic classification of verbs as transitive and intransitive extra-knowledge is needed. Two sources—VerbNet [119] and WordNet [139]—were used for compiling a dictionary of verbs having as key the verb in the infinitive form and as value a Boolean flag for the transitivity. Though the sources are different, they both share common information about the verbs: the frames. A frame illustrates how a verb could be used in a simple sentence [119]. The transitivity in VerbNet is considered true if the frame "NP V NP" is found among the verb's frames and if the first NP has the semantic role of Agent [119]. An agent is an active, intentional entity that carries out the activity introduced by the verb [119]. With the later condition some of the transitive verbs, which are rather states than actions, were excluded. Likewise, we parsed the WordNet [139] corpus and we added more verbs to the dictionary. In this case, the transitivity was judged depending on whether the frame "Somebody verb Something" was among the verb's frames. It can be noticed that "Somebody" is equivalent to a NP element with the semantic role of Agent. Two verbs were artificially changed to being intransitive: "be" and "have", because they do not represent activities. Moreover, during the experiments, when transitive verbs that were not in the dictionary were discovered, we added them manually to our file storing the verbs dictionary.

The verbs can be in active or passive forms. Consequently, the object of a verb can appear as object or subject. The form of the verb is not checked. Instead, we first check if the verb has an object. In the case it does not, we consider the subject being the verb's object. This is in general a valid assumption as we work with transitive verbs and passive voice is often used in reporting. For finding the object after the verb, we search for a NP element being the first right sibling of the VB sub-tree. For finding the object before the verb we look for the first left sibling of the VP or ADJP sub-tree or of any of their ancestors. A verb can have multiple objects in an enumeration or an object composed of multiple nouns. The algorithm discovers both cases.

Algorithm 4 Function to mine activities from a sentence.

Require:
sent—the input sentence;
Verbs—the dictionary of verbs.
Ensure:
Activities—the list of tuples (*sent*, *Verbs*).

```
function MINE_ACTIVITIES(convTrees, level)
    Activities ← []
    tagged ← TAG_SENTENCE(sent)
    treebank ← PARSE_TAGGED(tagged)
    for each subtree ∈ subtrees(treebank) with node label VP or ADJP do
        verbs ← FIND_VERBS(subtree, verbs)
        if |verbs| = 0 then
            continue
        objects ← FIND_OBJECTS_AFTER_VERB(subtree)
        if |objects| = 0 then
            objects ← FIND_OBJECTS_BEFORE_VERB(subtree)
            if |objects| = 0 then
                continue
        activities ← FORM_ACTIVITIES(sent, verbs, objects)
        extend Activities with the newly discovered activities
    return Activity
```

The result of *ActivityMiner* is a list of tuples where the first element of the tuple is the verb and the second element of the tuple is the list of objects. Both verbs and nouns are lemmatized using the WordNet lemmatizer [132]. The pseudo-code for the main algorithm of mining the activities from a sentence is summarized in Algorithm 4.

ActivityRelationshipMiner. The goal of *ActivityRelationshipMiner* is to mine the relationships between the discovered activities. As mentioned before, there are three types of relationship: sequence, parallelism and mutual exclusion.

In the methodology section of the archaeological report the writer reports the events in the order they happened. Therefore, the default relationship between two consecutive activities is sequence. Additionally, we also consider that two sentences are sequential by default. However, if the order is different then there are clues in the text which announce the change: for example temporal structures as "last year", "in 2011" or prepositions / conjunctions as "before", "after", "and", "or", "in order to".

ActivityRelationshipMiner is a rule-based method. The knowledge base represents a set of rules defined after analyzing the corpus. Some rules are presented in Table H.8. The activities extracted by *ActivityMiner* for each sentence together with the sentence are fed to the algorithm. First, the sentence is transformed by keeping only the tags of the verbs found among activities, the

S tags and their corresponding parentheses and other key structures as conjunctions, punctuation, prepositions etc. All the other words are replaced with "...". We show two examples:

- The sentence "The excavations were structured to accommodate the requirements of the developer." becomes "... 1.VBN to 2.VB ... .";
- The sentence "Particular areas were targeted by the second machining, including sondages across ditches and enclosure interiors." becomes "... 1.VBN ... , 2.VBG ... and ... .".

The numbers are used to relate the transitive verbs to the discovered activities.

Table H.8: Rules to mine activities relationships.

Rule Input	**Rule Output**
... 1.VB/VBN/VBD ... , ... 2.VBG ...	1 → 2 (independent clause)
... 1.VB/VBN/VBD 2.VBG ...	1 || 2 (dependent clause)
... where VB/VBD/VBN ... 2.VB/VBD/VBD ...	1? → 2 (decision)
... 1.VB/VBN/VBD ... , 2.VB/VBN/VBD ... or 3.VB/VBN/VBD ...	1 x 2 x 3 (branches)
... in order to 1.VB ... , ... 2.VB/VBN/VBD ...	2 → 1 (sequence)

The rules are always applied for a pair of verbs. For extracting the pairs, we take into consideration the sentence clauses and the dependencies of the verbs to each other. After discovering the relationship for each pair of verbs, the process fragment for the complete sentence is composed. Finally, the process fragments obtained from all sentences are put together to obtain the process instance. Currently, we consider the process fragments are in sequence but, in the future, we want to order the sentences and the independent clauses based on time structures. There is already research in this direction, one example being the Stanford Temporal Tagger [33] or the work of Muller [141].

We also consider the situation of the verbs having multiple objects. The activity is transformed in multiple parallel activities, one for each object, if the objects are enumerated with "and". The activity is transformed in multiple mutual exclusive activities, one for each object, if the objects are enumerated with "or". Finally, the algorithm is summarized (Algorithm 5).

H.4.3 Discussion

Though the presented technique achieved very good results on the selected corpus and during the validation, there are parts that could be improved.

First, *TextCleaner* removes the whole sentence if one of the negation structures is found. A better selection should be made, as there are situations when a complex sentence with negation contained also activities. One solution would be to remove only those clauses of the sentence that contain the negation instead of removing the whole sentence. Then, the filtering of negations would take place after the treebanks are produced.

Algorithm 5 Function to mine the process instance fragment from a sentence.

Require:

sent—the input sentence;
Activities—the list of *Activities* of the sentence.

Ensure:

ProcessFragment—a string with the activities and their relationships represented with the symbols from Table H.6.

```
function MINE_PROCESS_FRAGMENT(sent, Activities)
    transformed ← TRANSFORM_SENTENCE(sent, Activities)
    pairs ← EXTRACT_VERB_PAIRS(transformed)
    relationships ← []
    for each pair = (v1, v2) ∈ pairs do
        r ← APPLY_RULES(pair, transformed)
        add r to relationships
    ProcessFragment ← BUILD_STRING(relationships, transformed)
    return ProcessFragment
```

Second, *ActivityMiner* is capable of identifying the activities with high precision and accuracy. We have though to make the distinction between an activity and an activity name. For example "carry_out removal_of_topsoil" is a valid activity while its name is not necessary in the most appropriate form; it should be "remove topsoil". The automatic renaming of activities is a possibility to be explored. Another option would be to provide a way for the users to manually change the names of activities.

We still obtain false positives and false negatives mainly because of several reasons:

- The taggers and parsers do not work 100% correctly. For instance, when there are numbers in the sentence multiple errors appear;
- The verbs followed by a preposition pose problems. We are not able to tell at the moment if the construction verb + preposition is a transitive idiom (e.g. "look into the database") or not (e.g. "upload to the site"). Currently, these structures are considered as activities, exception made when the parser marks the word after the verb as particle (e.g. "carry out").
- Some activities belong to another process than the described one. They cannot be identified automatically. Similar to the activity names, we want to allow the users to manually remove false positives.

Further, a noun phrase (NP) can contain determiners ("this", "that" etc.) or pronouns ("it" etc.) instead of nouns. Currently, the determiners are replaced with the last encountered noun and the pronouns are kept as they are. We want to improve this feature in order to better identify the object referred by a determiner or by a pronoun.

Third, the authors, considering their thorough experience in process modeling, decided the reasoning rules of *ActivityRelationshipMiner*. An initial validation was made during one case. Other reports should be added to the corpus in order to check the existing rules and to discover new ones. The ordering of the sentences should also be better handled (we automatically consider them as sequential in this work).

Finally, lower priority was given to matters related to the performance or usability of the solution, but this is included in our future works.

H.5 Validation and Preliminary Evaluation

As specified in the Methodology section, we designed an evaluation to analyze the proposed technique in a practical case. We used the methodology section of the Villa Magna archaeological project [62].

H.5.1 Validation and Evaluation Setup

The protocol is designed according to the main objective: discover process models from textual methodologies.

