

Certainty Assessment in Informal Language

Liliana Mamani Sánchez

Thesis submitted for the Degree of Doctor of Philosophy

School of Computer Science & Statistics

Trinity College

University of Dublin

April, 2015

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Wherever there is published or unpublished work included, it is duly acknowledged in the text.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Liliana Mamani Sánchez

Summary

This research studied linguistic hedges in informal language style. Hedges are expressions that show a weakened commitment of speakers with respect to what they are saying. This modification in their commitment is made for various reasons: to show their personal stance, to show their state of knowledge or to show politeness as a deference to potential readers. At the same time, their motivations stem from their caring about how they are perceived and from their wish to be more precise in their evaluations.

Earlier research around hedging for linguistic studies purposes and in particular, research in automatic methods for identification of hedges have centred around formal registers of language and targeted to determining whether a hedge is used in a speculative sense or not, without taking into consideration characteristics of the speaker experiencing the speculation. These limitations in the state of the art of hedging led me to propose an annotation scheme for manual annotation of hedges in domains where an informal language style is used, such as web forums, and to create a dataset of web forum posts annotated according to this scheme.

The annotation scheme comprises four categories of hedges, from which SINGLE-hedges and NOT-CLAIMING-KNOWLEDGE first person epistemic phrases are the most important categories. SINGLE-hedges mostly conform to the traditional conception of hedges and comprise expressions such as *maybe*, *suggest*, and *probably*. NOT-CLAIMING-KNOWLEDGE first person epistemic phrases are composed of a first person article and other modal and lexical epistemic modals. Expressions such as *I think*, *I don't know* and *IMHO* are Not-Claiming-knowledge expressions frequently occurring in web forums.

Particularly, NOT-CLAIMING-KNOWLEDGE epistemic phrases were found to be a distinctive semantic category of hedges that comprise expressions of 'lack of commitment' and 'weak commitment'. This kind of epistemic phrases was found to have a wide range of linguistic realisations and to be less ambiguous than other kinds of hedges since they include pronouns in the first person, which makes them show the speaker's involvement in the situation being uttered. The utilization of NOT-CLAIMING-KNOWLEDGE epistemic phrases as linguistic features also offers some benefits for the processing of informal language, which may help in the building of automatic methods for hedging identification in this kind of language register.

Apart from the entity representing a hedge, the Source and Scope of hedging are entities also comprehended in the annotation scheme. The Source refers to the entity in a

proposition where the hedge occurs that is experiencing or/and expressing the mental state conveyed by the hedge. The Scope refers to the part in the proposition that is affected by the hedge.

Further, linguistic hedges are studied in the context of an online web forum to explore their interactions with other forum features such as user categories, whether users like posts or not (high ratings), and polarity of sentiment in posts. This set of features, along with hedges in posts are relevant to identify prominent individuals in online communities, i.e. web forum users who are trusted by their peers. In this scenario, trust is not only fostered by the knowledge and expertise those individuals hold in their particular domains, but also it is fostered by other users' perception of benevolence of these individuals. In this dissertation, hedge use is deemed to convey this sort of interest: of providing accurate and helpful advice, and being benevolent towards other users.

Based on these considerations, this dissertation describes three kinds of studies of hedges in a corpus consisting of posts from a particular web forum domain: first, an empirical description of how the entities involved in hedging are distributed in the corpus; second, some pragmatic uses of hedges according to an existent pragmatic taxonomy where hedges are classified into content-oriented and reader-oriented classes; and third, empirical descriptions and statistical models of how hedge occurrence interacts with other features in posts.

Overall, in this study it was mainly found that posts containing hedges are more likely to be assigned high ratings than posts with no hedges, or posts with no hedges mixed with negative sentiment or no sentiment expressed.

This dissertation describes future paths of research for improving methods for automatic detection of hedges in informal styles of language, and for using hedges as features for natural language processing tasks in online communities corpora.

Acknowledgement

I would like to thank the people and institutions that helped me during this academic period of my life.

Special thanks to my supervisor Prof. Carl Vogel for his guidance and support during these years. Here, I feel words would not be enough to express my gratitude to him.

Thanks to my PhD examiners Prof. Walter Daelemans and Prof. Lucy Hederman who read my thesis and provided many useful suggestions for making it better.

Thanks to my family, my parents Elida and Dionicio and my siblings Paúl, Licelly and Jorge Luis for their support and encouragement during these years. I thank particularly my mother who from very early on imbued me with the love for learning and to whom I dedicate this thesis.

Thanks to my extended family whom always were supportive of my choice to be researcher. In particular thanks to my uncle Clemente to whom I dedicate this thesis as well.

Thanks to Dr. Martin Emms who greatly contributed to my PhD. education.

Thanks to Mikhail Timofeev who helped me to gain access to the data I worked with for this research.

Thanks to my friends and colleagues who contributed closely to my PhD research experience by providing ideas, encouragement, suggestions, food, and laughter. I learned a lot from them and was inspired a great deal by their own individual qualities: Héctor Franco Peña, Stephan Schlögl, Gerard Lynch, Alfredo Maldonado Guerra, Anne Schneider, Abigail Parisaca Vargas, Lizeth Tapia Tarifa, Carmen Klaussner, Roman Atachians, Gurpreet Singh, Erwan Moreau, Melikeh Sah, Baoli Li, Francesca Bonin, Jesús Mena Chalco, Einar Broch Jonhsen, Guina Sotomayor, Alejandra López Fernandez, Hilary McDonald, and Chiara Leva. I do not say anything in particular about them because I hope they know the extent to which I value them.

Thanks to other friends who contributed in indirect ways to my well being during the time as PhD student: Prof. Eoin O'Neill, Yenny Picha, Hilda Zevallos, Gustavo Salazar Torres, Pavel Gonzalez, Esther Salazar Gonzales.

Thanks to my housemates who always found ways to keep my spirits up: Anna and Darek Haranicz.

Thanks to the administrative personnel from the School of Computer Science and Statistics for their support, in particular to the lovely ladies from the Accounts Unit.

Thanks to the personnel from the Trinity College Library, some of who I never met and might never meet, but who let me have access to a lot of great books during these years.

Thanks to Trinity College Dublin for their Research Scholarship Program and to Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Trinity College Dublin.

Contents

1	Introduction	1
1.1	Research description	1
1.1.1	Motivation	1
1.1.2	Domain overview	2
1.1.3	Problem	6
1.1.4	Research questions	8
1.2	Outline	8
2	Theoretical Background	9
2.1	Introduction	9
2.2	Hedging Research	10
2.3	Essential concepts related to hedging	11
2.4	Dimensions of Hedging Categorization	13
2.4.1	According to the scope of the modal meaning	13
2.4.2	Semantic classifications	14
2.4.3	Lexical syntactic classifications	15
2.4.4	A Pragmatic model	20
2.5	Annotation schemes for hedging	23
2.6	Elements and Attributes intervening in the Annotation of Hedges	28
2.6.1	Hedging expression	28
2.6.2	Source or Perspective	29
2.6.3	Scope	30
2.7	Computational approaches	31
2.8	Conclusions	39
3	Building an Annotation Scheme for Hedging	41
3.1	Introduction	41
3.2	Need for a hedging annotation scheme	42
3.3	Annotation Elements	46
3.3.1	Single hedges	46
3.3.2	Source	49
3.3.3	Scope	51

3.3.4	Epistemic phrases	52
3.3.5	Syntactic hedges	60
3.3.6	Other hedges	64
3.3.7	Relations	65
3.3.8	Summary	65
3.4	Annotation work	66
3.4.1	Annotation task boundaries	66
3.4.2	Annotation tools	67
3.4.3	Annotation procedure	68
3.4.4	Annotation strategies	72
3.4.5	Overlapping issues	73
3.5	Conclusions	74
4	Theoretical and empirical issues about hedging in-domain	77
4.1	Forum dataset	77
4.2	Profiling hedges in the forum dataset	80
4.2.1	Single hedges	86
4.2.2	Not-Claiming-Knowledge Epistemic Phrases	89
4.2.3	Syntactic hedges	100
4.2.4	OTHER category of hedges	101
4.2.5	Ambiguous expressions	104
4.2.6	Conflicting cases of hedging	108
4.3	Interactions between hedging elements	109
4.3.1	Interaction of negation and speculation phenomena	109
4.3.2	Co-occurrence of multiple hedging markers	109
4.4	Profiling Other elements of hedging	111
4.4.1	Source	111
4.4.2	Scope	115
4.5	Hyland's pragmatic model in web forum dataset	117
4.6	Conclusion	123
5	Hedges and forum features interaction	127
5.1	User categories	128
5.2	Ratings	130
5.3	Signals of emotion	132
5.4	Post labelling strategies based on hedges	133
5.5	Correlation between post characterizations	134
5.5.1	Methods of analysis	135
5.5.2	Correlation of hedging and user categories	135
5.5.3	Correlation of kudos and hedges in posts	141
5.5.4	Interaction with signals of emotion	155

5.6	Conclusions	158
6	Future of Hedging Detection in Web text	165
6.1	Improvement of Automatic Detection of Hedges	165
6.1.1	Dealing with noisy text	165
6.1.2	Scope	166
6.1.3	Exploring association to other post features	167
6.1.4	Enriched information about hedges	167
6.1.5	Hedges and Questions	168
6.1.6	Degrees of uncertainty	168
6.2	Applications of automatic hedge detection	169
6.2.1	Sentiment analysis	169
6.2.2	Hedging in dialogue	169
6.2.3	Extending Analysis to Other Domains and Social Media Platforms .	171
6.3	Conclusions	172
7	Conclusion	173
7.1	Contributions	173
7.2	Summary of concluding remarks and research findings	174
7.2.1	State of the art in the study of hedging and automatic methods . . .	174
7.2.2	Annotation scheme for hedges in informal language style	175
7.2.3	Not-Claiming-knowledge expressions of hedging	176
7.2.4	Lexicons of hedging expressions	177
7.2.5	Other important elements in hedging	178
7.2.6	Use of hedges in web forums	178
7.2.7	Future research paths	181
A	My relevant publications	197
B	Forum information	199
C	Lexicons and Patterns	201
C.1	Lakoff hedges	201
C.2	Holmes syntactic patterns	201
C.3	Rubin's lexicon	201
D	Lexical items found in this research	205
D.1	Single hedges	205
D.2	NCK hedges	213
D.3	Other hedges	222

E	Data Preparation	229
E.1	Substitution by wildcards	229
E.2	Skipping files	230
E.3	Anonymisation	230
F	Description of statistical models	231
F.1	Confidence intervals	231

List of Tables

2.1	Semantic classifications for modality.	14
2.2	Most frequent grammatical classes used to express epistemic modality. . . .	16
2.3	Frequencies of hedging items in three different corpora.	17
2.4	Hyland’s summary of hedging functions and main hedging realisation devices.	23
2.5	Summary of counts from Bioscope annotations of hedges and negation words.	26
2.6	Summary of inter-annotator agreements for the Bioscope corpus annotation of hedge keywords and the scope of hedges.	26
2.7	Source of modality for various modal values.	29
2.8	Data about annotated abstracts in Light et al.’s work.	32
2.9	Performance scores for Light et. al. system.	33
2.10	Results achieved by the hedge cue word identification system by Morante and Daelemans.	34
2.11	Results achieved by the hedge scope identification system by Morante and Daelemans.	34
2.12	Best results in CoNLL 2010 Shared Tasks.	37
3.1	Words conveying uncertainty extracted from Rubin’s work	47
3.2	Most frequent epistemic markers in Kärkkäinen’s work.	56
3.3	Classification of epistemic phrases according to their main epistemic constituent.	58
3.4	Extended list of epistemic phrases that convey uncertainty.	59
4.1	Distribution of sentences per type in the annotation dataset.	80
4.2	Percentage of posts containing hedges from each hedge category in the annotation dataset (Annset) and reduced training dataset (RTD).	81
4.3	Mean, standard deviation (sd) and coefficient of variation (cv) of hedges in the Annset	82
4.4	Percentage of posts containing hedge types.	85
4.5	Frequencies for Single-hedge types in the annotation dataset	86
4.6	Frequencies for Non-Claiming-Knowledge phrasal types	91
4.7	Comparison of hedges with explicit and elliptical subject	98

4.8	Frequencies of co-occurrences of <i>I don't know</i> with other hedges in the reduced training dataset.	100
4.9	Frequency of Syntactic hedges in the Annotation dataset.	101
4.10	Frequencies for <i>Other</i> hedges in the annotation dataset.	102
4.11	Comparison of hedges and potential hedges.	105
4.12	Distribution of count of hedges per sentence in the Annset	110
4.13	Frequencies of multiple hedges per sentence	110
4.14	Frequency of co-occurrence patterns of hedging in the Annset.	111
4.15	Distribution of Non-Explicit and Explicit Source.	112
4.16	Frequency of NCK phrases Source.	113
4.17	Distribution of Inner Source attribution.	113
4.18	Frequency of single hedge occurrence that have 'Other' as Inner Source. . .	114
4.19	Single hedge types without associated scope.	116
5.1	Frequencies of users and posts across user categories	130
5.2	Most frequent character-based signals of emotion in the RTDS.	133
5.3	Most frequent pictorial smilies in the RTDS.	134
5.4	Distribution of posts according to hedge and user categories in the Annset. .	138
5.5	Distribution of hedge tokens and types per user category.	139
5.6	Average and standard deviation of post's size in number of words across user categories in the Annset and RTD.	140
5.7	Distribution of users and posts at a sampling point.	141
5.8	Overall distribution of posts according to co-occurrence of hedges and kudos.142	
5.9	Posts categories considered in model AllCats1.	143
5.10	Probabilities of a post being <i>kudoer</i> in pairwise comparisons.	147
5.11	Probabilities of a post being <i>kudoer</i> in pairwise comparisons in AllbutSynt1	150
5.12	Probabilities of a post being <i>kudoer</i> in pairwise comparisons in SinglevsNCK1	151
5.13	Results from applying the AIC test over the models for <i>kudoer</i> posts.	152
5.14	Coefficients and standard errors for three explanatory models of the number of kudo-giving events	154
5.15	Goodness of fit comparison of models to account for number of kudos events.155	
5.16	Overall distribution of posts according to co-occurrence of emoticons and hedges.	156
B.1	Number of posts and users for full and training datasets.	199
C.1	Lakoff's hedges list.	202
C.2	Keywords expressing absolute certainty proposed by Rubin.	202
C.3	Keywords expressing high certainty proposed by Rubin.	203
D.2	Frequencies for Non-claiming-Knowledge phrasal types	213
D.3	Frequencies for <i>Other</i> hedges in the annotation dataset.	222

E.1	Wildchars used for text standaridazation in web forum posts.	229
F.1	Confidence intervals for probabilities in pairwise comparisons for model AllCats1 at 95% of confidence.	232
F.2	Confidence intervals for probabilities in pairwise comparisons for model AllbutSynt1 at 95% of confidence.	233
F.3	Confidence intervals for probabilities in pairwise comparisons for model SinglevsNCK at 95% of confidence.	233

List of Figures

1.1	Hierarchy of user ranks based on expertise	5
2.1	Hyland’s categorization of scientific hedges.	21
2.2	Four dimensional model for uncertainty-certainty.	24
2.3	Types of expressions for annotation formulated by Wiebe et al.	25
3.1	Annotation process.	66
3.2	Visual interface of the annotation of example (109).	67
3.3	Stand-off annotations.	67
3.4	Coding of the annotation scheme for hedging	68
3.5	Pre-processing pipeline.	71
4.1	Distribution of posts by frequency of hedges across different hedge categories in the Annset.	82
4.2	Count of non-normalised hedge type frequencies (log scale).	84
4.3	Count of non-normalised hedge type frequencies (log scale) in the four hedge categories.	84
4.4	Percentage of hedge per type occurrence in posts from the annotated dataset.	85
5.1	Hierarchy of user ranks based on expertise (Abbreviated)	129
5.2	Distribution of the number of kudos-giving events in posts in logarithmic scale.	131
5.3	Cases of false matches for the emoticons :), 8) and *) in posts from the web forum under study.	133
5.4	Cross-distribution of posts by user category and kudos in the annotation (Annset) and reduced training dataset (RTD) datasets.	137
5.5	Cross-distribution of posts by user category and hedges in the annotation and RTD datasets.	138
5.6	Vocabulary growth curve for user categories	140
5.7	Vocabulary growth curves (hedges) for three user categories: gurus, ranked and unranked in the annotation dataset.	141
5.8	Cross-distribution of posts by kudos and hedges in the annotation (Annset) and reduced training dataset (RTD) datasets.	142

5.9	Significant Pairwise comparisons in model <code>AllCats1</code>	145
5.10	Confidence intervals for difference of estimates of effects for <code>AllCats1</code>	147
5.11	Significant Pairwise comparisons in model <code>AllbutSynt1</code>	148
5.12	Confidence intervals (left) and probabilities (right) for difference of means of effects for model <code>AllbutSynt1</code>	149
5.13	Significant pairwise comparisons in model <code>SinglevsNCK</code>	150
5.14	Plots of confidence intervals and probabilities for model <code>SinglevsNCK</code>	151
5.15	Distribution of the number kudos-giving events in posts in logarithmic scale in HEDGED and UNHEDGED posts.	154
5.16	Hasse diagram for the model <code>BinHedgesEmots</code>	157
5.17	Hasse diagram for the model <code>SingNCKEmots</code>	159

Now, mostly dead is slightly alive.

In "The Princess Bride" by William Goldman.

Chapter 1

Introduction

1.1 Research description

1.1.1 Motivation

The main motivation for this research on hedges came from the desire of characterising individuals participating in online communities such as web forums. In principle, discourse markers of uncertainty would be a non-desirable feature of language used by prominent individuals in these communities as it would point out lack of authority or lack of expertise. Markers of uncertainty, however have many-fold semantic and pragmatic interpretations beyond simply lack of certainty. The term ‘hedge’ has been used to cover this gamut of interpretations and other ones that do not deal necessarily with uncertainty but also with the degree of involvement of a speaker in a proposition.

Nonetheless, any Natural Language Processing (NLP) and Information Retrieval (IR) system relying on information about discourse would benefit from the availability of accurate certainty assessments on linguistic constructions. The expression of certainty in text is realised by the use of many lexical forms and grammatical constructions. Particularly, epistemic modals is an important category of expressions that is utilized to convey certainty or uncertainty because they help the expression of possibility. The study of epistemic modals has a central role in the study of not only uncertainty but hedges in general as they constitute the core of any study of hedges.

Although a large amount of research on hedges has been done for discourse in scientific and formal language styles, it still requires further study particularly in domains where the use of language does not necessarily follow the conventions required in academic articles and the like. Discourse contexts such as oral conversations, e-mails, chat, blogs and, in the case of this study, web forums allow a less structured and more heterogeneous uses of language.

Therefore, interest in studying use of hedges in language originated in online web forums entails more than sole interest in the linguistic underpinnings of the study of hedges in informal language, entails also the interest of finding out how hedges are used to characterise individuals using such hedging devices.

These things being equal, the study of the use of hedging expressions in web forums has a concrete motivation and goal: the automatic identification of hedges in informal language in web forums.

An automatic assessment endeavour of any linguistic phenomenon needs to be based on a model of such a phenomenon. In this case the model should be represented by a significant amount of text annotated with the main properties an abstraction of such a phenomenon would comprehend. In this way, building a model for linguistic hedging entails the need to determine what this set of properties would be. My proposal is the creation of an annotation scheme that includes the set of properties that represent the phenomenon of linguistic certainty as a concrete starting point for the study of hedging expressions in web forums. A corpus annotated with this scheme will allow for the study of hedges in relation to other linguistic and non-linguistic features in an online community.

1.1.2 Domain overview

The dataset used in this research corresponds to conversations and informative announcements in an online Web forum. The Web forum is part of the customer service facilities provided by a major multinational software company.¹ The forum's main purpose is providing a communication channel between the corporate side and customers. This channel was designed to provide customers an alternative manner of having their product-related issues addressed without appealing to traditional channels (eg. talking to a call-centre representative). All activities and types of communication via the forum are coadjutant to this main purpose.

Because of their informal nature, text extracted from web forum conversations (posts) is frequently noisy. Noisy text comprises non-cannonical writing elements such as typographical errors, misspellings, non-standard abbreviations, use of emoticons and inclusion of non-linguistic elements besides non-grammatical constructions. This feature of web forum content makes the language style used there inherently different from other more formal domains such as academic prose, newspaper articles and the like². The term "noisy text" will be often used in this document to point to this feature of web forum posts.

In the Web forum, I classified the roles that users play into four metaroles: Advice-Seeker, Advice-Giver, Commenter, and Facilitator.³ Each of these roles can be accomplished by the same individual in diverse contexts. The Advice-Seeker is the main meta-role in the Web forum, as all other meta-roles are dependent on this one. Advice-Seeker is the role normally played by customers who use the forum to look for solutions to technical

¹The company was involved as a research collaborator, but remains unnamed in this document.

²Additional insights about noisy text are given in Section 6.1.1.

³These metaroles are designed based on the assumption that forum users follow or are expected to follow Gricean maxims of quantity, quality, relation and manner in their communications. Other metaroles that could emerge from the forum interaction (e.g. internet trolls) are not considered for this description. For instance, an internet troll violates the principle of cooperation underlying these maxims when "contributes" to the forum with non-relevant topics or comments.

problems related to the use of the company software products. Depending on their level of expertise and knowledge about the products, their search can be more or less active or passive, achieved by searching existing information, describing issues and asking questions. Nonetheless, these roles can be potentially played by individuals who usually play other roles, although in different levels and contexts. Advice-Giver is the role where an individual directly addresses the questions raised by Advice-Seekers. Facilitators' hallmark is guaranteeing and providing the means for communication in the forum to run in an effective and smooth manner. They may mediate interactions between individuals playing some of the other roles, chiefly Advice-Seekers and Advice-Givers. Commenters' participation in the forum is to provide information or comments about new features in the web forum or in the products provided by the company. The information thus provided was not requested in a dialogue situation but it is seen as a way of contributing to the forum community.

Also because of its difference from other domains such as academic articles, news reports or Wikipedia articles, language style in web forums has a more varied nature. Forum users in their participation within the forum assume one or more meta-roles and use different clause types to express a broad range of speech acts that correspond with those meta-roles. This difference comes from the type of interaction within such communities of users; academic articles do not reflect direct interaction with the academic community while web forum posts reflect a dialogue setup, where some utterances such as questions can be specifically directed to a particular individual. Questions in academic articles generally address the whole academic community. Some instances of sentences that can be expected from users taking specific metaroles are showed in examples (1) to (7).⁴ A Commenter can utter the declarative sentence (1) where a promise is made. Example (2) depicts a declarative sentence mixed with a question apparently used in a sarcastic manner. Advice is given in the imperative sentence in (3)⁵ likely answering to an Advice-Seeker. Also, an Advice-Giver utters the exclamative sentence is used in (4). Giving thanks in (5) was likely used by an Advice-Seeker for answers provided to him or her. Apologetic sentences such as (6) can be used by a user taking the Commenter role. Subjective expressions can be also used to apologise as in (7).⁶ These examples hopefully give a glimpse of the type of expressions that are particularly different to the ones expected in academic articles. These are specific observations to illustrate some differences between both styles (informal in web forum vs. formal in academic articles). This research does not focus on a thorough comparison of both

⁴This labelling of sentences represents a unique document identifier number given to each post in a first-posted first-assigned basis. Although examples mainly correspond to single sentences, in many cases multiple sentences extracted from a post are provided in one example. The trailing label attached to each example signalled by 'Post:' corresponds to the post identifier as found in the original web forum dataset. Examples that are not accompanied by a citation or any sort of label were provided by myself to illustrate a concept or to provide a comparison with another example.

⁵Through this document and for the sake of anonymisation, some organization and brand names subjected to the terms of non-disclosure agreements are replaced by a tag surrounded by square brackets ([]).

⁶Linguistic examples through this dissertation are given verbatim, eg. *I was clera enoug* in (7) is shown as found in the dataset.

styles, nonetheless relevant differences will be pointed out across this document as hedges in academic articles have been taken as a starting point in this research.

- (1) We are expecting a patch for this shortly, which I'll notify this thread about when it goes live. Post: 5671
- (2) They say patience is a virtue I wonder how you get it? Post: 31730
- (3) Uninstall [Product_name] completely, removing your setting also. Post: 4181
- (4) We would not be able to recommend such approach simply because it is NOT secure! Post: 20642
- (5) Thanks, in advance, for your response. Post: 733
- (6) Sorry to hear of your problem. Post: 54942
- (7) ps: I hope I was clara enoug :P I'm a little bit tired at the moment :D :D :D
Post: 35934

Users Forum users, like people in most other contexts, may be expected to be sensitive to how they are perceived. In particular, users who contribute on a voluntary basis playing the role of Advice-Givers are likely to seek to be considered as expert as the company employees who contribute. This predicts emulation of employees by the most proficient and professional forum “netizens”, and correspondingly, a convergence of linguistic and nonlinguistic features among postings of employees as participation unfolds (cf. [Goffman, 1956]).

The raw dataset as was provided, comprises information about users belonging to 29 fine-grained ranks. From these ranks, 25 are arranged in a hierarchical fashion (Figure 1.1). Normally, every user starting her or his participation in the forum is given a rank that corresponds to a novice and can be promoted to the next rank in the hierarchy. Promotion in relation to rank occurs relies on quantitative and qualitative metrics. The first kind of metrics are related to the number of messages answered by the user, ratings the user's messages are given by other users, time online, etc. The second type of metrics includes level of expertise, knowledge about the topics of interest in the forum and other subjective measures decided by the forum moderators. The 4 remaining roles cannot be reached through means of promotion as they normally include the company's employees and ex-employees, and third-party collaborators who are given the function of moderators and administrators.