First, we wanted to validate the obtained logs—the list of activities (the partial output in Figure H.1). We targeted the correctness, completeness and soundness of the activities. The correctness allows us to know if all activities are well identified in the log. The completeness determines if the set of activities in the log capture all activities from the report methodology or from the real process. Finally, the soundness assesses the degree to which the log reflects the process activities described in the report.

Secondly, we wanted to validate the process instance models created from the log (the final output in Figure H.1). We used the final output to draw a process model manually, using the BPMN formalism [227]. Since we evaluated the obtained process model and not a modeling formalism, we consider we did not introduce any bias in this step.

Then, we focused on the preliminary evaluation of the technique: the correctness, comprehension and utility of the process instance model, from the point of view of the real authors of the texts. They are specialists in Archaeology and potential users of the presented approach. The comprehension allows us to know if these potential users are able to understand the activities presented in the process instance model and their relationships. We also wanted to evaluate the whole approach in terms of its utility: does it allow the specialists to share knowledge with colleagues and make decisions based on the created methodological model? Nielsen [145] considers that utility is "synonymous with relevance or efficacy". In this particular case, we wanted to evaluate if the model allows the specialists to achieve a better understanding of the followed archaeological methodology.

According to other works in textual and discourse analysis, a sound evaluation should involve the participation of the authors of the texts themselves in order to avoid erroneous interpretations or bias [96]. Thus, the subject of this evaluation was the author of the report, being also one of the archaeologists responsible for the excavation works.

We designed a protocol consisting of six steps:

1. Fill in the first part of the first questionnaire with personal background information;
2. Read the report in order to recapitulate the archaeological project and the specific text;
3. Access the log generated from the report;
4. Fill in the second part of the first questionnaire regarding the log validation;
5. Access the process instance model drawn according to the generated model and analyze it;
6. Fill in the second questionnaire, validating and evaluating the process instance model.

The questionnaires were available online to provide easy access to the author of the Villa Magna report. The first part of the first questionnaire requires personal information about the participant: the name and current affiliation, his/her professional background and the number of years of archaeological experience. The second part of the first questionnaire comprises questions regarding the correctness, completeness and soundness of the activity log: the activities that are correctly/incorrectly identified or named (true/false positives), the activities that are missing from the log but were described in the report methodology; the extent to which the log reflects the process described in the report. The second questionnaire contains questions concerning the correctness, comprehension and utility of the process instance model, its strengths, weaknesses and possible improvements. All questions are single-answer accompanied by a textbox that enables the author to freely express his/her opinions, and, for us, to extract qualitative information.

H.5.2 Results

Using *TextProcessMining* on the methodology text of the Archaeology case, we obtained 34 activities. A fragment of the log is presented in Figure H.2. Out of these, the author reported in the first questionnaire:

- 4 false positives: "9.plan unit", "10.print transparent polyester sheet", "11.use millimetre grid" and "24.find deposit";
- 3 activities with wrong names: "11.model millimetre grid" instead of "model sheets", "13.allow recording of sequence" instead "allow recording of overlays" and "14.keep find" instead of "keep finds";

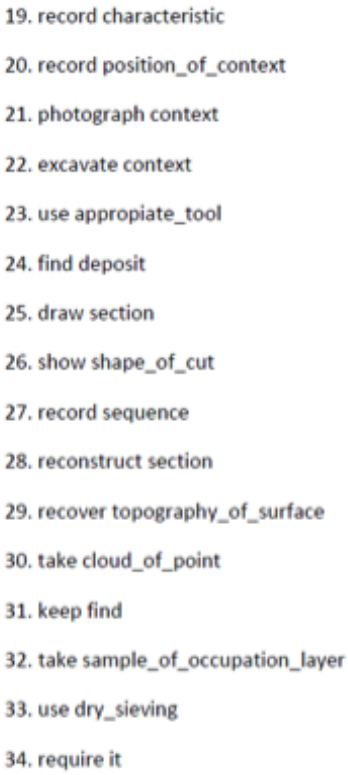


Figure H.2: Fragment of the discovered log.

- 1 activity missing: "wet sieving".

The achieved precision is very satisfactory for the validation case with 88% correctly discovered activities. Regarding the missing activity, though the statement of the author is totally valid, the methodology text does not contain any sentences regarding it. The author assumes previous archaeological knowledge to reason about the convenience of choosing a wet or a dry sieving in function of specific criteria. Consequently, we are able to extract activities explicitly referred in the text, but we are not able to discover implicit activities.

Further, the naming of the activities could be improved first by improving the object identification and second by not lemmatizing the objects. In conclusion, the correctness reflected by the precision of the log is very satisfactory. The completeness is also very good, just one activity being reported as missing from the real process. In the final solution, the user will be able to review the log, exclude the false positives or add missing activities. Regarding the soundness, the author finds the real process activities "adequately" reflected in the log.

Concerning the relationships between activities, the archaeologist agrees with the discovered process instance model (a fragment is presented in Figure H.3). However, as she states: "what does not emerge is the situation of conditionality". The mutual exclusion is identified in multiple cases but not always (correctly). For example, "25.draw section" happens only in some situations

in order to execute the activity "26.show shape of cut" and "27.record sequence" happens only under other situations when "25.draw section" is not required. An appropriate re-modeling could be summarized thus: if certain situations then "25.draw section" → "26.show shape of cut" else "27.record sequence". Similarly, in the process fragment below (Figure H.3), the activity "28.reconstruct section" does not happen all the times but only under certain circumstances. Another error concerns the activity "33.use dry sieving", which should come after the activity "22.excavate context". However, the archaeologist acknowledged it was a consequence of the order the activities were reported in text.

Regarding the overall comprehensiveness of the process instance model, the archaeologist is positive about its strengths and states that it "flows like the archaeological site investigation process". However, although the archaeologist reports that the model reflects adequately the report methodology, she states she would not use the model for knowledge sharing with other colleagues or for process guidance. She would only use the model for teaching or for disseminating knowledge to non-specialists. These results about the utility of the process instance model might be grounded in the traditional way of working in Humanities but a more thorough evaluation is required from our side too.

H.6 Related Works

Automatic analysis of text for discovering models has been researched for decades. Extensive work done in this field is related to requirements engineering where textual specifications are often created before the design and construction of any information system. Years ago, Rolland [174] identified four types of strategies to address the relationship between text and conceptual modeling: support the generation of models from text, support the model paraphrasing, help with the general understanding of texts and improve the text quality. Our work is related to the first strategy as we aim at discovering process models from textual methodologies. In a later work, Rolland [172] classifies the current techniques to support the discovery of conceptual models form text considering the following model-related characteristics: static or dynamic, rule-based or ontology-based. Regarding the static aspects of the conceptual models, there are tools to discover and generate object models or class models from textual requirement specifications [89, 111]. Regarding the dynamic aspects, Rolland highlighted works to extract use cases and scenarios from texts [122, 173, 179]. Finally, regarding the rule-based or ontology-based aspects, there are approaches looking for discovering business process models from business rules [92, 167].

In addition to the later mentioned works, we have found other approaches to discover processes from: clinical documents [201], emails [193], stories extracted from collaboration [80] and business process documentation [69, 76]. Compared to our technique, [201] is domain dependent and fully-supervised, relying on prior annotation of the corpus. [80, 193] use a technique similar to ours but the key elements of an activity in their case are the actors and the verbs. The authors use

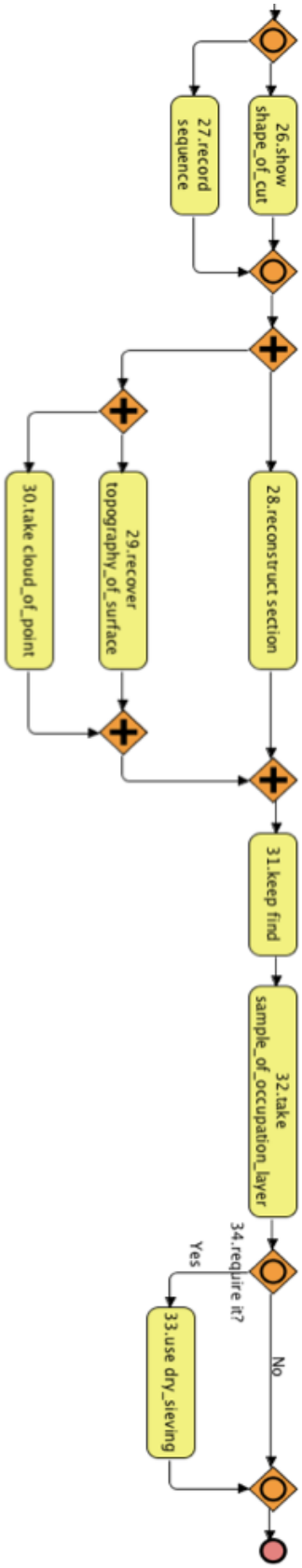


Figure H.3: Fragment of the discovered process instance model.

a template for activity extraction that does not appear to handle passive-voice sentences and according to the exhibits in the articles there is no fine-grained identification of the objects. [76] first defines a list of templates that are used for identifying process-related sentences; second, they extract activities. No technical details are given and no indication of discovering a full process is found but rather process fragments. Finally, the solution proposed by [68, 69] is probably the closest to ours. The difference relies in the understanding of the verb semantics: we handle the passive/active sentences implicitly; we avoid false positives by enforcing the discovery of only transitive verbs while the authors use a manually defined list of "weak verbs"; we can extract activities from treebanks directly while the authors rely on the grammatical relationships exposed as Stanford Dependencies [136]. Nonetheless, [68, 69] provide a mechanism for associating the pronouns to the corresponding concepts.