Additionally, some users from the upper levels in the hierarchy are given the additional role of guru. Individuals belonging to this group are considered the most qualified in terms of knowledge about the topics arising in the forum dynamics:

USER: Our guru [Username] is superb with handling the tool (only computer gurus like [Username] should handle [Productname], by the way...it's not something to play around with!!!), and, to my knowledge, currently the

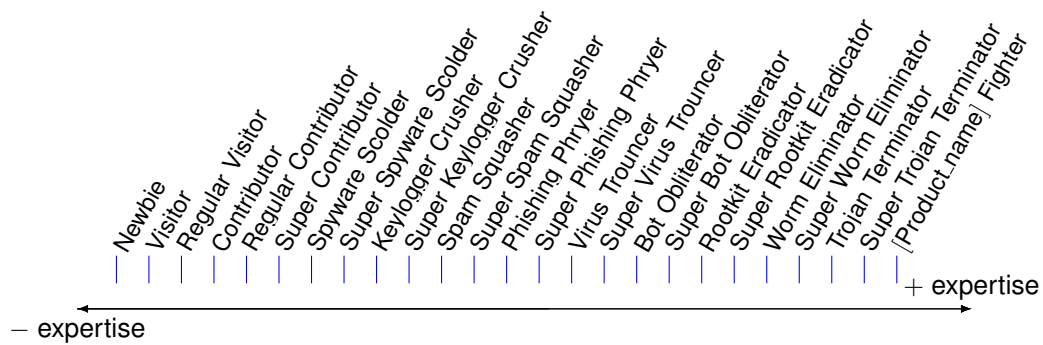


Figure 1.1 – Hierarchy of user ranks based on expertise drawn out of Table B.1.

only person who can do so in this forum. [...]

Post: 138894

Also the quality of their communication and interaction is expected to be excellent. Users from other ranks have normally a positive attitude to them and consider them to be sources of reliable information:

USER: And you think that the gurus were gurus from the very start?

[...] They worked themselves up the ranks and kept on learning things on the way [...]

The gurus [...] are modest, extremely friendly and patient [...]

Post: 118429

USER: [...] in a way u guys really r like superheroes out saving innocents for the common good.

Post: 122705

Features and Metadata Other pieces of information provided include:

- **Kudos:** Each time a user considers a post useful or finds the information given in it relevant (normally), or for any other reason accords the badge of *Kudos* to the posts. It is a qualitative feature that can be quantified: A post with n amount of kudos has been given kudos by n different users.

- Username: for each user we were provided with the login name used during his or her participation in the forum.
- Date of publishing: the original date a post was published in the online forum.
- Views: number of views a post has received after its publication in the web forum.

Types of knowledge In the dialogue scenario the forum domain belongs to, there are sub-domains that can affect the pragmatic interpretation of propositions. The conversational background⁷ of a proposition where hedging occurs might be referring to knowledge of a different nature. Winograd [1983] conceived three types of knowledge that compose the knowledge base from where individuals can appropriately select language resources to achieve his/her communicative goals: a) Knowledge of the language, b) Knowledge of the world, and c) Knowledge of the situation. Particularly the knowledge of the world and of the situation determine the kind of inferences that can be made in this research about how writers use hedging expressions. Knowledge of the world in the forum domain comprehends knowledge to the different artifacts and entities that are subject to discussion in the forum, for instance software products, other software companies, inherent properties of the software, objects in the upper domain i.e. information technologies in general, etc. Knowledge of the current situation refers to information about events where entities from the world are referred. Looking at the hedge *probably* uttered in propositions (8) and (9), in (8) *probably* is used to hedge the user's knowledge about the world, while (9) refers to a particular situation.

(8) **Probably**, a startup repair will make Win7 boot.

(9) **Probably**, I have overlooked the answer you say you wrote earlier today.

Although this distinction is not implemented in the annotation process, in Chapter 4, I will relate these concepts to the observations and conclusions drawn out of this study.

1.1.3 Problem

Previous computational approaches aiming towards detecting speculation in scientific and academic language have reached a certain level of accuracy (see Section 2.7), however neither specific methods nor categorization models have been proposed which address informal and everyday language.

Observe the following extracts in a technical support forum domain:

May be, some of the previous updates were not installed properly which caused problem when you ran subsequent LiveUpdates

⁷Term coined by Kratzer [2008].

Hello [username]

Please check your [product_name1], [product_name2], [product_name3] and you [product_name4] to make sure you are running the latest versions. All of these programs are updated often for security reasons. If these programs are old, they **can** affect other programs also and **may** have been the source of your trojan attacks that you mentioned in your post.

These fragments correspond to posts in a forum thread. The forum has the purpose of providing technical support to users of specific software applications. Users post their questions and observations by opening a topic or thread and the posts are answered by the software providers or by other users who had similar experiences. In this way, a post may be written as an answer to another post previously posted stating a specific technical problem. It can be observed that the previous posts do not convey 100% certainty about what they are stating in contrast with the following fragment:

Hi [username], Better update your current [product_name5] to the latest version - [product_name6], and then check for this problem. If you have a current subscription to any version of [product_name7], you can directly download the [product_name8] here: [product_name9] [link1] [product_name10]: [link2] [product_name11] is also available at the [product_name12]. When you install the update, it will remove your previous [product_name13] product and apply the subscription information to [product_name11]. After installing the [product_name14], run [product_name15] repeatedly until you see the message "No more updates" and then restart the computer. Let us know the results.

The language used in this message is direct and free of speculative markers or hedges like the ones in the two previous text fragments (*may, can*).

The ambiguity of speculation markers makes the task of identifying speculative sentences hard. Look at the examples (10) that contain the *sort of* named speculation marker when their presence in the sentence introduces a speculative sense. However the presence of these particles do not indicate speculation in all cases, such as in (10a), where *sort of* has the sense of 'kind of', while it conveys doubt when used in (10b) and in (10c).

- (10) a. I welcome this **sort of** post for everyone else reading this thread
 b. Thanks!! I was **sort of** thinking that..
 c. Problem 1 is **sort of** resolved! :(

This ambiguity present in the technical forum scenario is evidenced by Ozgur & Radev [2009] in the biomedical domain as well. They reported that speculative markers are used in speculative contexts in less than 50% of the cases in abstracts from the Bioscope corpus [Szarvas et al., 2008]. Since more than 50% of markers are used in a non-speculative fashion, it makes sense creating a method for disambiguating the sense of a speculative marker.

1.1.4 Research questions

The research questions that emerged for this research are:

- Which expressions of uncertainty are used in a web forum domain?
- How is uncertainty expressed in web forum posts?
- What categorization models for certainty assessment exist so far that can be improved upon?
- Is the expression of uncertainty in this domain related to user categories and other features in web forum posts?
- Which methods can be used to automatically produce accurate certainty assessment from discourse in this specific domain?

1.2 Outline

Through this document, I will frequently draw comparisons with similar annotation works or theoretical research to highlight what has already been done and to show practical differences between various domains. Chapter 2 describes theoretical work in hedging from their conception as modals and state of the art in the research of hedges in computational linguistics. Chapter 3 addresses the problem of annotating linguistic hedges in the domain under study towards meeting the goal of providing a scheme to be used in automatic identification of hedges. To this end, an annotation scheme and the necessary steps for manual annotation assisted by automatic mechanisms are described in that chapter. Chapter 4 presents findings found in a dataset where hedging expressions are annotated applying the proposed annotation scheme and elaborates on empirical observations of how hedges from different categories are used in web forum posts. Chapter 5 shows various observations of how hedges co-occur with other non-linguistic web forum features around posts and describes some statistical models formulated with explanatory purposes of finding out associations between hedges and these features. Chapter 6 elaborates on how the results from this research can be used in similar domains to the one under study and how some aspects of my research could be improved as future work. Finally, a summary of this research and conclusions are provided in Chapter 7.

Chapter 2

Theoretical Background

2.1 Introduction

Although a large amount of work addressing the concept of modality and speculation has been conducted in language studies and the idea of how people use some instances of language to make their speech less categorical has already been addressed, it was only since Lakoff [1973] named words denoting speculation as “hedges” that this category of words obtained an important status.

Going years backwards before Lakoff’s contribution, the role of speculation in discourse has been studied since at least as far back as Aristotle. In his *Rhetoric* [trans. 1991], he covers topics related to probability arguments [trans. 1991, 1400a] and purity in discourse [trans. 1991, 1407a,1407b]. He considered one of the rules for ensuring purity was the avoidance of ambiguous terms or the use of general terms. He compared this kind of language to the one used by those who “having nothing to say, yet pretend to say something” and soothsayers who by talking in ambiguous and general terms, have less chance of making mistakes, i.e. soothsayers do not define the time when a prediction is going to take place. After Lakoff’s work was published, the coined denomination has been used widely in linguistics and the concept dealt with in similar magnitude. In this chapter, concepts related to the study of hedging and contributions of academics who have described important aspects of hedging relevant to this dissertation will be pointed out.

This chapter is structured in the following manner: Section 2.2 relates a summary of the original contribution of people who have described the essentials of the phenomenon of hedging. Section 2.3 attempts to describe an abridged taxonomy of concepts related to hedging as further chapters will frequently point at these concepts. Section 2.5 summarizes the main contributions that addressed the creation of annotation schemes for the phenomenon of hedging. While the basics of linguistic phenomena related to hedging will be described in these sections, in later chapters I will engage in deeper discussion and reference to studies that may have not been mentioned in this section. Particularly, this will be done in Chapters 3 and 4 that address results of the annotation work and corpus study. In Section 2.7, advances in the automatic identification of hedges will be described and compared to some

extent. Conclusions on the concepts presented in this chapter will be given in Section 2.8.

2.2 Hedging Research

Lakoff [1973] designed hedges as “words whose meaning implicitly involves fuzziness - words whose job is to make things fuzzier or less fuzzy” and they not only reveal degree of category membership but very important dimensions of meaning. He started by establishing connections between different typicality judgments and degrees of truth attributed to them according to fuzzy set theory concepts such as degree of membership and fuzzy logic operations [Zadeh, 1965].

To Lakoff, the study of hedges is important since it raises interesting questions in the exploration of vagueness and fuzziness inside a formal semantics approach. By analysing different hedge instances on a case-by-case basis, Lakoff was able to present different interactions which hedges may be subjected to, such as context, degrees and respects of similarity, degrees and respects of truth, intensity, and other modifiers.

He defined four types of criteria for category membership, where the hedges: *technically*, *strictly speaking*, *loosely speaking* and *regular* determine which kind of criteria (for categorisation) is being established in a proposition.

- (A). Definitional
- (B). Primary
- (C). Secondary
- (D). Characteristic though incidental

Definitional criteria are capable of conferring membership to a certain degree depending whether other primary and secondary criteria are met. Lakoff [1973, p. 238] uses (11) and (12) to depict distinctions between the different types of criteria. In these propositions, *technically* asserts that definitional criteria are met but still some important criterion for category membership is not met, and *strictly speaking* requires both types of criteria, as Kay [2005, p. 690] defines: “When the words fit the facts and the rules are followed, one speaks strictly”.

- (11) a. A whale is **technically** a mammal. (true)
- b. **Strictly speaking** a whale is a mammal. (false)
- c. **Loosely speaking**, a whale is a fish. (true)
- (12) a. Nixon is **technically** a Quaker. (true)
- b. **Strictly speaking**, Richard Nixon is a Quaker. (false)
- (13) a. My brother is a **regular** fish. (true)
- b. **Loosely speaking**, my brother is a fish. (false)
- (14) A whale is a fish.

In (11a) and (11b) there is not such distinction since by definition a whale is a mammal. (12a) is true since Nixon may be a Quaker in a definitional sense, on the other hand (12b) is false because he does not meet other criteria that would make him strictly speaking a Quaker, Lakoff considers that Nixon does not meet the religious and ethical views proper of Quakers.¹

Examples (11b) and (11c) illustrate the distinction between primary and secondary criteria, propositions where example (14) is hedged. Important criteria that make whales members of the mammal class are considered when (11b) is asserted. The same does not happen in (11c) where secondary properties like the fact they can swim are considered making a coherent proposition out of (14). What works for a whale does not work for *my brother* in (13b), that shows how some properties, despite referring to some primary characteristic, do not confer any degree of membership. Although (13a) is acceptable as *my brother* has some incidental characteristic of a fish, the former proposition implies that *my brother* in some degree is member of the category fish, which is not true. The fact that incidental characteristics do not confer membership on their own made them relevant for understanding at least one type of metaphor. Imitating Lakoff in this example, (13a) is a metaphor to indicate that *my brother* swims well or maybe has a weird smell. In these metaphorical cases, when some properties do not confer category membership, they might be used in a literal sense. Subsequently, Lakoff in ‘Metaphors we live by’ [Lakoff and Johnson, 1980, p. 124] consolidated this conjecture when asserting that hedges as instruments reveal the open-endedness of categories used in metaphors since someone could always choose to confer some properties to an object in order to assign it a category.

All these criteria compose an important categorization scheme of hedge properties and their influence in the meaning of a proposition. Lakoff [1973] has also offered an alternative thinking to philosophers who considered that pragmatic aspects of meaning are irrelevant to the assignment of truth values. For him, semantics cannot be taken independently of pragmatics.

Lakoff’s work has greatly influenced natural language research by calling attention to hedges. However, he was not originally concerned with the communicative value of the use of hedges like the ones mentioned above,² but with their logical properties. Research in fuzzy logic hedges and algebra of hedges [Kochen and Badre, 1974, Hersh and Carmazza, 1976, Ho and Nam, 2002] is an area closer to Lakoff’s original conception of hedges, however it is out of the scope of this study.

2.3 Essential concepts related to hedging

After Lakoff’s work on hedges, the concept of hedges has been used in various forms and interpretations. For instance, Brown & Levinson [1987, pag. 145] defined a hedge as “a

¹Kay pointed out another case where an interrogated group of people hear *technically* and *strictly speaking* as synonyms.