In other fields, complementary to Information Systems, there are works to identify statistical models from textual sources, with applications in biomedicine [198]. The resulting models are essentially mathematical. Consequently, they lack the semantic richness achieved through conceptual modeling approaches, as previously explained.

We have identified also attempts to apply natural language processing techniques for text analysis in Humanities and Social Sciences disciplines, including Archaeology [24]. However, these approaches mainly rely on previous annotation of the texts by experts or require the use of auxiliary domain-dependent ontologies.

H.7 Conclusion and Future Works

In this paper, we proposed an automatic technique to extract activity logs and mine process instance models from textual methodologies. We have used natural language processing techniques focusing on the verb semantics for activities mining and a rule-based system for activity relationships mining. The validation and preliminary evaluation in an Archaeology case show that:

- The majority of the discovered activities are correct;
- The mined process instance model is satisfactory in comparison to the real enacted process;
- The archaeologist is positive concerning the comprehension of the mined model but she is reserved regarding its usability for knowledge sharing or process guidance.

Although the results are promising, the technique can be improved to overcome its limitations. We highlight the following improvement areas: the identification of negations; the activity discovery with idiomatic verbs; the activity naming; the ordering of sentences and independent clauses taking in consideration the time clues; the rule-based method especially the mutual exclusion; the performance and usability. Additionally, we want to enrich the model by discovering other

relationships such as iterations or implicit mutual exclusion and other information regarding an activity such as its actor.

We might also generate a XES log [1] instead of a log text file in order to enable the loading of the file in ProM [211] or Disco [84]. In this way, the users would be provided with an interactive application, which could support the process review, by deleting or renaming activities, and its representation using different formalisms. However, the main issue regarding the generation of the XES logs is the conservation of the activity relations captured from text. A possible solution to this matter would be the artificial injections of events and traces in the log.

As we mentioned in the beginning, this paper presents only the first step of our research aiming at mining a process instance model. However, in the second step we want to be able to mine multiple process instances from the same or different text and to aggregate those in one process model. This requires: (1) a method to identify if two process instances refer to the same process; (2) a method to identify if two activities are equivalent or related; (3) a method to aggregate two or more process instances of the same process. Regarding phase (3) there are already mature approaches dealing with process variability management and process model similarity, matching and merging in process model repositories [175] that might be reused and adapted.

Finally, an extensive validation and evaluation including other domains apart from Archaeology (Medicine, Business, Information Systems etc.) and other data sources (business reports, ISO standards, software documentation etc.) is necessary in order to establish the generalization and usefulness of the proposed technique. The challenge will be to handle the threats to validity considering the characteristics of various domains and the background of the various workers following processes.

BIBLIOGRAPHY

[1] W. V. D. AALST, *Process mining*, Springer, 2011.

[2] G. ADOMAVICIUS AND A. TUZHILIN, *Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*, Transactions on Knowledge and Data Engineering, 17 (2005), pp. 734–749.

[3] C. C. AGGARWAL, *Recommender Systems: The Textbook*, Springer Publishing Company, Incorporated, 1st ed., 2016.

[4] A. AHMED, C. H. TEO, S. V. N. VISHWANATHAN, AND A. SMOLA, *Fair and Balanced : Learning to Present News Stories*, in WSDM, 2012, pp. 333–342.

[5] I. AJZEN, *The theory of planned behavior*, Organizational Behavior and Human Decision Processes, 50 (1991), pp. 179 – 211.

[6] M. ALLAMANIS AND C. SUTTON, *Mining idioms from source code*, in Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, FSE 2014, New York, NY, USA, 2014, ACM, pp. 472–483.

[7] J. ALLEN AND M. CORE, *Draft of damsl: Dialog act markup in several layers*, 1997.

[8] S. ANANTHAKRISHNAN, P. GHOSH, AND S. NARAYANAN, *Automatic classification of question turns in spontaneous speech using lexical and prosodic evidence*, in 2008 IEEE ICASSP, March 2008, pp. 5005–5008.

[9] G. ANSCOMBE, *Intention*, Harvard University Press, 1957.

[10] J. ARGUELLO AND K. SHAFFER, *Predicting speech acts in mooc forum posts.*, in Ninth International AAAI Conference on Web and Social Media, 2015, pp. 2–11.

[11] S. ASUR AND B. A. HUBERMAN, *Predicting the future with social media*, in Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '10, Washington, DC, USA, 2010, IEEE Computer Society, pp. 492–499.

[12] J. L. AUSTIN AND J. URMSON, *How to Do Things with Words. The William James Lectures Delivered at Harvard University in 1955.*, Clarendon Press, 1962.

[13] B. BAYAT, C. KRAUSS, A. MERCERON, AND S. ARBANOWSKI, *Supervised speech act classification of messages in german online discussions*, in Proceedings of the 29th International FLAIRS Conference, AAAI, 2016.

[14] S. BHATIA, P. BIYANI, AND P. MITRA, *Classifying user messages for managing web forum data*, in Proceedings of the 15th International Workshop on the Web and Databases, ACM, 2012, pp. 13,Äì–18.

[15] F. BICKENBACH AND E. BODE, *Evaluating the Markov Property in Studies of Economic Convergence*, International Regional Science Review, 26 (2003), pp. 363–392.

[16] D. BILLSUS AND M. J. PAZZANI, *Adaptive News Access*, in The Adaptive Web, P. Brusilovsky, A. Kobsa, and W. Nejdl, eds., Springer, 2007, ch. 18, pp. 550–570.

[17] C. M. BISHOP, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, 2006.

[18] R. BLUMBERG AND S. ATRE, *The problem with unstructured data*, Dm Review, 13 (2003), p. 62.

[19] A. BOOTH, D. PAPAIOANNOU, AND A. SUTTON, *Systematic Approaches to a Successful Literature Review*, SAGE Publications, 2012.

[20] W. BOSMA AND E. ANDRÉ, *Exploiting emotions to disambiguate dialogue acts*, in Proceedings of the 9th International Conference on Intelligent User Interfaces, IUI '04, New York, NY, USA, 2004, ACM, pp. 85–92.

[21] M. BRATMAN, *Intention, plans, and practical reason*, Harvard Univ. Press, 1987.

[22] S. BRINKKEMPER, *Method engineering: engineering of information systems development methods and tools*, Information and Software Technology, 38 (1996), pp. 275–280.

[23] J. BROWN, A. J. BRODERICK, AND N. LEE, *Word of mouth communication within online communities: Conceptualizing the online social network*, Journal of Interactive Marketing, 21 (2007), pp. 2 – 20.

[24] M. P. D. BUONO, M. MONTELEONE, P. RONZINO, V. VASSALLO, AND S. HERMON, *Decision making support systems for the archaeological domain: A natural language processing proposal*, in 2013 Digital Heritage International Congress (DigitalHeritage), vol. 2, Oct 2013, pp. 397–400.

[25] J. K. CALVERT, *An ecological view of internet health information seeking behavior predictors: Findings from the chain study*, Open AIDS J., 7 (2013), pp. 42–46.

[26] M. CAPELLE, A. HOGENBOOM, A. HOGENBOOM, AND F. FRASINCAR, *Semantic News Recommendation Using WordNet and Bing Similarities Categories and Subject Descriptors*, in Symposium on Applied Computing, 2013, pp. 296–302.

[27] J. CARLETTA, *Assessing agreement on classification tasks: The kappa statistic*, Comput. Linguist., 22 (1996), pp. 249–254.

[28] V. R. CARVALHO AND W. W. COHEN, *On the collective classification of email "speech acts"*, in 28th ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, New York, NY, USA, 2005, ACM, pp. 345–352.

[29] ——, *Improving "email speech acts" analysis via n-gram selection*, in Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech, ACTS '09, Stroudsburg, PA, USA, 2006, Association for Computational Linguistics, pp. 35–41.

[30] M. CASTELLANOS, M. HSU, U. DAYAL, R. GHOSH, M. DEKHIL, C. CEJA, M. PUCHI, AND P. RUIZ, *Intention insider: Discovering people's intentions in the social channel*, in Proceedings of the 15th International Conference on Extending Database Technology, EDBT '12, New York, NY, USA, 2012, ACM, pp. 614–617.