²For sake of illustration, the original list of Lakoff’s hedge expressions is included in C.1.

particle, word or phrase that modifies the degree of membership of a predicate or a noun phrase in a set; it says of that membership that it is partial or true only in certain respects, or that it is more true and complete than perhaps might be expected” while in most of the articles in computational linguistics, a hedge has the connotation of uncertainty expression device.

Thus, two main interpretations emerge: the first one comes more directly from Lakoff’s use of hedges as modifiers that intensify or de-intensify the commitment in a proposition, and the second one refers to the set of modifiers that make a proposition less certain. Skelton [1997] has named both interpretations as “traditional hedges” and bare “hedges” respectively.

Some concepts, namely: subjectivity, epistemic modality and hedging, are fundamental for an accurate understanding of how the state of art related to the weak expression of certainty has evolved and been addressed in different disciplines. Also, these concepts will hopefully provide a frame to better illustrate the linguistic phenomena addressed in my research.

Wiebe et al. [2001] defined subjectivity as a set of “aspects of language used to express opinions and evaluations” in contrast to objectivity, which focuses on the aspects to present factual information. Evaluation and speculation are the main types of subjectivity. Observe sentences (15a) and (15b): they present negative and positive judgements, emotions, opinions or evaluations. In contrast, sentence (16) shows a case of speculation containing the speculative expression *may be*. According to Wiebe et. al. the speculation category is composed of “anything that removes the presupposition of events occurring or states holding, such as speculation and uncertainty” [Wiebe et al., 2001, p. 2]. For instance, in sentence (16) the presupposition of absolute certainty that the subject *he* is merely more mindful of athletics than of aesthetics at the present time is not valid because of the presence of *may be*. Furthermore, examples (17) and (18) illustrate the contrast between subjectivity and objectivity respectively. Sentence (17) does not refer to a fact but a complaint (in quotation marks) whereas sentence (18) refers to factual information about a lawsuit.

- (15) a. I had in mind your facts, buddy, not hers.
 b. We stand in awe of the Woodstock generation’s ability to be unceasingly fascinated by the subject of itself.
 [Wiebe et al., 2001, p. 2]
- (16) But it **may be**, also, that he is merely more mindful of athletics than of esthetics at the present time.
 [Francis and Kucera, Revised 1989, C09]
- (17) “The cost of health care is eroding our standard of living and sapping industrial strength,” complains Walter Maher, a Chrysler health-and-benefits specialist.
 [Wiebe et al., 2001, p. 2]
- (18) Northwest Airlines settled the remaining lawsuits filed on behalf of 156 people killed in a 1987 crash, but claims against the jetliner’s maker are being pursued, a

federal judge said.

[Wiebe et al., 2001, p. 2]

Epistemic modality is another important concept related to the expression of certainty and central to the main components of my proposal in this study. As such, it will be discussed in Section 2.4.2.

Some other studies deal with hedges by positing them in a range of degrees that go from complete uncertainty to complete certainty. Formal semantic descriptions of the gradability in epistemic modals have been proposed, e.g. [Lassiter, 2011, Yalcin, 2010]; I do not, however, adopt a formal semantics approach nor do I intend to describe hedges by their degree of certainty, and therefore will not expand in providing details about these approaches. Rubin et al. [2005b] particularly uses a more empirical approach for describing expressions by their degree of certainty. Nonetheless, I will expand more on Rubin's contributions to the research in annotation of hedges in later sections where I discuss this topic.

2.4 Dimensions of Hedging Categorization

In this section, I will describe the main works that attempt a categorization of hedges. While hedges can be categorized according to their lexical and syntactic features, research on creating a taxonomy of hedges according to their semantic and pragmatic functions has also been carried out.

2.4.1 According to the scope of the modal meaning

Epistemic modals have been considered as expressions of the degree of the speaker's commitment to the proposition where these modals are used. There is, however, no consensus on how this degree of commitment is realised by epistemic modality. While some authors favor an approach where epistemic modals can be assigned to discrete categories, others do not share this view as they state that humans when asserting their commitment, do not think in terms of a scale [Nuyts, 2001]. A classification of the main types of modality is needed to put epistemic modality into context. Portner [2009] mentions three types of modality according to the scope it covers in discourse:

- Sentential modality. The modal affects the whole sentence. Modals in this category comprise traditional modal auxiliaries and sentential adverbs such as *maybe*.
- Sub-sentential modality. The modal meaning is expressed in constituents smaller than the whole sentence or full clauses. Some representatives of this class are modal adjectives and nouns such as *possible*, *necessary*, *possibility*, etc. There is a lot of overlap in the use of modals at sentential and sub-sentential level. Portner gives the example of *possible* as representative of sub-sentential modality but when enclosed in the expression *It is possible that S*, it transcends to the sentential level as this expression is deemed to have identical semantic function as *may*.

Table 2.1 – Semantic classifications for modality. From [Portner, 2009, p. 140]

Traditional	Epistemic Epistemic		Deontic	Root	Dynamic	X
Portner [2009]	Epistemic	Deontic	Priority Bouletic	Teleological	Dynamic Volitional	Quantificational
Brennan [1993]	Epistemic		Deontic	Root	Dynamic	Quantificational
Hacquard [2006]	Epistemic	True deontic	Goal-oriented	Root	Ability	X

- Discourse modality. Portner points to evidentiality, clause types, performativity and some types of sentential modality as aids to the expression of discourse modality such as modal subordination.

This classification covers both deontic and epistemic modality. Other kinds of modality mentioned by Portner (Table 2.1) may conflict with epistemic modality. In the next section, I delve deeper into explaining some distinctions between epistemic and deontic modality.

2.4.2 Semantic classifications

Traditionally, modality has been divided between epistemic and deontic modality. In literature, a distinction between epistemic and root modality has also been done as to cover those cases where modals do not fall in either epistemic or deontic categories. Particularly Portner [2009] has summarized a broad range of types of modality which I show for sake of illustration in Table 2.1, I do not intend to explain all of them as only some of those categories are involved in the phenomenon of hedging.

The topics of modality and specifically epistemic modality have been addressed extensively and intensively in literature. Therefore, I will not try to make a thorough description of the topic beyond what is needed for the understanding of this research. Linguists such as Portner [2009], Kratzer [2008], and Hyland [1998], to mention but a few, have focused on exploring the different qualities of epistemic modality and I will refer to them when needed.

Epistemic modality is a category that is used to convey a speaker's attitude towards the truth or reliability of his or her assertions in contrast to deontic modality where her or she expresses obligation, permission or suggestion [Finegan, 1992]. Examples (19)-(21) show the contrast between sentences depicting epistemic modality and assertions [Finegan, 1992, p. 192]. Also, we can see the contrast in the use of epistemic and deontic modals in (19a) and (19b).

- (19) a. She has **probably** left town by now. (probability)
 b. She must leave town by now. (obligation)
 c. She has left town by now. (assertion)
- (20) a. Harry **must've** been very tall when he was young. (conjecture)

- b. Harry was very tall when he was young. (assertion)
- (21)
- a. They **may** come to the party. (possibility)
 - b. They are coming to the party. (assertion)

The way individuals express this attitude has to do with the knowledge they have about the domain their assertions are contained in. In traditional definitions, epistemic modality expressions were given a status of propositional content modifiers based on the writer's attitude. For Halliday [1970], epistemic modality:

“ is the speaker's assessment of probability and predictability. It is external to the content, being a part of the attitude taken up by the speaker: his attitude, in this case, towards his own speech role as 'declarer' ”.

This conception has evolved to definitions that bound epistemic modality to evidentiality. Evidentials are members of a semantic category that express the degree of reliability of information. Therefore, they are also a means to express certainty but they are more commonly known as a grammatical category in languages other than English. Other authors such as Aikhenvald [2004] support the idea there is no evidentiality in English. For instance, Gisborne & Holmes [2007] states that verbs of appearance are evidential because they indicate the evidential source for the proposition. Proposition (22a) says that Fulgencio's appearance is the reason for inferring that he is ill and (22b) says the sound he makes is the reason for inferring that he is ill.

- (22)
- a. Fulgencio looks ill.
 - b. Fulgencio sounds ill.

This text will cover some aspects of epistemic modality more intensively than others as it will be required by the nature of the domain under study. While I do not intend to fully characterize discourse in web forum conversations, and not even the particular phenomenon of epistemic modality, my main interest is the study of phenomenon of hedging in this domain, and how it is correlated to user categorizations that are the main themes underlying my research questions.

2.4.3 Lexical syntactic classifications

A large number of studies on the devices to express uncertainty has been done in the area of epistemic modality around modal auxiliaries. This category and the most important ones found in literature will be described in this section.

Holmes [1988] developed an analysis of linguistic devices for expressing doubt and certainty being taught in textbooks for learning English as a Second Language (ESL). Besides her instruction-related goals, she has explored diverse corpora and her work is broadly cited in studies of corpus analysis of epistemic modality. She explores the main grammatical parts of speech and their comparative frequency across corpora. The corpus compiled by Holmes

Table 2.2 – Most frequent grammatical classes used to express epistemic modality. These frequencies represent percentages found by [Holmes, 1988] in a corpus of 50,000 words.

Grammatical class	Speech	Writing	Total
Modal verbs	42.4	36.8	40.2
Lexical verbs	31.5	35.9	33.3
Adverbials	21.5	12.8	18.1
Nouns	2.3	7.7	4.5
Adjectives	2.3	6.6	4.0

comprises a written and a spoken English corpus, with 25,000 words each. Table 2.2 shows percentages of grammatical categories addressed in Holmes' work. These five classes are the main ones addressed in studies of lexical devices for epistemic modality and hedging in general because of the frequency of their use.

Hyland [1998] also focuses on these categories to study hedging in a corpus of journal research articles (Journal RAs) compiled by him and makes a comparative study with other corpora in American and British English: JDEST [Qiao and Huang, 1998] and Brown/LOB [Kennedy, 1987b]. He uses the term 'lexical hedges' to differentiate them from other grammatical categories. Each type of lexical hedge is closely explored to describe how they convey an epistemic meaning and discussion is provided on the contribution of these categories to nuances of uncertainty expression. The most frequent lexical types found in Hyland's study are shown in Table 2.3.

Varttala [1999] also provides a discussion of the lexical hedging categories in medicine articles. A set of 15 articles was extracted from the journal *Scientific American* and from *The New England Journal of Medicine*. He restricts to the study of 80 different lexical types belonging to these categories and provides in some cases additional sub-categorizations.³

In the next subsections, the main kinds of lexical hedges: modal verbs, lexical verbs, modal adverbs, epistemic adjectives, epistemic nouns and conditionals will be reviewed.

Modal verbs

Some academics (e.g. Portner [2009]) draw a distinction between the concept of modal verbs and modal auxiliaries; in the context of this research, both categories are considered to be equivalent. Hedging expressions in this category comprise: *can(not)*, *could*, *may(not)*, *might*, *must*, *need*, *shall*, *should*, *ought*, *would*, *will*.

Modal verbs constitute the core devices used to express epistemic modality and it is the most studied category in modality studies in linguistic, philosophy of language and logic. This kind of hedge is frequently ambivalent in the modal meaning conveyed as it can be used to express either epistemic or deontic modality. *must* in (23b) conveys a deontic meaning of obligation in contrast to (23c) that conveys a meaning closer to the epistemic modal *may*

³I will mention Varttala's work later in Section 3.3.2 as he engages in a relevant discussion about the occurrence of a hedging phenomenon in specific rhetorical structures.

Table 2.3 – Frequencies of hedging items in three different corpora as compared by Hyland [1998, p. 149]. Modal verbs are highlighted to differentiate them from other common lexical hedging types.

	Item	Journal RAs (75,000) ¹	JDEST (435,850) ²	Brown/LOB(J) (350,000) ³
1	indicate	10.8	3.2	4.3
2	would (not)	10.4	16.8	16
3	may (not)	9.2	8	4.4
4	suggest	9.1	3.7	3.9
5	could	6.4	3.2	0.4
6	about	4.0	*	*
7	appear	4.0	2.7	3.7
8	might(not)	3.6	2.4	4
9	likely	2.8	2.5	2.7
10	propose	2.8	1.6	0.9
11	probably	2.7	1.6	2.8
12	apparently	2.7	*	2.8
13	should	2.4	0.8	0.4
14	seem	2.3	4	7.7
15	possible	2.3	1.5	1.3

¹ The Journal RAs set is the corpus built by Hyland. It comprises 26 research articles from refereed journals in the fields of cell and molecular biology. These articles are written in American and British English and as a whole consist of 75,000 words. The journals from where the articles came were selected by consulting specialists in the area and the 1992 Journal Citation Reports [SCI, 1993], so articles with high number of citations could be selected.)

² The JDEST (Jiao Da English for Science and Technology) Corpus consists of texts randomly extracted from theses, textbooks, and other academic documents in British, American and other countries English. This corpus comprises 2,000 units of about 500 words each. It was created in 1985 by Jiao Tong University in Shanghai.

³ The Brown/LOB corpora comprehends texts from the category J or "learned" samples, that contain written English for academic purposes. The Brown corpus collects texts in American English and the LOB, British English. Both together produce a set of documents of 350 000 words equivalent to 1000 pages of text.

in (23a).