[31] D. CENTOLA, *The spread of behavior in an online social network experiment*, Science, 329 (2010), pp. 1194–1197.

[32] V. CHANDOLA, A. BANERJEE, AND V. KUMAR, *Anomaly detection: A survey*, ACM Comput. Surv., 41 (2009), pp. 15:1–15:58.

[33] A. X. CHANG AND C. MANNING, *Sutime: A library for recognizing and normalizing time expressions*, in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), European Language Resources Association (ELRA), 2012.

[34] R. CHARON, *Narrative medicine: A model for empathy, reflection, profession, and trust*, JAMA, 286 (2001), pp. 1897–1902.

[35] ——, *What to do with stories: The sciences of narrative medicine*, Canadian Family Physician J., 53 (2007), pp. 1265–1267.

[36] S. CHATTERJEE AND A. PRICE, *Healthy living with persuasive technologies: Framework, issues, and challenges*, JAMIA, 16 (2009), pp. 171–178.

[37] N. V. CHAWLA, K. W. BOWYER, L. O. HALL, AND W. P. KEGELMEYER, *Smote: Synthetic minority over-sampling technique*, J. Artif. Int. Res., 16 (2002), pp. 321–357.

[38] H. CHEN, W. CHUNG, J. J. XU, G. WANG, Y. QIN, AND M. CHAU, *Crime data mining: a general framework and some examples*, Computer, 37 (2004), pp. 50–56.

[39] R. J. W. CLINE, *Consumer health information seeking on the internet: the state of the art*, Health Education Research, 16 (2001), pp. 671–692.

[40] W. W. COHEN, V. R. CARVALHO, AND T. M. MITCHELL, *Learning to classify email into "speech acts"*, in Proceedings of EMNLP 2004, D. Lin and D. Wu, eds., Barcelona, Spain, July 2004, Association for Computational Linguistics, pp. 309–316.

[41] D. COMPAGNO, E. V. EPURE, R. DENECKERE-LEBAS, AND C. SALINESI, *Exploring digital conversation corpora with process mining*, Corpus Pragmatics, (2018).

[42] H. M. COOPER, *Organizing knowledge syntheses: A taxonomy of literature reviews*, Knowledge in Society, 1 (1988), p. 104.

[43] A. DAS, M. DATAR, A. GARG, AND S. RAJARAM, *Google News Personalization: Scalable Online*, in WWW, 2007, pp. 271–280.

[44] G. DE FRANCISCI MORALES, A. GIONIS, AND C. LUCCHESE, *From Chatter to Headlines: Harnessing the Real-time Web for Personalized News Recommendation*, in Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, New York, NY, USA, 2012, ACM, pp. 153–162.

[45] M. DE GEMMIS, P. LOPS, C. MUSTO, F. NARDUCCI, AND G. SEMERARO, *Semantics-Aware Content-Based Recommender Systems*, Springer US, Boston, MA, 2015, pp. 119–159.

[46] R. DENECKÈRE, C. HUG, G. KHODABANDELOU, AND C. SALINESI, *Intentional process mining: discovering and modeling the goals behind processes using supervised learning*, International Journal of Information System Modeling and Design, 5 (2014), pp. 22–47.

[47] S. DONOHO, *Early detection of insider trading in option markets*, in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, New York, NY, USA, 2004, ACM, pp. 420–429.

[48] D. DOYCHEV, A. LAWLOR, R. RAFTER, AND B. SMYTH, *An Analysis of Recommender Algorithms for Online News*, in CLEF (Working Notes), 2014, pp. 825–836.

[49] U. ECO, *The Role of the Reader: Explorations in the Semiotics of Texts*, Indiana University Press, 1979.

[50] E. V. EPURE, D. COMPAGNO, C. SALINESI, R. DENECKERE, M. BAJEC, AND S. ZITNIK, *Process models of interrelated speech intentions from online health-related conversations*, Artificial Intelligence In Medicine, (2018).

[51] E. V. EPURE, R. DENECKERE, AND C. SALINESI, *Analyzing perceived intentions of public health-related communication on twitter*, in Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings, A. ten Teije, C. Popow, J. H. Holmes, and L. Sacchi, eds., Cham, 2017, Springer International Publishing, pp. 182–192.

[52] E. V. EPURE, R. DENECKERE, C. SALINESI, B. KILLE, AND J. INGVALDSEN, *Devising news recommendation strategies with process mining support*, in Atelier interdisciplinaire sur les systèmes de recommandation/Interdisciplinary Workshop on Recommender Systems, 2017.

[53] E. V. EPURE, C. HUG, R. DENECKÉRE, AND S. BRINKKEMPER, *What shall i do next?*, in Advanced Information Systems Engineering, M. Jarke, J. Mylopoulos, C. Quix, C. Rolland, Y. Manolopoulos, H. Mouratidis, and J. Horkoff, eds., Cham, 2014, Springer International Publishing, pp. 473–487.

[54] E. V. EPURE, J. E. INGVALDSEN, R. DENECKERE, AND C. SALINESI, *Process Mining for Recommender Strategies Support in News Media*, in Proceedings of the 10th International Conference on Research Challenges in Information Science (RCIS), IEEE, June 2016, pp. 1–12.

[55] ——, *Process Mining for Recommender Strategies Support in News Media*, in Proceedings of the 10th International Conference on Research Challenges in Information Science (RCIS), IEEE, June 2016.

[56] E. V. EPURE, B. KILLE, J. E. INGVALDSEN, R. DENECKERE, C. SALINESI, AND S. ALBAYRAK, *Recommending personalized news in short user sessions*, in Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17, New York, NY, USA, 2017, ACM, pp. 121–129.

[57] E. V. EPURE, P. MARTIN-RODILLA, C. HUG, R. DENECKERE, AND C. SALINESI, *Automatic process model discovery from textual methodologies*, in 2015 IEEE RCIS, May 2015, pp. 19–30.

[58] E. V. EPURE, S. ZITNIK, D. COMPAGNO, R. DENECKERE, AND C. SALINESI, *Automatic analysis of online conversations as processes*, in JournéeS Analyse de Données Textuelles en Conjonction avec EDA 2017, 2017.

[59] C. ESIYOK, B. KILLE, B.-J. JAIN, F. HOPFGARTNER, AND S. ALBAYRAK, *Users' Reading Habits in Online News Portals*, in Proceedings of the 5th Information Interaction in Context Symposium, IIiX '14, New York, NY, USA, 2014, ACM, pp. 263–266.

[60] R. FELDMAN AND J. SANGER, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, New York, NY, USA, 2006.

[61] D. FENG, E. SHAW, J. KIM, AND E. HOVY, *Learning to detect conversation focus of threaded discussions*, in Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06, Stroudsburg, PA, USA, 2006, Association for Computational Linguistics, pp. 208–215.

[62] E. FENTRESS, *Archaeological fieldwork reports: Excavations at villa magna, 2009*, Papers of the British School at Rome, 78 (2010), pp. 324,Äì–325.

[63] O. FERSCHKE, *The quality of content in open online collaboration platforms: Approaches to NLP-supported information quality management in Wikipedia*, PhD thesis, Technische Universität, 2014.

[64] O. FERSCHKE, I. GUREVYCH, AND Y. CHEBOTAR, *Behind the article: Recognizing dialog acts in wikipedia talk pages*, in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, Stroudsburg, PA, USA, 2012, Association for Computational Linguistics, pp. 777–786.

[65] W. R. FISHER, *The narrative paradigm: In the beginning*, Journal of Communication, 35 (1985), pp. 74–89.

[66] J. L. FLEISS, J. COHEN, AND B. S. EVERITT, *Large sample standard errors of kappa and weighted kappa.*, Psychological Bulletin, 72 (1969), pp. 323–327.

[67] A. FRANKFORT-NACHMIAS, CHAVALEON-GUERRERO, *Social Statistics for a Diverse Society*, Pine Forge Press, 2006.

[68] F. FRIEDRICH, *Automated Generation of Business Process Models from Natural Language Input*, PhD thesis, School of Business and Economics, Humboldt-Universitat zu Berlin, 2010.

[69] F. FRIEDRICH, J. MENDLING, AND F. PUHLMANN, *Process model generation from natural language text*, in Proceedings of the 23rd International Conference on Advanced Information Systems Engineering, CAiSE'11, Berlin, Heidelberg, 2011, Springer-Verlag, pp. 482–496.

[70] A. G. O'KEEFFE, G. AMBLER, AND J. BARBER, *Sample size calculations based on a difference in medians for positively skewed outcomes in health care studies*, 17 (2017).

[71] Q. GAO, F. ABEL, G.-J. HOUBEN, AND K. TAO, *Interweaving Trend and User Modeling for Personalized News Recommendation*, in IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, IEEE, 2011, pp. 100–103.