- (23) a. Morten **may** have left town by now. (possibility)
 b. Morten **must** leave town by now. (obligation)
 c. Morten **must**'ve left the town yesterday. (conjecture)

Hyland [1998] provides a comprehensive description of these modals, their frequencies in text from different sources extracted from the state of the art in epistemic modality, and how they are used in academic texts to express epistemic modality. In Table 2.3 Hyland shows them alongside lexical verbs as the most frequent hedging items in diverse corpora.

Lexical verbs

This is another category of hedging devices that has been thoroughly explored in epistemic modality and are “the most transparent means of coding the subjectivity of the epistemic source and are used to hedge either commitment or assertiveness” [Hyland, 1998, p. 119]. Verbs such as *seem*, *think*, *believe*, *suggest* are used for diverse epistemic purposes. Varttala [1999] divides them into two categories in a context of research articles: epistemic reporting verbs and semi-auxiliaries.

Epistemic reporting verbs have the function of reporting the writer's or somebody else's mental states so the writer can express the tentativeness of his/her propositions as in (24) and (25). Semi-auxiliaries such as *appear*, *seem* and *tend* are used to introduce another verb in a way they indicate epistemic possibility, as in (26).

- (24) Post **proposed** that manic-depressive illness progresses in a similar fashion, each episode facilitating the next one.
- (25) Our findings **suggest** that quantitative techniques to measure volume are essential to define the subtle abnormalities of schizophrenia.
- (26) The efforts of several laboratories have together yield at least one technique that **seems** to work well [...].
 [Source: Varttala [1999, p. 186]]

Lexical verbs are also amongst the most frequent hedging items found by Hyland (cf. rows 1,3,7,10, and 14 in Table 2.3); 10.8 and 9.1 occurrences per 10,000 words for *indicate* and *suggest* respectively. [Rubin, 2006] reports 56 out of 1,727 of markers of certainty correspond to verbs of mental states and attribution to denote discrete levels of certainty that go from uncertainty to absolute certainty (see Section 2.5).

Modal adverbs

Also called epistemic adverbs, they indicate epistemic possibility. Frequent ones mentioned in various sources [Holmes, 1988, Fraser, 2010, Varttala, 1999] are: *perhaps*, *probably*, *apparently*. [Hyland, 1998] found 36 different types of adverbs in his corpus of journal

articles, most frequent reported: 2.8 occurrences per 10,000 words for each *apparently* and *probably*, 2.4 for each *relatively* and *essentially* and 2.1 occurrences for *generally*.

They normally mark sentential modality as they can appear in different positions in the sentence or verbal clauses they occur in [see (27)].

- (27) a. The end of human civilization is coming soon, **probably**.
 b. **Probably**, the end of human civilization is coming soon.
 c. The end of human civilization is, **probably**, coming soon.

Epistemic adjectives

Epistemic or modal adjectives such as *probable*, *(un)likely* and *possible* are used to express uncertainty. According to [Holmes, 1988], they occur more frequently in written than in spoken discourse. They are sub-sentential modifiers of certainty when used accompanying a noun as in (28) or as a sentential modal as illustrated by (29).

(28) In the **unlikely** event of meteorite impact, we will be able to know it in advance.

(29) It is **unlikely** a meteorite will hit this town.

Epistemic nouns

This is a less frequently used category of hedges. In the corpus compiled by Holmes [1988], nouns such as *assumption*, *belief*, *doubt* and *idea* have each from 1 to 5 occurrences in 50,000 words as they compose 4.5% of found lexical epistemic devices (cf. Table 2.2). Hyland mentions only *estimate* as an epistemic noun occurring rarely in his corpus. Rubin does not mention nouns as particular means of expressing uncertainty, however she counted *possibility* with 5 occurrences out of 1,727 items for expressing various degrees of certainty. Nouns expressing tentativeness behave in the same way as adjectives when they transcend to a sentential modality level as in (30).

(30) There is the **possibility** that a meteorite will hit this town.

Conditionals

Hyland [1998, p. 145] places conditionals in a strategic category of hedges as they are used to refer to 'limitations of model theory and method' in research articles. He described conditionals as achieving three general functions with relation to hedging: a) By presenting hypothesis as conditions that may be unlikely to be fulfilled or true, and therefore the hedged consequence will not occur (31). b) The truth of the condition in the if-clause remains as an open question, consequently the accuracy of theoretical or descriptive claims is hedged (32), and c) they do not hedge the efficacy of theory or models but present real world contingencies or hedge the precision of results (33). However, Hyland does not pinpoint clear features of when a conditional poses as a hedge or when it does not.

- (31) These results suggest that if a flavonoid mutant with unaltered sinapate accumulation were available, it would be more sensitive to UV-B than is tt 4. (=but it is not available)
[Hyland, 1998, p. 146]
- (32) **If** we assume that the antisense gene is also active in guard cells, it seems that the stomata do not use the Calvin cycle (either in the guard cells or neighbouring mesophyll cells) to monitor p_l to set the appropriate conductance.
- (33) Initially, the gradient of the variable fluorescence curve will increase if there is a sigmoidal component present but will decrease if the curve is ..
- (34) ... indicating that only very small amounts, **if any**, of additional carotenoids like antheraxanthin could be present.
Hyland [1998, p. 147]⁴

2.4.4 A Pragmatic model

An important linguistically motivated categorization of hedges was done by Hyland [1996, 1998] in a domain of scientific research articles. To Hyland, hedging represents the writer's attitude within a particular context and cannot be fully understood if social and institutional contexts have not been taken into account. He designed a functional framework of hedges based on the analysis of a set of 26 articles (75,000 words) from leading journals in the field of cell and molecular biology. Hyland defined two types of statements used by writers to back up their claims: a) statements accepted by the discourse community as truths about the world that are expressed as categorical assertions and b) hedged or non-factive statements used to assert knowledge until certain degree. As Hyland points out, most of the work in science is of non-factive character, referring to what is possibly true rather than what is certain, for instance:

- (35) **I suggest** therefore that D1 degradability **must be** causally linked to Q_B site occupation which in turn determines PEST region accessibility to
[Hyland, 1996, p. 435]

The diagram in Figure 2.1 summarizes the different statement categories that Hyland described. The first hedge categorization is the division of non-factive statements into Content-oriented and Reader-oriented hedges. Content-oriented are divided into accuracy-oriented and writer-oriented and finally accuracy-oriented hedges have a further classification into attribute and reliability hedges. Notwithstanding this categorization, Hyland advocates for a fuzzy model that he calls polypragmatic because of the nature of hedging devices that convey different various meanings and interpretations: "Particular forms often convey more than one function and a complex overlap of usage suggests that the precise motivation for employing a hedge may not always be clear" [Hyland, 1996, p. 437].

⁴The three last examples were borrowed from Hyland's work and from that specific page.

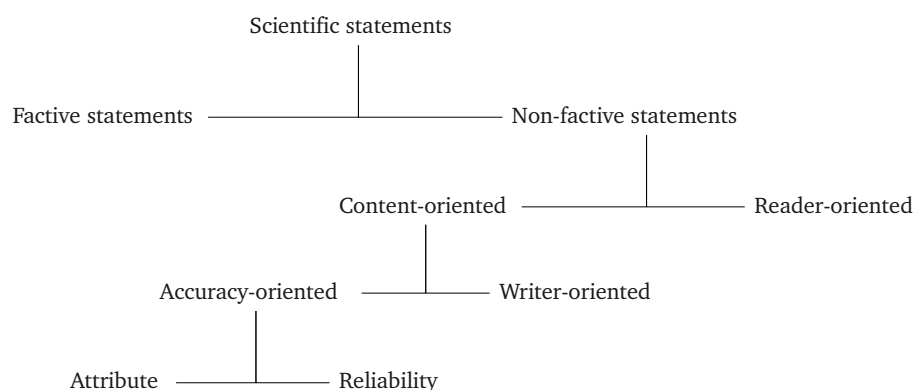


Figure 2.1 – Hyland’s categorization of scientific hedges.

The most specific categories (Reader-oriented, Writer-oriented, Attribute and Reliability) are worth describing and illustrating since Hyland suggested their importance in determining core cases of hedging.

Attribute hedges have the purpose of communicating the accuracy conveyed by the described phenomena as they highlight a deviation from an ideal model where the certainty is 100% when a particular phenomenon is reported. This deviation is expressed by describing actual attributes of the phenomenon and accuracy is achieved by hedging the description of these attributes. For instance, the following examples used by Hyland show this deviation as each claim intends to give an accurate description of the phenomena involved. The response reported in (36) has not an absolute feature, therefore insights about the response are hedged to indicate non-conformity with a simplistic description of its nature. Similarly, in example (37), *approximately* accompanied by a hedge in the form of a number in percentage is used to provide an accurate description of a temperature value.

(36) The response of the assembly of PSII proteins to the solute environment is **unique in some ways, but quite normal and predictable in others**.

(37) ... decreases by **approximately** 60% at 44°C.

[Hyland, 1996, p. 440]

Reliability hedges have the purpose of conveying the writer’s confidence in an assertion being true, according to Hyland such a user means: ‘I do not speak from secure knowledge’. Most reliability hedges are realised by epistemic modal verbs, epistemic adjectives, nouns and adverbs. The following examples convey doubt (38) and show knowledge limitations explicitly (39). In scientific articles, where a reported phenomenon’s features are not completely known to researchers, hedges are used to convey this limitation in their knowledge.

(38) This modification could **possibly** play a role in substrate binding.

[Hyland, 1996, p. 442]

(39) **It is not known whether** such a weak temperature response

[Hyland, 1996, p. 443]⁵

Writer-oriented hedges are different from accuracy-oriented hedges, they intend to prevent criticism against the writer by diminishing his/her commitment to the proposition that contains the hedge. One of the strategies to limit full commitment is by using general claims. Hyland explains that these hedges can be confused with reliability hedges but points out that writer-oriented hedges include an additional reluctance to make a commitment to what it is reported. However, Hyland does not specify which kind of linguistic element helps in the assertion of this additional reluctance. He also highlights that making an accurate distinction between hedging writer's confidence in the accuracy of an assertion and hedging writer's personal commitment is not always possible if looking at a particular hedging expression in an isolated fashion. However, analysing the context that circumscribes it should help to decide to which pragmatic category the hedge belongs to. In example (40), Hyland shows how the writer avoids explicit responsibility for an assertion, while example (41) shows a passive construction that realises the lack of writer agency.

(40) **The present work indicates** that the aromatic nng to which the carboxyl group is bound is not necessary, provided that a bulky substituent is present.

(41) The BS fraction **is assumed** to originate from the center of the . . .

[Hyland, 1998, p. 171]

Reader-oriented hedges are used to show commitment to the proposition being asserted with the purpose of ensuring, for the writer, the benefits of his/her work being acknowledged by the reader. For instance, in example (42), the writer makes a statement about his or her results trying to avoid conflict at the same time. In another example (43), the writer intends to weaken the criticism by avoiding attribution to a particular source. In (44), *I believe* is used as personal attribution to soften the illocutionary force of the writer's assertion with the purpose of easing its acceptance. A clear difference from writer-oriented hedges is that in these the writer agency is absent: (44) could be derived into a claim like (45).

(42) **We do not know the reason for the discrepancy** between our results and those of Ngerprintsin et al, but **it might reflect** genetic differences in the cultivars employed

(43) **In spite of its shortcomings, the method has been widely employed** to evidence this type of modification in a number of genomes including plastid and nuclear ones

(44) **I believe** that the major organisational principle of thylakoids is that of continuous unstacking and restacking of sections of the membrane

[Hyland, 1996, p. 447]

(45) **These data indicate** that the major organisational principle of thylakoids is that of continuous unstacking and restacking of sections of the membrane

⁵Copied verbatim from Hyland's work.

Content-oriented		Reader-oriented
Accuracy-oriented	Writer-oriented	
Hedges propositional content	Hedges writer commitment	Hedges assertiveness
Attribute type	Epistemic lexical verbs:	Epistemic lexical verbs:
Precision adverbs	judgemental	judgemental
content disjuncts	evidential	deductive
style disjuncts	Impersonal expressions:	Personal attribution
downtoners	passive voice	Personal reference to
Reliability type	abstract rhetors	methods
Epistemic lexical verbs	“empty” subjects	model
Epistemic modal adjectives	Modal verbs	Assume shared goals
Epistemic modal nouns	Thematic epistemic device	Hypothetical
Content disjunct adverbs	Attribution to literature	conditionals
Limited knowledge	Impersonal reference to	<i>would</i>
	method	Involve reader
	model	direct questions
	experimental conditions	refer to testability

Table 2.4 – Hyland’s summary of hedging functions and main hedging realisation devices [Hyland, 1998, p. 186].

For Hyland a fuzzy-set model is the most appropriate medium of characterizing the indeterminacy caused by hedging. This model enables the identification of core cases of hedging that exhibit salient elements of membership, in contrast to individual cases of hedging, which convey multiple meanings (polypragmatism). Table 2.4 depicts hedging functions and tentative realisation devices in core cases.