[72] P. GARCIA, *Meaning in academic contexts: A corpus-based study of pragmatic utterances*, PhD thesis, University of Northern Arizona, 01 2004.

[73] F. GARCIN, K. ZHOU, B. FALTINGS, AND V. SCHICKEL, *Personalized News Recommendation Based on Collaborative Filtering*, in Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, WI-IAT '12, Washington, DC, USA, 2012, IEEE Computer Society, pp. 437–441.

[74] R. GASPARINI, D. PANATTO, P. L. LAI, AND D. AMICIZIA, *The "urban myth" of the association between neurological disorders and vaccinations*, Journal of Preventive Medicine and Hygiene, 56 (2015), pp. E1–E8.

[75] A. GELMAN, *Scaling regression inputs by dividing by two standard deviations*, Statistics in Medicine, 27 (2008), pp. 2865–2873.

[76] A. GHOSE, G. KOLIADIS, AND A. CHUENG, *Process discovery from model and text artefacts*, in 2007 IEEE Congress on Services (Services 2007), July 2007, pp. 167–174.

[77] A. K. GODEA, C. CARAGEA, F. A. BULGAROV, AND S. RAMISETTY-MIKLER, *An analysis of twitter data on e-cigarette sentiments and promotion*, in AIME 2015, J. H. Holmes, R. Bellazzi, L. Sacchi, and N. Peek, eds., Cham, 2015, Springer, pp. 205–215.

[78] M. K. GOLD, *Debates in the Digital Humanities*, University of Minnesota Press, ned - new edition ed., 2012.

[79] J. GOLDSTEIN AND R. E. SABIN, *Using speech acts to categorize email and identify email genres*, in Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 03, HICSS '06, Washington, DC, USA, 2006, IEEE Computer Society, pp. 50.2–.

[80] J. C. D. A. R. GONCALVES, F. M. SANTORO, AND F. A. BAIAO, *A case study on designing business processes based on collaborative and mining approaches*, in The 2010 14th International Conference on Computer Supported Cooperative Work in Design, April 2010, pp. 611–616.

[81] C. GONZALEZ-PEREZ, P. MARTIN-RODILLA, C. PARCERO-OUBINA, P. FABREGA-ALVAREZ, AND A. GUIMIL-FARINA, *Extending an abstract reference model for transdisciplinary work in cultural heritage*, in Metadata and Semantics Research, Springer, 2012, pp. 190–201.

[82] C. GONZALEZ-PEREZ AND C. PARCERO-OUBINA, *A conceptual model for cultural heritage definition and motivation*, in Revive the Past: Proceedings of the 39th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Amsterdam University Press, 2015, pp. 234–244.

[83] J. A. GULLA, B. YU, Ö. ÖZGÖBEK, AND N. SHABIB, *Third International Workshop on News Recommendation and Analytics (INRA 2015)*, in Proceedings of the 9th ACM Conference on Recommender Systems, ACM, 2015, pp. 345–346.

[84] C. W. GÜNTHER AND A. ROZINAT, *Disco: Discover your processes*, in BPM, 2012.

[85] C. W. GÜNTHER AND W. M. P. VAN DER AALST, *Fuzzy mining: Adaptive process simplification based on multi-perspective metrics*, in 5th Int. Conference on Business Process Management, Berlin, Heidelberg, 2007, Springer-Verlag, pp. 328–343.

[86] I. GUYON, J. WESTON, S. BARNHILL, AND V. VAPNIK, *Gene selection for cancer classification using support vector machines*, Mach. Learn., 46 (2002), pp. 389–422.

[87] M. HACK, *Petri net language*, tech. rep., Cambridge, MA, USA, 1976.

[88] C. L. HANSON, J. WEST, R. THACKERAY, M. D. BARNES, AND J. DOWNEY, *Understanding and predicting social media use among community health center patients: A cross-sectional survey.*, Journal of Medical Internet Research, 16 (2014), p. e270.

[89] H. M. HARMAIN AND R. GAIZAUSKAS, *Cm-builder: A natural language-based case tool for object-oriented analysis*, Automated Software Engg., 10 (2003), pp. 157–181.

[90] A. HASHAVIT, R. LEVIN, I. GUY, AND G. KUTIEL, *Effective trend detection within a dynamic search context*, in Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM, 2016, pp. 817–820.

[91] L. HEMPHILL, J. OTTERBACHER, AND M. SHAPIRO, *What's congress doing on twitter?*, in 2013 Conference on Computer Supported Cooperative Work, CSCW '13, New York, NY, USA, 2013, ACM, pp. 877–886.

[92] H. HERBST, *Business Rule-Oriented Conceptual Modeling*, Physica-Verlag, 1997.

[93] A. R. HEVNER, S. T. MARCH, J. PARK, AND S. RAM, *Design science in information systems research*, MIS Q., 28 (2004), pp. 75–105.

[94] B. HIDASI, A. KARATZOGLOU, L. BALTRUNAS, AND D. TIKK, *Session-based recommendations with recurrent neural networks*, in Proceedings of the International Conference on Learning Representations, 2015.

[95] T. HILLESUND, *Digital text cycles: From medieval manuscripts to modern markup*, Journal of Digital Information, 6 (2005).

[96] J. HOBBS, *On the coherence and structure of discourse*, tech. rep., 1985.

[97] B. HOLLERIT, M. KRÖLL, AND M. STROHMAIER, *Towards linking buyers and sellers: Detecting commercial intent on twitter*, in Proceedings of the 22Nd International Conference on World Wide Web, WWW '13 Companion, New York, NY, USA, 2013, ACM, pp. 629–632.

[98] W. HONG, L. LI, AND T. LI, *Product Recommendation with Temporal Dynamics*, Expert systems with applications, 39 (2012), pp. 12398–12406.

[99] J. HOUGHTON, M. SIEGEL, AND D. GOLDSMITH, *Modeling the influence of narratives on collective behavior case study*, in Int. System Dynamics Conf., 2013.

[100] J. HU, R. J. PASSONNEAU, AND O. RAMBOW, *Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units*, in Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '09, Stroudsburg, PA, USA, 2009, Association for Computational Linguistics, pp. 357–366.

[101] C. HUG, C. SALINESI, R. DENECKERE, AND S. LAMASSE, *Process modeling for humanities: tracing and analyzing scientific processes*, in Revive the Past: Proceedings of the 39th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Amsterdam University Press, 2011, pp. 245–255.

[102] W. IJSSELSTEIJN, Y. DE KORT, C. MIDDEN, B. EGGEN, AND E. VAN DEN HOVEN, *Persuasive technology for human well-being: Setting the scene*, in 1st Conf. on Persuasive Technology for Human Well-being, Heidelberg, 2006, Springer-Verlag, pp. 1–5.

[103] E. IVANOVIC, *Dialogue act tagging for instant messaging chat sessions*, in Proceedings of the ACL Student Research Workshop, ACLstudent '05, Stroudsburg, PA, USA, 2005, Association for Computational Linguistics, pp. 79–84.

[104] M. JAMALI AND H. ABOLHASSANI, *Different aspects of social network analysis*, in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI '06, Washington, DC, USA, 2006, IEEE Computer Society, pp. 66–72.

[105] A. JAVARI AND M. JALILI, *A Probabilistic Model to Resolve Diversity–Accuracy Challenge of Recommendation Systems*, Knowledge and Information Systems, 44 (2015), pp. 609–627.

[106] M. JEONG, C.-Y. LIN, AND G. G. LEE, *Semi-supervised speech act recognition in emails and forums*, in 2009 Conf. on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2009, ACL, pp. 1250–1259.

[107] Y. JO, M. YODER, H. JANG, AND C. ROSE, *Modeling dialogue acts with content word filtering and speaker preferences*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 2179–2189.

[108] N. JONNALAGEDDA, S. GAUCH, K. LABILLE, AND S. ALFARHOOD, *Incorporating Popularity in a Personalized News Recommender System*, PeerJ Computer Science, 2 (2016), p. e63.

[109] S. JOTY, G. CARENINI, AND C.-Y. LIN, *Unsupervised modeling of dialog acts in asynchronous conversations*, in 22nd Int. Joint Conf. on Artificial Intelligence, IJCAI'11, AAAI Press, 2011, pp. 1807–1813.

[110] D. JURAFSKY, R. BATES, N. COCCARO, R. MARTIN, M. METEER, K. RIES, E. SHRIBERG, A. STOLCKE, P. TAYLOR, AND C. VAN ESS-DYKEMA, *Automatic detection of discourse structure for speech recognition and understanding*, in 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, Dec 1997, pp. 88–95.

[111] N. JURISTO, J. L. MORANT, AND A. M. MORENO, *A formal approach for generating oo specifications from natural language*, J. Syst. Softw., 48 (1999), pp. 139–153.

[112] S. KEIZER, R. OP DEN AKKER, AND A. NIJHOLT, *Dialogue act recognition with bayesian networks for dutch dialogues*, in Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue - Volume 2, SIGDIAL '02, Stroudsburg, PA, USA, 2002, Association for Computational Linguistics, pp. 88–94.

[113] J. KEMÉNY AND J. SNELL, *Finite Markov Chains*, University Series in Undergraduate Mathematics, Van Nostrand, 1960.

[114] G. KHODABANDELOU, C. HUG, R. DENECKÈRE, AND C. SALINESI, *Unsupervised discovery of intentional process models from event logs*, in Proceedings of the 11th Working Conference on Mining Software Repositories, MSR 2014, New York, NY, USA, 2014, ACM, pp. 282–291.

[115] ——, *Unsupervised discovery of intentional process models from event logs*, in Proceedings of the 11th Working Conference on Mining Software Repositories, MSR 2014, New York, NY, USA, 2014, ACM, pp. 282–291.

[116] Y. KIAM TAN, X. XU, AND Y. LIU, *Improved Recurrent Neural Networks for Session-based Recommendations*, arXiv preprint arXiv:1606.08117, (2016).

[117] R. KIBBLE, *Speech acts, commitment and multi-agent communication*, Computational and Mathematical Organization Theory, 12 (2006), pp. 127–145.

[118] S. N. KIM, L. CAVEDON, AND T. BALDWIN, *Classifying dialogue acts in one-on-one live chats*, in 2010 Conf. on Empirical Methods in Natural Language Processing, EMNLP '10, Stroudsburg, PA, USA, 2010, ACL, pp. 862–871.

[119] K. KIPPER, B. SNYDER, AND M. PALMER, *Using prepositions to extend a verb lexicon*, in Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, CLS '04, Stroudsburg, PA, USA, 2004, Association for Computational Linguistics, pp. 23–29.

[120] K. KIRA AND L. A. RENDELL, *The feature selection problem: Traditional methods and a new algorithm*, in Proceedings of the 10th National Conference on Artificial Intelligence, AAAI'92, AAAI Press, 1992, pp. 129–134.

[121] D. KLEIN AND C. D. MANNING, *Accurate unlexicalized parsing*, in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, Stroudsburg, PA, USA, 2003, Association for Computational Linguistics, pp. 423–430.

[122] L. KOF, *Scenarios: Identifying missing objects and actions by means of computational linguistics*, in 15th IEEE International Requirements Engineering Conference (RE 2007), Oct 2007, pp. 121–130.

[123] P. KRAL AND C. CERISARA, *Dialogue act recognition approaches*, Computing and Informatics, 29 (2010), pp. 227–250.

[124] A. LAMPERT, R. DALE, AND C. PARIS, *Classifying speech acts using verbal response modes*, in Australasian Language Technology Workshop, 2006, p. 34.

[125] J. LI, A. RITTER, C. CARDIE, AND E. HOVY, *Major life event extraction from twitter based on congratulations/condolences speech acts*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, pp. 1997–2007.

[126] L. LI, W. CHU, J. LANGFORD, AND R. E. SCHAPIRE, *A Contextual-Bandit Approach to Personalized News Article Recommendation*, in Proceedings of the 19th International Conference on World Wide Web, WWW '10, New York, New York, USA, 2010, ACM Press.

[127] L. LI, L. ZHENG, F. YANG, AND T. LI, *Modeling and Broadening Temporal User Interest in Personalized News Recommendation*, Expert Systems with Applications, 41 (2014), pp. 3168–3177.

[128] T.-P. LIANG AND H.-J. LAI, *Discovering user interests from web browsing behavior: An application to internet news services*, in HICSS, IEEE Computer Society, 2002, p. 203.

[129] B. LIU, M. HU, AND J. CHENG, *Opinion observer: Analyzing and comparing opinions on the web*, in 14th Int. Conf. on WWW, NY, USA, 2005, ACM, pp. 342–351.

[130] J. LIU, P. DOLAN, AND E. R. PEDERSEN, *Personalized News Recommendation Based on Click Behavior*, in Proceedings of the International Conference on Intelligent User Interfaces, 2010, pp. 31–40.

[131] A. LOMMATZSCH, *Real-time news recommendation using context-aware ensembles*, in Proceedings of the 36th European Conference on IR Research, ECIR '14, Springer, 2014, pp. 51–62.

[132] E. LOPER AND S. BIRD, *Nltk: The natural language toolkit*, in ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Stroudsburg, PA, USA, 2002, ACL, pp. 63–70.

[133] W. MAALEJ AND H. NABIL, *Bug report, feature request, or simply praise? on automatically classifying app reviews*, in 2015 IEEE 23rd International Requirements Engineering Conference (RE), Aug 2015, pp. 116–125.

[134] C. D. MANNING AND H. SCHÜTZE, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, USA, 1999.

[135] M. P. MARCUS, M. A. MARCINKIEWICZ, AND B. SANTORINI, *Building a large annotated corpus of english: The penn treebank*, Comput. Linguist., 19 (1993), pp. 313–330.

[136] M. MARNEFFE, B. MACCARTNEY, AND C. MANNING, *Generating typed dependency parses from phrase structure parses*, in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), 2006.

[137] P. MARTIN-RODILLA, *An empirical approach to the analysis of archaeological discourse*, in Across Space and Time: 41st Computer Applications and Quantitative Methods in Archaeology Conference, Amsterdam University Press, 2015, pp. 319–325.

[138] J. MILDINHALL AND J. NOYES, *Toward a stochastic speech act model of email behavior*, in CEAS 2008 - The Fifth Conference on Email and Anti-Spam, 2008.

[139] G. A. MILLER, *Wordnet: A lexical database for english*, Commun. ACM, 38 (1995), pp. 39–41.

[140] N. MORGAN, D. BARON, J. EDWARDS, D. ELLIS, D. GELBART, A. JANIN, T. PFAU, E. SHRIBERG, AND A. STOLCKE, *The meeting project at icsi*, in Proceedings of the First International Conference on Human Language Technology Research, HLT '01, Stroudsburg, PA, USA, 2001, Association for Computational Linguistics, pp. 1–7.

[141] P. MULLER AND A. REYMONET, *Using inference for evaluating models of temporal discourse*, in 12th International Symposium on Temporal Representation and Reasoning (TIME'05), June 2005, pp. 11–19.

[142] J. MURPHY, *The role of clinical records in narrative medicine: A discourse of message*, The Permanente Journal, (2016).

[143] J. MURPHY AND M. ROSER, *Internet*, tech. rep., OurWorldInData.org, 2017. https://ourworldindata.org/internet/.

[144] N. NEWMAN, *Overview and key findings of the 2016 report*, 2017. http://digitalnewsreport.org/survey/2016/overview-key-findings-2016/.

[145] J. NIELSEN, *Usability Engineering*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[146] R. K. NIELSEN, *People Want Personalised Recommendations (Even as they Worry about the Consequences)*, in Digital News Report, 2016.

[147] S. OKADA, Y. OHTAKE, Y. I. NAKANO, Y. HAYASHI, H.-H. HUANG, Y. TAKASE, AND K. NITTA, *Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets*, in Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, New York, NY, USA, 2016, ACM, pp. 169–176.

[148] D. O'KEEFE, J. JENSEN, AND D. JAKOB, *The relative persuasiveness of gain-framed loss-framed messages for encouraging disease prevention behaviors*, Health Communication J., 12 (2007), pp. 623–644.

[149] A. OMUYA, V. PRABHAKARAN, AND O. RAMBOW, *Improving the quality of minority class identification in dialog act tagging*, in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, June 2013, Association for Computational Linguistics, pp. 802–807.

[150] S. ORABY, P. GUNDECHA, J. MAHMUD, M. BHUIYAN, AND R. AKKIRAJU, *"how may i help you?": Modeling twitter customer serviceconversations using fine-grained dialogue acts*, in Proceedings of the 22Nd International Conference on Intelligent User Interfaces, IUI '17, New York, NY, USA, 2017, ACM, pp. 343–355.

[151] O. OWOPUTI, C. DYER, K. GIMPEL, N. SCHNEIDER, AND N. SMITH, *Improved part-of-speech tagging for online conversational text with word clusters*, in NAACL, 2013.

[152] B. PANG AND L. LEE, *Opinion mining and sentiment analysis*, Found. Trends Inf. Retr., 2 (2008), pp. 1–135.

[153] C. PAPADOPOULOS, A. CHRYSANTHI, AND P. MURRIETA-FLORES, *Thinking beyond the Tool: Archaeological Computing and the Interpretive Process*, 01 2012.

[154] M. J. PAUL, *Mixed membership markov models for unsupervised conversation modeling*, in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, Stroudsburg, PA, USA, 2012, Association for Computational Linguistics, pp. 94–104.