2.5 Annotation schemes for hedging

In this section, relevant research on annotation of hedging will be described in a general manner. Specific details about the elements composing such annotation schemes will be later pointed out in Section 2.6. Annotation is often a pre-processing step prior to building sophisticated systems that deal with natural language and need linguistic descriptions in various levels of granularity. For instance, the Penn Treebank Corpus [Marcus et al., 1993] aims to annotate part-of-speech constituents and therefore its descriptions are morpho-syntactic. Annotations of hedging generally fall into lexical, semantic and pragmatic categories of linguistic description.

Similar to annotation of hedging, Herbelot and Copestake [2008] conducted an annotation study of noun phrases that have a generic relation with other constituents in the same sentence. This includes factual and non-specific statements in contrast to hedged statements which describe subjective situations. Nonetheless, their manual annotation scheme is relevant to the present study as their annotation process starts by drawing intuitions out from corpus observations. The annotation labels were adjusted after a certain number of itera-

tions in the annotation process. They point out this kind of annotation is not a trivial task as every noun phrase in a corpus has to be examined to see if it potentially is enclosed in a generic relation. The annotation of generics and specifics is centered around entities, while in annotation of hedging the focus is on the hedging relation between entities. This relation to be annotated is realised by, to some extent, more regular lexical markers than in the annotation of noun entities, therefore there is no advantage in pre-annotating text for annotating genericity. In a sense, this is a simpler task as only noun phrases have to be examined in contrast to hedges that belong to a more varied set of grammatical classes (cf. Section 2.4.3).

One multi-dimensional model for explicitly identifiable certainty was proposed by Rubin et. al. [2010]. This model, developed in the journalistic text domain, was originally intended to be an annotation scheme for manual categorization of hedges and comprises four dimensions, each one with several categories. A diagram taken from the work by Rubin et. al. is shown in Figure 2.2.

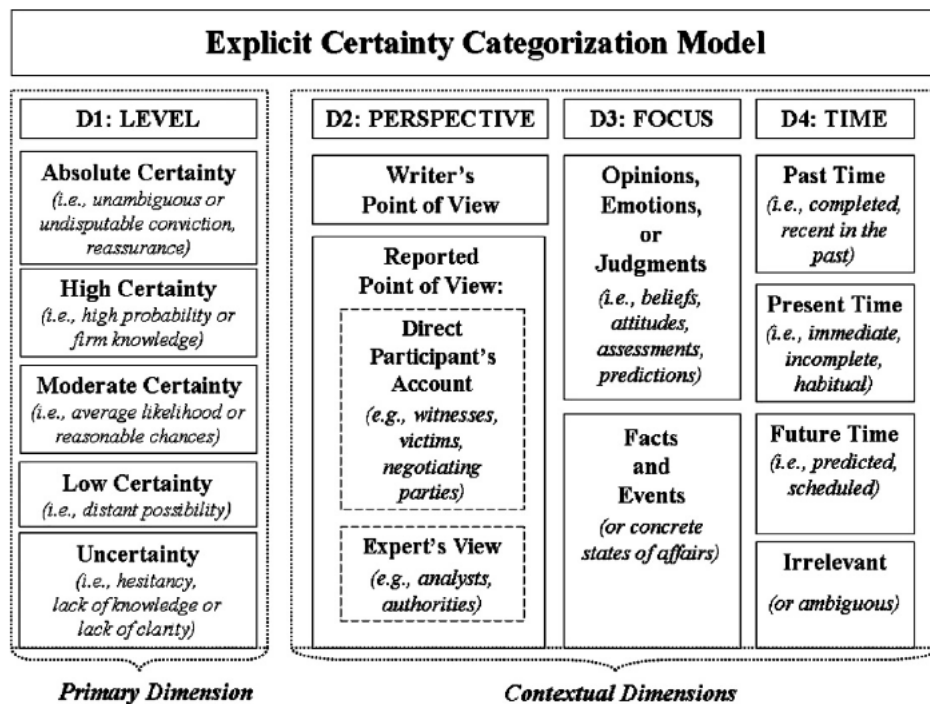


Figure 2.2 – Uncertainty-certainty continuum four dimensional model for journalistic text by Rubin [2010].

This model is an upgrade of a model previously proposed by Rubin et al. [2005b], where the dimension D1 contained only four categories, as the model intended to identify certainty levels, and total lack of certainty (uncertainty) was not considered. This categorization of certainty markers focuses on the classification of degrees of certainty at the sentence level. The perspective dimension is related to the writer and reported points of view (the writer is the author of a report and the reported subcategory encloses the points of view of reported entities that either take part or are subjects of a reported event). The Focus dimension

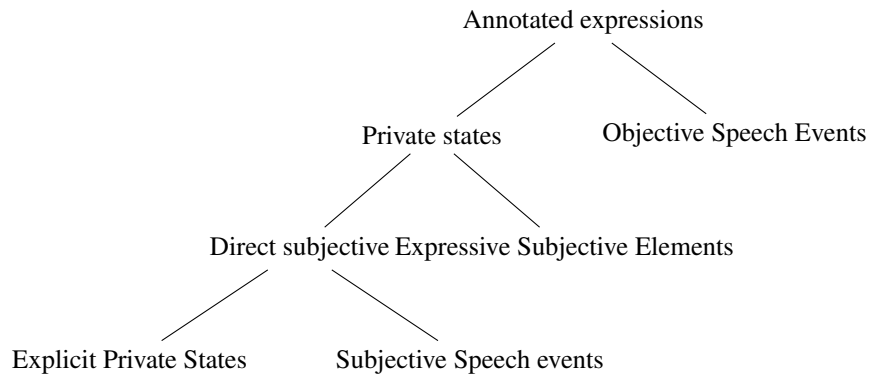


Figure 2.3 – Types of expressions for annotation formulated by Wiebe et al. [2005].

divides the information between abstract and factual categories. Lastly, the Time dimension describes the information according to the time of writing: if time is relevant in the situation being informed, the values assigned in this dimension are chosen from {past, present, and future}.

Wiebe et al. [2005], Wilson and Wiebe [2005] carried out a more complete and complex annotation work of opinions in general than sole annotation of speculations in journalistic text. This annotation was centered around the concept of Private States. They use this term following up the definition by Quirk et al. [1985] who asserts that these states are not open to observation or verification, so it includes opinions, beliefs, thoughts, feelings, emotions, goals, speculations, evaluations, and judgments. Figure 2.3 shows the kind of expressions their annotation scheme was designed to support. Private states are expressed either by Direct Subjective or by Expressive Subjective Elements. The former type of expression are mentions of the private state while in the later, the private state is underlying the expression. For instance, in (46), *fears* is an Explicit mention of a Private State and *said* expresses a Subjective Speech event. Both are Direct subjective expressions that differ from the Expressive Subjective Element ‘full of absurdities’ in in (47). All these private states in annotation are assigned an attribute of Attitude that represent the polarity of the private state and can take any of these values: {positive, negative, speculative, other} [Wilson and Wiebe, 2005]. Expressions for Objective Speech Events are also annotated as the contrast case for Subjective Speech Events, for instance *said* is used objectively in (48) in comparison to how it is used in (46).

- (46) “The U.S. **fears** a spill-over,” said Xirao-Nima.
 (47) “The report is **full of absurdities**,” Xirao-Nima **said**.
 (48) Sargeant O’Leary said the incident took place at 2:00pm.
 [Wiebe et al., 2005, p. 5]⁶

Additionally, the annotation scheme by Wiebe et al. [2005] comprises other elements such as the Source that represent the entity expressing the private state and the Target that

⁶Examples copied verbatim from the work by Wiebe et. al.

Table 2.5 – Summary of counts from Bioscope annotations of hedges and negation words.

	Clinical free-texts	Full articles	Article abstracts
No. of documents	1,954	9	1,273
No. of sentences	6,383	2,624	11,872
% hedge sentences	13.4	22.29	17.69
% negation sentences	6.6	13.76	13.45

Table 2.6 – Summary of inter-annotator agreements for the Bioscope corpus annotation of hedge keywords and the scope of hedges.

		Clinical free-texts	Full articles	Abstracts
Keyword	Avg.	88.78	83.30	85.03
	3 Ann.	(84.01/89.86/92.37)	(79.12/83.92/92.05)	(77.60/81.49/90.81)
Full Scope	Avg.	86.77	72.96	83.61
	3 Ann.	(81.90/82.88/95.54)	(76.72/80.07/94.04)	(62.50/66.72/89.67)

represents the topic of the private state.

Bioscope [Szarvas et al., 2008, Vincze et al., 2008], a Biomedical database of sentences tagged with speculation and negation information was created because of the surge of interest in an automatic and corpus-based statistical solution to the uncertainty and negation detection problems, specifically in medical and biomedical domains. Szarvas et. al. reported these problems are raised from the need to determine whether relations between entities convey uncertainty, whether these relations are factual or there is not real relatedness (e.g. propositional negation).

They annotated sentences which contain uncertain and negative language. As an additional outcome, they produced a set of guidelines and conventions followed during the annotation process they carried out. These guidelines specify the main forms of hedging covered in the process and some technical details such as the annotation format followed while carrying on the task.

They do not reveal further analysis on why interrogative sentences (questions) although “inherently suggest uncertainty” are not marked as uncertain. One of their findings is that over 20% of the sentences contain a modifier (speculative or negative) that in some way changes their semantic content. Table 2.5, created from the counts reported by Szarvas et. al., states the distributions of 3,263 documents from where 20,000 sentences were collected.

Inter-annotator agreements for speculation annotation are summarized in Table 2.6, when comparisons around the hedging keyword or the hedging scope are made. The annotations were carried out by two junior annotation and a chief annotator who produced the gold standard annotations. Therefore, Table 2.6 shows 3 different agreement ratios and the average of these ratios. The first ratio is produced by comparing annotation of the junior annotators and the two other ratios follow from comparing the annotations of each one of these annotators with the chief annotator ones.

Szarvas et al. [2008] state they followed a minimalist strategy for marking keywords

and their scope, that is, a keyword is the minimal unit expressing hedging or negation. The annotation accounts for single and complex keywords, as shown in (50) and (49); keywords between < and > and scope marked by round brackets.

- (49) The picture most (<likely> reflects airways disease).
 (50) Mild bladder wall thickening (<raises the question of> cystitis).
 [Szarvas et al., 2008, p. 40]

As they identified several cases where potentially speculative and negative keywords did not actually imply speculation/negation, this resource provides disambiguated speculation/negation keywords. This is especially compelling because some keywords have different levels of propensity to be speculative depending on the domain. Szarvas et. al. pointed out the example of the keyword *or* which in scientific abstract is labelled as speculative in 11.2% of the cases while in clinical texts, this occurs in 97.86% of the cases.

Bongelli et al. [2012] focus on the annotation of articles on a biomedical domain, they proposed an approach built around the concepts of Certainty/Uncertainty.

Konstantinova et al. [2012] focused on the annotation of speculation, negation and their scope in review articles. Regarding speculation they do not delve into further refinements about categories of speculative markers.

Hendrickx et al. [2012] proposed a scheme for annotation of various types of modality including deontic and epistemic modality for text in Portuguese. The annotated dataset was composed of around 2,000 sentences extracted from the Reference Corpus of Contemporary Portuguese Online (Corpus de Referência do Português Contemporâneo) [Généreux et al., 2012].

Wikipedia text has also been addressed for annotation of uncertainty related phenomena. I describe Wikipedia text in a distinctive manner as the nature of its annotation originated differently from other types of text. Originally, Ganter & Strube [2009] exploited the concept of “weasel word” in Wikipedia articles to detect hedges in sentences. A weasel word, term apparently originated in political discourse,⁷ is “an equivocating or ambiguous word which takes away the force or meaning of the concept being expressed” [Oxford English Dictionary, 2000]. Expressions like *some people say, it is believed, many are of the opinion, most feel, experts declare, research has shown, science says*, etc. are discouraged in Wikipedia edition.⁸

Ganter & Strube considered that these expressions have the same status as hedges in discourse and take advantage of weasel word tags to build a dataset from an automatically extracted dataset from Wikipedia enriched with a small hand annotated dataset.

The case of Wikipedia dataset and Bioscope is particularly interesting because there is not a consistent tagging of the text particle that conveys hedging. For instance, some cases of tagged particles are: *most, most of the people, many* and *many Muslim*; in some of these

⁷ cf. [Watson, 2004]

⁸http://en.wikipedia.org/wiki/Wikipedia:Avoid_weasel_words. Last accessed on 28/06/2011.

cases only the determiner has been tagged as cue and in some others the determiner and a dependent noun are marked as a single speculation cue.

The lack of consistency is an issue to be seriously considered, particularly if the study of the speculation scope is addressed, i.e. determining which part of the sentence is being affected by the speculation phenomenon. This is also important to be considered if portability to other domains is targeted, for instance *many Muslim* is not a speculative expression to be found in every domain in contrast to *many*.

Recently, Vincze [2013] has divided speculative cues in the Wikipedia dataset into three types: weasels, hedges and peacocks. Particularly weasels are cues that signal uncertainty regarding an argument identity. For instance, the uncertainty in ‘some other’ is caused by the lack of specification (51). Hedges here are taken as the regular conception although limited to expressing uncertainty. Peacocks are terms that signals various sorts of subjective judgements such as ‘ardent’ and ‘most distinguished’ in (52).

(51) While the Skyraider is not as iconic as **some other** aircraft, it has been featured in some Vietnam-era films such as *The Green Berets* [...] and *Flight of the Intruder* [...].

(52) Through the **ardent** efforts of Rozsnyai, the Philharmonia Hungarica quickly matured into one of Europe’s **most distinguished** orchestras. [Vincze, 2013, p. 40]

Posteriorly, Farkas et al. [2010] followed Ganter & Strube’s strategy to create a set of Wikipedia sentences where weasels were annotated as uncertainty expressions. However, they do not annotate the scope for these expressions as they claim the main motivation of the marking of weasels by Wikipedia editors is encouraging writers to improve their articles to express factual information.

This section has highlighted the main studies for annotation of hedges or speculative expressions. The main elements comprising the annotation scheme in each study were roughly described. I also pinpointed the domains they addressed, some limitations of their annotation schemes and which kind of hedging expression was targeted. More details on the elements commonly found in annotations schemes for hedges will be described in next section.

2.6 Elements and Attributes intervening in the Annotation of Hedges

This section gives an overview of the main elements that are often found in annotation schemes of hedging. In Section 3.3 from next chapter, I will go into more detailed descriptions of how some annotation elements are related to the state of the art.

2.6.1 Hedging expression

This element is the central entity in any hedging annotation work. It is the minimal annotation unit, therefore all other entities and attributes in an annotation scheme subordinate to

Modal value	Source of the modality
epistemic knowledge	who has the knowledge
epistemic belief	who has the belief
epistemic doubt	who has the doubt
epistemic possibility	who thinks something is possible
epistemic interrogative	who asks the question
...	...

Table 2.7 – Source of modality for various modal values [Hendrickx et al., 2012, p.1807].

it. Its various linguistic realisations were covered in Sections 2.4 and 2.5.

2.6.2 Source or Perspective

The source or perspective of a hedge refers to the entity whose point of view is being expressed in a hedging event. Normally, it is the writer who expresses that point of view, however there is an additional interpretation for the source of hedging: that it corresponds to the experiencer of the hedging event. Therefore, in literature authors have either divided or organized their ideas around these two interpretations.

The term “perspective” is mainly used by Rubin [2010] in her proposed model for annotation of certainty levels (Section 2.5). In her work, perspective is a dimension in the expression of certainty. She divides this dimension into: Writer’s certainty and reported point of view. She points out as well that the writer’s opinion may or may not coincide with the third’s party opinion [Rubin, 2006, p. 36]. However, Rubin is not concerned with identifying the experiencer in the situation where certainty is assessed.

In their proposal for annotation of modality,⁹ Hendrickx et al. [2012, p.1807] use the term “source of the event mention” in contrast to the “source of modality”; the source of modality was coined in this work to make a distinction between the case where the agent or experiencer of modality is one or a group of individuals and the case where the source is who writes expressing this modality. For instance, in their example (53), the source of the event mention is the sentence’s writer, while the source of modality is the noun phrase *Portuguese people*. Since not only the case of epistemic modality is covered, the source of event mention is reported to be usually the speaker or writer and only in 6% of the cases this is textually represented in the sentence. They report that in 70% of the cases the source of modality is present in the text and the other 30% refers to the speaker or writer and therefore the two types of sources refer to the same entity. It is not clear if in the 70% are included cases where the speaker or writer is the source (i.e. it is explicitly present in the sentence).

- (53) Portuguese people need, on average, 180 thousand escudos per month to support a family of four people.

⁹Hendrickx et.al. consider different types of modality being epistemic modality one of them.

Hyland [1998] uses the term “source of the epistemic judgement” , and he distinguishes four kinds of sources: a) explicitly attributed to the writer as in (44), b) evidential (40), c) intertextual (54) or c) non-explicit (55).

- (44) **I believe** that the major organisational principle of thylakoids is that of continuous unstacking and restacking of sections of the membrane
- (40) **The present work indicates** that the aromatic nng to which the carboxyl group is bound is not necessary, provided that a bulky substituent is present.
- (54) **Trifonov (38) has suggested** that the 530 loop ...
- (55) The inhibitor **is thought to be** one of the inducible ...
[Hyland, 1998, p. 45]

Particularly, the intertextual source would match the concept of ‘nested sources’ proposed by Wiebe et.al [2005] in their study of subjectivity expressions seen earlier in this chapter. The idea of nested sources was proposed as they found that sometimes the writer tells about other people’s private states. And such as in (46) the explicit mention of the private state *fears* has as first source *The U.S.* then *Xirao-Nima* as second source and the proposition’s writer as third source.

- (46) “The U.S. **fears** a spill-over,” said Xirao-Nima.

This section has shown that the concept of Source or perspective of hedging and any expression of subjectivity in general has been recognized as an important element in various studies related to modality and subjectivity. In particular, the concept of multiple sources has emerged and makes evident that, at least from a linguistic point of view, the distinction between who is expressing a mental state with who is experiencing that mental state is relevant for the interpretation of this sort of expressions.

2.6.3 Scope

The linguistic scope of the hedging phenomenon in a proposition is the span of text in the sentence that is affected by a hedge or falls under implicit hedging event. For instance, the whole sentence in (56) falls under the scope of *probably*. A single sentence can comprise one or more scopes corresponding to the hedges they are related to, such as in (57). Szarvas et al. [2008] study was the first one to address the annotation of the scope of hedging and their work has inspired thereafter research on the automatic identification of a hedge’s scope (see Section 2.7). They consider that the study of the hedged scope is important because the information that falls under this scope cannot be considered factual.

The concept of a hedge’s scope as such has not fallen into the focus of hedging or epistemic modality studies in linguistics, although the levels of modality described by Portner (see Section 2.4.1) can be related to this, since for instance sub-sentential epistemic modals such as adjective *probable* affect normally a noun, while sentential epistemic modals such

as might affect the whole sentence they are contained within. However, Portner’s focus was on how these levels of discourse influence the semantic interpretation of epistemic modals and not on the scope of epistemic modals per se.

- (56) [_{SCOPE-of-PROBABLY} The chimaeric oncoprotein **probably** affects cell survival rather than cell growth]
- (57) These findings [_{SCOPE-of-MIGHT} **might** be chronic] and [_{SCOPE-of-MAY} **may** represent reactive airways disease].

Source: Szarvas et al. [2008]

Szarvas et al. [2008], in their annotation work of biomedical domain articles, pointed out that the scope can be determined based on syntax depending on the hedge’s grammatical class. Nonetheless, they describe various problems in the manual identification of hedging scope as in some cases the hedge’s ambiguity itself cannot be resolved and this leads to being linked to a dubious scope. They also pointed out the difficulty of identifying the scopes in longer sentences.

The proposal of the 2010 CoNLL Shared Task [Farkas et al., 2010], that aimed at the automatic identification of hedging cues and their scope, promoted more interest in this topic and in the same way as hedge identification, the scope of hedges has received further interest.

Additional insights on the state of the art in the study of hedge’s scope is described in Section 2.7. Also, further discussion is done in Section 3.3.3 for giving account of the annotation of hedges’ scope in informal language style.

2.7 Computational approaches

This section presents automatic computational approaches relevant to the problem of hedging detection in language, particularly when the hedge conveys speculation.

Light et al. [2004] were the first to explore a machine learning approach based on bag-of-words using Support Vector Machine (SVM) classifier for the task of speculation detection. Their work pointed out different expressions of belief used by scientists when there is no certainty about a conclusion: the expression of hypotheses, tentative conclusions and speculations.

They focused on MEDLINE¹⁰ abstracts as their estimation of the proportion of sentences containing speculative fragments in this corpus is about 11%. Four people annotated the set of abstracts, which are distributed into the topics shown in Table 2.8. This table shows the number of annotated sentences and annotators per topic.

¹⁰MEDLINE is the National Library of Medicine’s premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the pre-clinical sciences, available at: <http://www.ncbi.nlm.nih.gov/pubmed/>. Last accessed on 30/11/2011.

The first topic is about gene regulation in molecular biology research, the second one is about research of associated genes to Crohn’s disease, and the third one is about research on turmeric, a spice with analgesic and curative properties.

Table 2.8 – Data about annotated abstracts in Light et al.’s work.

Topic	Abstracts	Sentences	Annotators
Gene regulation	63	547	2
Gene regulation additional	47	344	1
Crohns (last 2 sent.)	100	200	2
Crohns additional(last 2 sent.)	400	800	2
Turmeric	100	738	1
Turmeric additional (last 2 sent.)	400	800	1

Three categories of sentences were defined for annotation: Low Speculative, High Speculative and Definite. The kappa coefficient for inter-annotator agreement results were not outstanding as the amount of data redundantly annotated was small and many items had only one annotator. In this respect, they concluded that although kappa coefficients should be taken as trends because of their values, distinction between speculative and definitive sentences could be made with some reliability. However, it does not seem possible to make a reliable distinction between high and low speculation. Apparently because of this, Light et. al. only referred to the terms “speculative sentence” and “speculative fragment”, and not to any term related to degrees of speculation.

A system implementing an automatic method for detecting speculation was built using the SVM^{light} package [Joachims, 1999] with its default settings. The set of features used for training was composed of vector representations obtained by processing the abstracts with the SMART retrieval system [Zadeh, 1965].

Light et. al. evaluated their system using two datasets: one dataset with only the last 2 sentences of every abstract in the full dataset and a second one composed of articles abstracts where all the sentences were annotated (Gene regulation, Gene regulation additional and Turmeric). At the same time, two baseline systems were proposed to compare the proposed method with: a) a majority-based classifier and b) a system that looks for sentences with sub-strings matching with strings inside the set { *suggest, potential, likely, may, at least, in part, possibl*¹¹, *potential, further investigation, unlikely, putative, insights, point toward, promise, propose* } and mark as speculative in case of a match occurring and as definitive otherwise.

The scores for the three systems together with the baselines scores (Precision (P), Recall(R), Accuracy(A)) for both kinds of dataset (last two sentences and whole dataset) are depicted in Table 2.9.

As the majority method marked every sentence as definitive, it did not receive preci-

¹¹For instance: *possible, possibly*.

Table 2.9 – Performance scores for Light et. al. system.

System	Last2			All		
	P	R	A	P	R	A
SVM	71	39	85	84	39	92
Substring	55	80	87	55	79	95
Majority	-	-	82	-	-	89

sion and recall scores.¹² The SVM system’s precision is good compared to the sub-string baseline, but the recall and accuracy values are lower. Light et. al. concluded that as only words were featured in the training phase, the amount of data was not sufficient to reach better scores. Although these results were still preliminary, this work showed it was possible to categorize sentences as either speculative or definite in an automatic way with little linguistic knowledge.

The first work focusing specifically on the hedge concept and their identification using a machine learning approach is the method developed by Medlock & Briscoe [2007]. They proposed a weakly supervised probabilistic model, using bootstrapping of manually annotated sentences as seed set. Firstly, their system classified sentences by ranking features according to their probability of “hedge-cue-ness”, as specified by the following formula:

$$P(spec|x_k) = \frac{P(x_k|spec) \cdot P(spec)}{\sum_{n=1}^N P(y_n)P(x_k|y_n)}$$

where *spec* is the class containing words that convey speculation, x_k is a feature, y_n is a target class. Secondly, a threshold σ is set to control the precision/recall balance:

$$X_j \rightarrow spec \quad \text{if} \quad P(spec|X_j) > \sigma$$

Additionally, the package SVM^{light} [Joachims, 1999] was used as a reference method. Both classifiers get around 0.76 for break-even-point measure ($bep = \frac{recall}{precision}$).

Subsequently, Morante & Daelemans [2009] addressed the problem of hedges and their scope identification in sentences extracted from the Bioscope corpus by using the IGTRee algorithm implemented in the TiMBL package [Daelemans et al., 2010] and gain ratio for feature weighting. They first focused on identifying hedge cues, where they concluded that the lemma and word of the hedging expression are the most informative features followed by one word to the right and left of the hedge. For scope identification, they build three classifiers to identify each one three labels assigned to every word in a sentence: F to first word in a scope, L to the last one and NONE if the word is neither the first nor the last one in the hedging scope.

For the identification of hedge cues they get the highest scores of F1 for abstracts (84.77% in abstracts) as Table 2.10 shows. Besides, they get better scores of correctness

¹²These measures are based on speculative sentences detection, however, when the sentences were deemed definitive, there are no positive answers of speculative sentences.

Table 2.10 – Results achieved by the hedge cue word identification system by Morante and Daelemans [2009] comparing with results over a pre-processed test set.

Corpus	Pre-proc.	Prec.	Recall	F1	% Correct
Abstracts	No	90.81	79.84	84.77	78.67
	Yes	60.74	94.83	74.05	96.03
Papers	No	75.35	68.18	71.59	69.86
	Yes	56.56	84.03	67.61	88.60
Clinical	No	88.10	27.51	41.92	33.36
	Yes	71.25	52.33	60.34	64.49

Table 2.11 – Results achieved by the hedge scope identification system by Morante and Daelemans [2009]. PCS and PCS-2 are measures for the number of tokens within the scope that are identified correctly.

Corpus	Prec.	Recall	F1	PCS	PCS-2
Abstracts	85.77	72.44	78.54	65.55	66.1
Papers	67.97	53.16	59.66	35.92	42.37
Clinical	68.21	26.49	38.16	26.21	27.44

when the sentences are pre-processed by matching them with the set of Bioscope speculative cues 96.03% for abstracts. They claim that in abstracts, the speculative cue that leads to the highest number of false positives is *or* (83.32% of the cases), which leads to a low recall in other subsets (sentences from clinical reports). The contribution of each hedging cue to the measures of precision and recall varies depending on the subset of articles under evaluation. Table 2.10 also shows results using a pre-processed dataset, however, the system perform worse on this dataset than in the set that was not pre-processed.