[155] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, J. VANDERPLAS, A. PASSOS, D. COURNAPEAU, M. BRUCHER, M. PERROT, AND E. DUCHESNAY, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[156] K. PEFFERS, T. TUUNANEN, M. ROTHENBERGER, AND S. CHATTERJEE, *A design science research methodology for information systems research*, J. Manage. Inf. Syst., 24 (2007), pp. 45–77.

[157] A. PERRIN, M. DUGGAN, AND S. GREENWOOD, *Social media update 2016*, tech. rep., PEW, 2017.
http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

[158] K. PERUMAL AND G. HIRST, *Semi-supervised and unsupervised categorization of posts in web discussion forums using part-of-speech information and minimal features*, in Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech, NAACL-HLT 2016, Stroudsburg, PA, USA, 2006, Association for Computational Linguistics, pp. 100–,Äì108.

[159] K. PETERSEN, R. FELDT, S. MUJTABA, AND M. MATTSSON, *Systematic mapping studies in software engineering*, in Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, EASE'08, Swindon, UK, 2008, BCS Learning & Development Ltd., pp. 68–77.

[160] O. PHELAN, K. MCCARTHY, AND B. SMYTH, *Using Twitter to Recommend Real-time Topical News*, in Proceedings of the Third Conference on Recommender Systems, RecSys '09, New York, NY, USA, 2009, ACM, pp. 385–388.

[161] A. POPESCU-BELIS, *Dialogue acts: One or more dimensions*, tech. rep., ISSCO Working paper n. 62, 2005.
https://pdfs.semanticscholar.org/f42a/050b37b5ca5f8dcfa6054cba2c62a9030faf.pdf.

[162] V. M. PRIETO, S. MATOS, M. ALVAREZ, F. CACHEDA, AND J. L. OLIVEIRA, *Twitter: A good place to detect health conditions*, PLOS ONE, 9 (2014), pp. 1–11.

[163] A. QADIR AND E. RILOFF, *Classifying sentences as speech acts in message board posts*, in Conf. on Empirical Methods in Natural Language Processing, EMNLP '11, Stroudsburg, PA, USA, 2011, ACL, pp. 748–758.

[164] J. J. RANDOLPH, *A guide to writing the dissertation literature review*, Practical Assessment, Research & Evaluation, 14, pp. 1–13.

[165] S. RANGANATH, X. HU, J. TANG, S. WANG, AND H. LIU, *Understanding and identifying rhetorical questions in social media*, ACM Trans. Intell. Syst. Technol., 9 (2018), pp. 17:1–17:22.

[166] S. RAVI AND J. KIM, *Profiling student interactions in threaded discussions with speech act classifiers*, in Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, Amsterdam, The Netherlands, The Netherlands, 2007, IOS Press, pp. 357–364.

[167] P. RAYSON, L. EMMET, R. GARSIDE, AND P. SAWYER, *The revere project: Experiments with the application of probabilistic nlp to systems engineering*, in Natural Language Processing and Information Systems, M. Bouzeghoub, Z. Kedad, and E. Métais, eds., Berlin, Heidelberg, 2001, Springer Berlin Heidelberg, pp. 288–300.

[168] J. READ, B. PFAHRINGER, G. HOLMES, AND E. FRANK, *Classifier chains for multi-label classification*, in Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II, ECML PKDD '09, Berlin, Heidelberg, 2009, Springer-Verlag, pp. 254–269.

[169] J. RECKER, *Opportunities and constraints: the current struggle with bpmn*, Business Process Management Journal, 16 (2010), pp. 181–201.

[170] J. R. RIDPATH, C. J. WIESE, AND S. M. GREENE, *Looking at research consent forms through a participant-centered lens*, J. of Health Promotion, 23 (2009), pp. 371–375.

[171] A. RITTER, C. CHERRY, AND B. DOLAN, *Unsupervised modeling of twitter conversations*, in HLT 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, ACL, pp. 172–180.

[172] C. ROLLAND, *Conceptual Modeling and Natural Language Analysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 57–61.

[173] C. ROLLAND AND C. B. ACHOUR, *Guiding the construction of textual use case specifications*, Data Knowl. Eng., 25 (1998), pp. 125–160.

[174] C. ROLLAND AND C. PROIX, *A natural language approach for requirements engineering*, in Advanced Information Systems Engineering, P. Loucopoulos, ed., Berlin, Heidelberg, 1992, Springer Berlin Heidelberg, pp. 257–277.

[175] M. L. ROSA, M. DUMAS, C. C. EKANAYAKE, L. GARCIA-BANUELOS, J. RECKER, AND A. H. T. HOFSTEDE, *Detecting approximate clones in business process model repositories*, Information Systems, 49 (2015), pp. 102–125.

[176] N. SAHOO, P. V. SINGH, AND T. MUKHOPADHYAY, *A Hidden Markov Model for Collaborative Filtering*, MIS Q., 36 (2012), pp. 1329–1356.

[177] S. SAKURAI AND K. UENO, *Analysis of daily business reports based on sequential text mining method*, in 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), vol. 4, Oct 2004, pp. 3279–3284.

[178] M. SAPPELLI, G. PASI, S. VERBERNE, M. D. BOER, AND W. KRAAIJ, *Assessing e-mail intent and tasks in e-mail messages*, Information Sciences, 358-359 (2016), pp. 1 – 17.

[179] K. P. SAWANT, S. ROY, S. SRIPATHI, F. PLESSE, AND A. S. M. SAJEEV, *Deriving requirements model from textual use cases*, in Companion Proceedings of the 36th International Conference on Software Engineering, ICSE Companion 2014, New York, NY, USA, 2014, ACM, pp. 235–244.

[180] L. SBAFFI AND J. ROWLEY, *Trust and credibility in web-based health information: A review and agenda for future research*, Med. Internet Research, 19 (2017), p. e218.

[181] M. SBISA, *Cognition and narrativity in speech act sequences*, Pragmatics & Beyond New Series, 103 (2002), pp. 71,Äì–97.

[182] M. SBISÀ, *How to read austin*, Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA), 17 (2007), pp. 461–473.

[183] E. A. SCHEGLOFF, *Sequence organization in interaction: Volume 1: A primer in conversation analysis*, vol. 1, Cambridge University Press, 2007.

[184] J. R. SEARLE, *Speech acts*, Cambridge University Press, 1 ed., 1969.

[185] J. R. SEARLE, *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, 1979.

[186] J. R. SEARLE, *Conversation*, in (On) Searle on conversation, J. R. Searle, H. Parret, and J. Verschueren, eds., John Benjamins, Amsterdam, 1992, ch. 2, pp. 7–30.

[187] J. R. SEARLE, H. PARRET, AND J. VERSCHUEREN, *(On) Searle on conversation*, John Benjamins, 1 ed., 1992.

[188] G. SHANI, D. HECKERMAN, AND R. I. BRAFMAN, *An MDP-Based Recommender System.*, Journal of machine Learning research, (2005).

[189] S. SHEN AND J. KIM, *Modeling the process of online q&a discussions using a dialogue state model*, in Artificial Intelligence in Education, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, eds., Berlin, Heidelberg, 2013, Springer Berlin Heidelberg, pp. 674–678.

[190] F. SHI AND C. GHEDIRA, *Improving recommender systems with an intention-based algorithm switching strategy*, in Proceedings of the Symposium on Applied Computing, SAC '17, New York, NY, USA, 2017, ACM, pp. 1668–1673.

[191] E. SHMUELI, A. KAGIAN, Y. KOREN, AND R. LEMPEL, *Care to comment?: Recommendations for commenting on news stories*, in Proceedings of the 21st International Conference on World Wide Web, WWW '12, New York, NY, USA, 2012, ACM, pp. 429–438.

[192] E. SHRIBERG, R. DHILLON, S. BHAGAT, J. ANG, AND H. CARVEY, *The icsi meeting recorder dialog act (mrda) corpus*, in 5th SIGdial Workshop on Discourse and Dialogue, M. Strube and C. Sidner, eds., Cambridge, Massachusetts, 2004, ACL, pp. 97–100.

[193] D. C. SOARES, F. M. SANTORO, AND F. A. BAIÃO, *Discovering collaborative knowledge-intensive processes through e-mail mining*, J. Netw. Comput. Appl., 36 (2013), pp. 1451–1465.

[194] I. SONG AND J. DIEDERICH, *Intention extraction from text messages*, in Neural Information Processing. Theory and Algorithms, K. W. Wong, B. S. U. Mendis, and A. Bouzerdoum, eds., Berlin, Heidelberg, 2010, Springer Berlin Heidelberg, pp. 330–337.

[195] W. STILES, *Describing Talk: A Taxonomy of Verbal Response Modes*, SAGE Series in Interpersonal Communication, SAGE Publications, 1992.

[196] A. STOLCKE, N. COCCARO, R. BATES, P. TAYLOR, C. VAN ESS-DYKEMA, K. RIES, E. SHRIBERG, D. JURAFSKY, R. MARTIN, AND M. METEER, *Dialogue act modeling for automatic tagging and recognition of conversational speech*, Comp. Linguist., 26 (2000), pp. 339–373.

[197] P. F. STRAWSON, *Intention and convention in speech acts*, The Philosophical Review, 73 (1964), pp. 439–460.

[198] W. SUN, A. RUMSHISKY, AND O. UZUNER, *Annotating temporal information in clinical narratives*, Journal of Biomedical Informatics, 46 (2013), pp. S5–S12.
2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data.

[199] M. TAVAFI, Y. MEHDAD, S. JOTY, G. CARENINI, AND R. NG, *Dialogue act recognition in synchronous and asynchronous conversations*, in Proceedings of the SIGDIAL 2013

Conference, Metz, France, August 2013, Association for Computational Linguistics, Association for Computational Linguistics, p. 117–121.

[200] M. THOMPSON, *The Challenging New Economics of Journalism*, in Digital News Report, 2016.

[201] C. THORNE, E. CARDILLO, C. ECCHER, M. MONTALI, AND D. CALVANESE, *Process fragment recognition in clinical documents*, in Proceeding of the XIIIth International Conference on AI*IA 2013: Advances in Artificial Intelligence - Volume 8249, New York, NY, USA, 2013, Springer-Verlag New York, Inc., pp. 227–238.

[202] E. THORSON, *Changing Patterns of News Consumption and Participation*, Information, Communication & Society, 11 (2008), pp. 473–489.

[203] A. TIWANA, *The Knowledge Management Toolkit: Practical Techniques for Building a Knowledge Management System*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[204] K. TOUTANOVA AND C. D. MANNING, *Enriching the knowledge sources used in a maximum entropy part-of-speech tagger*, in Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13, EMNLP '00, Stroudsburg, PA, USA, 2000, Association for Computational Linguistics, pp. 63–70.

[205] G. TRAN, M. ALRIFAI, AND E. HERDER, *Timeline summarization from relevant headlines*, in European Conference on Information Retrieval, Springer International Publishing, 2015, pp. 245–256.

[206] D. R. TRAUM, *20 questions on dialogue act taxonomies*, Journal of semantics, 17 (2000), pp. 7–30.

[207] V. TUDINI, *Conversation Analysis of Computer-Mediated Interactions*, John Wiley & Sons, Inc., 2012.

[208] A. K. VAIL AND K. E. BOYER, *Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes*, in Intelligent Tutoring Systems, S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia, eds., Cham, 2014, Springer International Publishing, pp. 199–209.

[209] D. VALLET, S. BERKOVSKY, S. ARDON, A. MAHANTI, AND M. A. KAFAAR, *Characterizing and predicting viral-and-popular video content*, in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, 2015, pp. 1591–1600.

[210] W. VAN DER AALST AND A. TER HOFSTEDE, *Yawl: yet another workflow language*, Information Systems, 30 (2005), pp. 245 – 275.

[211] B. F. VAN DONGEN, A. K. A. DE MEDEIROS, H. M. W. VERBEEK, A. J. M. M. WEIJTERS, AND W. M. P. VAN DER AALST, *The prom framework: A new era in process mining tool support*, in Proceedings of the 26th International Conference on Applications and Theory of Petri Nets, ICATPN'05, Berlin, Heidelberg, 2005, Springer-Verlag, pp. 444–454.

[212] D. VANDERVEKEN, *Meaning and speech acts*, Cambridge University Press, 1 ed., 1990.

[213] S. VOSOUGHI AND D. ROY, *Tweet acts: A speech act classifier for twitter*, in Proceedings of the 10th International AAAI Conference on Weblogs and Social Media, ICWSM'16, AAAI, 2016.

[214] G. A. WANG, H. J. WANG, J. LI, A. S. ABRAHAMS, AND W. FAN, *An analytical framework for understanding knowledge-sharing processes in online q&a communities*, ACM Trans. Manage. Inf. Syst., 5 (2014), pp. 18:1–18:31.

[215] J. WANG, G. CONG, W. X. ZHAO, AND X. LI, *Mining user intents in twitter*, in the 29th AAAI Conf., AAAI Press, 2015, pp. 318–324.

[216] S. WANG, B. NAN, B. ROSSET, AND J. ZHU, *Random lasso*, The Annals of Applied Statistics, 5 (2011), pp. 468–485.

[217] N. WEBB, *Cue-based dialogue act classification*, PhD thesis, University of Sheffield, 3 2010.

[218] A. V. D. WEEL, *Changing Our Textual Minds: Towards a Digital Order of Knowledge.*, Manchester Univ. Press, 2011.

[219] H. WEIGAND AND A. DE MOOR, *Workflow analysis with communication norms*, Data and Knowledge Engineering, 47 (2003), pp. 349–369.

[220] A. J. M. M. WEIJTERS, W. M. VAN DER AALST, AND A. ALVES DE MEDEIROS, *Process mining with the heuristics miner-algorithm.*, Tech. Rep. WP 166, Technische Universiteit Eindhoven, 2006.

[221] L. R. WHEELESS, R. BARRACLOUGH, AND R. STEWART, *Compliance-gaining and power in persuasion*, Annals of the Int. Communication Assoc., 7 (1983), pp. 105–145.

[222] E. WHELAN, R. TEIGLAND, E. VAAST, AND B. BUTLER, *Expanding the horizons of digital social networks: Mixing big trace datasets with qualitative approaches*, Information and Organization, 26 (2016), pp. 1–12.

[223] R. WIERINGA, *Design science as nested problem solving*, in Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, DESRIST '09, New York, NY, USA, 2009, ACM, pp. 8:1–8:12.

[224] ——, *Design science methodology: Principles and practice*, in Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering - Volume 2, ICSE '10, New York, NY, USA, 2010, ACM, pp. 493–494.

[225] R. WIERINGA AND A. MORALI, *Technical action research as a validation method in information systems design science*, in Proceedings of the 7th International Conference on Design Science Research in Information Systems: Advances in Theory and Practice, DESRIST'12, Berlin, Heidelberg, 2012, Springer-Verlag, pp. 220–238.

[226] A. WIERZBICKA, *English speech act verbs: a semantic dictionary*, Academic Press, 1987.

[227] P. Y. H. WONG AND J. GIBBONS, *Formalisations and applications of bpmn*, Sci. Comput. Program., 76 (2011), pp. 633–650.

[228] T. WU, F. M. KHAN, T. A. FISHER, L. A. SHULER, AND W. M. POTTENGER, *Posting Act Tagging Using Transformation-Based Learning*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 319–331.

[229] X. WU, F. XIE, G. WU, AND W. DING, *Personalized News Filtering and Summarization on the Web*, in Proceedings of the 23rd International Conference on Tools with Artificial Intelligence, IEEE, 2011, pp. 414–421.

[230] D. YANG, T. CHEN, W. ZHANG, Q. LU, AND Y. YU, *Local Implicit Feedback Mining for Music Recommendation*, in Proceedings of the sixth ACM conference on Recommender systems, ACM, 2012, pp. 91–98.

[231] F.-C. YANG, A. J. LEE, AND S.-C. KUO, *Mining health social media with sentiment analysis*, J. Med. Syst., 40 (2016), pp. 1–8.

[232] Q. YANG, J. FAN, J. WANG, AND L. ZHOU, *Personalizing Web Page Recommendation via Collaborative Filtering and Topic-Aware Markov Model*, in Proceedings of the International Conference on Data Mining, ICDM '10, IEEE Computer Society, 2010, pp. 1145–1150.

[233] R. YIN, *Applications of Case Study Research*, Applied social research methods series, Sage Publications, 2003.

[234] R. ZHANG, D. GAO, AND W. LI, *What are tweeters doing: Recognizing speech acts in twitter*, in 5th AAAI Conf. on Analyzing Microtext, AAAI Press, 2011, pp. 86–91.

[235] ——, *Towards scalable speech act recognition in twitter: Tackling insufficient training data*, in Proceedings of the Workshop on Semantic Analysis in Social Media, Stroudsburg, PA, USA, 2012, Association for Computational Linguistics, pp. 18–27.

[236] H. ZHOU, M. HUANG, AND X. ZHU, *Context-aware natural language generation for spoken dialogue systems*, in Proceedings of the 26th International Conference on Computational Linguistics, COLING 2016, Stroudsburg, PA, USA, 2016, Association for Computational Linguistics, pp. 2032–,Äì2041.

[237] M. ZIMMERMANN, A. STOLCKE, AND E. SHRIBERG, *Joint segmentation and classification of dialog acts in multiparty meetings*, in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, May 2006, pp. I–I.