To evaluate the performance of their automatic scope identification system, Morante and Daelemans [2009] proposed two methods. One method (PCS) takes into account whether all the words within the scope have been correctly identified. The second method (PCS-2) only assesses the correctness in the identification of nouns and verbs that fall within the speculation scope. Table 2.11 shows their overall results for the identification of hedging scope. Highest values for all the measures are obtained in the subset of abstracts. The results although compelling, show that the system effectiveness on identifying the scope varies across the various data subsets since the choice of words and consequently of hedge cues is heterogeneous between subsets.

Kilicoglu & Bergler [2008] intended to create a more linguistically motivated system and proposed a semi-automatic approach incorporating syntactic and some semantic information such as lexical cues and syntactic patterns that suggest non-speculative contexts (non-hedges). Some of these patterns will be described further in this section.

They used the dataset created by Medlock & Briscoe [2007]. While Kilicoglu & Bergler considered the manually annotated set provided by Medlock & Briscoe useful, they did not find the bootstrapped dataset suitable because “speculative instances overemphasize certain

hedge cues used as seed terms (*suggest, likely*)”.

They proposed a categorization model based on the principal realization devices proposed by Hyland and created a dataset of speculative sentences annotated with such a categorization scheme. This dataset was built by taking the original example sentences given by Hyland (63 hedging cues identified). They developed a number of strategies to augment this set of epistemic terms:

- (A). By exploring lexical relations from WordNet [Fellbaum, 1998]. They obtained 66 additional lexical terms found in the biomedical dataset by browsing on synsets related to the original terms.
- (B). Finding nominalizations of verbs and adjectives from the extended term set in the UMLS SPECIALIST Lexicon [McCray et al., 1994], they obtained 48 additional terms. Nominalizations with predicative function were considered important hedge conveyors after molecular biology corpus analysis. From this analysis 5 lexical terms were identified and expanded using the strategy described in item a.
- (C). Identifying “unhedgers” (factives), terms expressing strong certainty in non-negation scopes such as *know, demonstrate, prove* and *show*.

From these procedures, they obtained a dictionary of 190 features (lexical terms). These features were assigned hedging strength values from 1 to 5 (1 for the lowest and 5 for the highest) based on Hyland [1998] categorizations for core terms; derived terms were assigned weights 1 less than their corresponding original terms. As Hyland also referred to “harmonic” combinations where two or more hedge cues are combined in a sentence, Kilicoglu & Bergler calculated the sentence hedging strength score by accumulating their individual hedging features weights.

One additional strategy used by Kilicoglu & Bergler is the extraction of syntactic patterns of hedging. The sentences were analysed with the statistical Stanford Lexicalized Parser [Klein and Manning, 2003], in order to obtain typed dependency relations. Clausal complement, infinitival clauses and negation relations are identified and their related components are used to create the syntactic patterns. For instance, a sentence containing a finite clausal complement (*ccomp*) with a complementizer (*complm*) <VB> headed by ‘that’ is parsed as:

ccomp(<EPISTEMIC VERB>, <VB>)

complm(<VB>, *that*)

For instance, the syntactic pattern

<EPISTEMIC VERB> *that* (comp) <VB>

was created from these relations.

The performance of Kilicoglu & Bergler’s system was measured against two baseline systems: one with a substring keyword matching approach proposed by Light et al. [2004] with 14 keywords, and a second one with a feature selection approach where the top 15 keywords (features) selected by using learning and classifier models based on a function $P(spec|x_j)$ [Medlock and Briscoe, 2007]. These baseline systems obtained F1 scores of 0.53 and 0.60 respectively. The best values of accuracy and F1 were obtained with the overall hedging score (accumulated weight values) threshold $t = 3$, are accuracy of 0.93 and F1 of 0.85. The difference between results of baseline and proposed systems with this threshold is statistically significant with $p < 0.01$.

Kilicoglu and Bergler obtained some interesting conclusions through error-analysis of their linguistically motivated approach, such as the influence of negation particles on lexical ‘unhedgers’. For instance, the following sentence is labelled by the system as non-speculative because of the presence of ‘known’ with threshold $t = 0$, unaware of the presence of the negative quantifier ‘little’:

(58) Little was **known** however about the specific role of the roX RNAs during the formation of the DCC.

In their work, they follow a strong linguistic approach, considering at the same time Hyland’s findings. They attempted improvement over other systems by using syntactic and semantic information. However, it needed a good deal of manual intervention for building the classification model. But this is, to my knowledge, the first computational approach that includes a weighting mechanism for speculation classification.

Recent work on computational approaches to hedge detection and the availability of corpora, such as Bioscope and Wikipedia- derived datasets [Ganter and Strube, 2009], inspired and made it possible for the research community to undertake the CoNLL-2010 Shared Task [Farkas et al., 2010]: “Learning to detect hedges and their scope in natural language text”. It aimed at building automatic systems for speculation detection. The aforementioned corpora were proposed in order to test the automatic systems in two different domains: for biomedical domain, Bioscope and for a more general domain, a Wikipedia-derived dataset.

In this challenge, two tasks were proposed: a) Determine if a sentence contains uncertain information; and b) if this is found to be speculative, detect the uncertainty boundaries in-sentence. In the first task multiple evaluations were proposed:

- Closed: Performed only on the datasets provided by the CoNLL Task. the training and evaluation stages has to be done separately in each domain.
- Cross-domain: Training and testing is allowed in a cross-domain fashion, e.g. training in biomedical dataset and testing on Wikipedia dataset and vice versa, or a union of both datasets for training.
- Open: additional resources different to those provided are allowed in the training and evaluation stages.

Many of the systems compared their own results with a baseline system where a sentence containing a hedge cue was classified as speculative and as non-speculative if that sentence did not contain any hedge keyword. Overall best results are shown in Table 2.12. These results are from the closed (C) domain experiments.

Table 2.12 – Best results in CoNLL 2010 Shared Tasks.

Evaluation domain	P	R	F	Type
Task1 Wikipedia	72.0	51.7	60.72	C
Task1 Biomedical	85.0	87.7	86.4	C
Task2	59.6	55.2	57.3	C

The best results for the Biomedical domain were reached by the system proposed by Tang et. al. [2010]. They covered both subtasks, proposing a layered approach of Conditional Random Fields (CRF) [Lafferty et al., 2001] classifiers. The hedge detection module is built in two layers of classifiers, where the predictions from a CRF system and large margin based-system layer are used as input for another CRF system layer. The CRF classifiers are implemented using the CRF++ tool¹³ and the large margin classifier using SVM^{hmm} [Joachims et al., 2009].

Tang et. al. used lexical-syntactic features in both layers inspired by previous work done on Name Entity Recognition (NER) in the biological domain [Tsai et al., 2005] and [Sun et al., 2007]. The word shape of the lemma feature is quite distinctive in this approach: this feature attempts to be a unique representation of different shapes a word can take following the intuition that words belonging to the same category may look similar (e.g. IL-4 and IL-5 in the name entities context). This unique representation is obtained by applying a normalization method where the capitalised characters are represented by ‘X’, non-capitalised characters by ‘x’, digits by ‘0’, and other characters by ‘_’. For instance, the entity ‘Kappa-B’ is normalised as ‘Xxxxx_X’ and further normalised as ‘Xx_X’. This shape normalization method looks interesting especially in the biomedical domain or any domain that includes a considerable ratio of named entities.

Other particular features are the prefix and suffix of keyword tokens, useful as well for characterising named entities. Furthermore, a context feature is included, the context chunk feature in this approach is represented by a BIOS chunk tag together with the position of the chunk in relation to the keyword. The BIOS tag points out tokens at the beginning (B), inside (I), and outside (O) of a chunk. S indicates a token representing a chunk. The BIO tagging produced by the Gennia Tagger [Tsuruoka et al., 2005] is used to form the BIOS chunks.

Though the cascaded method showed the best results during the training stage for the biomedical dataset, the individual CRF component produced the best results in the hedge detection task, reaching a F-measure of 86.79%. CRF reached a F-measure of 50.54% surpassed by the results of cascade and large margin methods, 55.05% in the Wikipedia

¹³<http://crfpp.sourceforge.net/>. Last accessed on 25/08/2011.

dataset. These later results are particularly low due to low recall (35.77% for CRF and 41.66% for cascade and large margin methods). The results produced using a cross-domain dataset (Biomedical plus wikipedia) showed worse F1 than the in-domain results.

Tang et. al. suggest the addition of some complex features such as context free grammar may be useful for detecting the scope of hedges in a sentence.

Georgescul [2010] addressed the task by using a SVM classifier with Gaussian Radial Basis Kernel function (RBF) over a bag-of-features approach. A preliminary experiment with a bag-of-words approach showed that in the Wikipedia domain the SVM system performed worse than the baseline system described earlier in this section. In order to overcome this problem and adjust SVM parameters to get a better performance, she followed a 10-fold cross-validation strategy. By performing a grid search she tested the system with different values for the RBF Kernel parameters, width (gamma) and regularization parameter C .

One interesting characteristic of this system is that they considered the unbalanced categories issue, considering that 18% of the training sentences on the biomedical domain belong to the ‘speculative’ category, and attributed different weights to both classes in both domains:

	Biomedical	Wikipedia
uncertain	0.8198	0.7764
certain	0.1801	0.2235

Besides the bag-of-words, they used hedge keyword frequencies in each sentence, bigrams and trigrams of the hedge keyword as inputs. The best score obtained by this system was F-score of 60.17% on the Wikipedia dataset and 78.5% in the biomedical domain. After the evaluation dataset was released, they carried out additional tuning, getting an F-score of 61.91%. However, in the Wikipedia domain these values are still low compared to the 59% F-score of the baseline system. In order to narrow this gap, Georgescul suggest that the addition of lexical information about hedge tokens could be used to find new potential speculative keywords. Lexical information such as synonyms of verbs, adjectives and adverbs conveying speculations would be extracted from an ontology. This approach was quite simple. Since it only uses simple features, the good performance of the system relies on the parameter tuning stage.

The only system participating in the open evaluation category is the one proposed by Kilicoglu & Bergler [2010] based on their previous work on detecting speculative language in biomedical domain (cf earlier in this section). In the latter approach they introduced some modifications to adjust their method to CONLL domains and data. Some speculative word categories in their lexicon (based on Hyland’s [1998]) were eliminated as these kinds of words were not annotated in the CoNLL datasets as hedges. For instance, approximative verbs such as ‘generally’, ‘largely’ and ‘partially’ were eliminated for the biomedical domain and verbs and nouns related to tendencies such as ‘tend’ and ‘inclination’ were dropped for the Wikipedia dataset. Another amendment was the inclusion of an additional Wikipedia category, vagueness quantifiers. This category includes words such as ‘some’,

‘several’, ‘many’ and ‘various’. Finally some adjustments to weight assignments were made in view of differences between datasets from Kilicoglu & Bergler and CoNLL tasks.

It follows that current computational approaches of hedging detection deal with this problem as one of word disambiguation (cf. Section 2.7), however Lyons [1977, p.791] asserted that the explanation of the ambiguity of words like *must* or *may* is not that they have various meanings. This suggests further linguistic study of hedges towards improving automatic identification systems is needed.

In addition to methods described in this section, Morante and Sporleder [2012] provide a comprehensive description of computational linguistic resources and methods for the treatment of modality in general and hedging in particular. Apart from modality in general, they also address the phenomenon of negation in interaction with modality and further how both phenomena interact with others such as mood and tense.

2.8 Conclusions

In this chapter, I provided an overview of the main concepts related to hedging and the main contributions to the characterisation, annotation and automatic processing of hedges. This overview presented various nuances in the definition of a hedge since this research was started considering that hedges would strictly and straightforwardly be used to express speculation coming from the writer using a hedge. However, this review of the state of the art has shown various definitions and interpretations of this phenomenon, and that hedges are still a research topic that is increasingly being studied.

Lakoff thought of hedges as linguistic devices to define criteria for membership in definitional categories of concepts. Some linguistic hedges originally presented by Lakoff, such as: *a true*, *very* or *extremely* are no longer considered hedges. Subsequent studies have taken two conceptions: one where hedges are used to modify the commitment expressed in a proposition, and other one where they are used to express some degree of uncertainty.

There are many dimensions of hedge categorizations: level of discourse scope, lexical syntactic realisations, semantics, and according to their pragmatic functions. Particularly the distinction between sentential and sub-sentential modality allows to categorize hedges in a different dimension that transcends vertically through lexical and grammatical categories. For instance, when an epistemic adjective such as *unlikely* that is normally associated with sub-sentential modality can come to affect epistemic modality at a sentential level as in *It is unlikely a meteorite will hit this town*. These distinctions in levels of discourse scope are relevant for the interpretation of hedges, particularly in the identification of the scope of hedging, a subtask in the manual annotation to be described in Chapter 3.

There are four main studies in the annotation and characterisation of hedges that have influenced subsequent research in this area. Hyland proposed a pragmatic model of hedges organized around reader-oriented and content-oriented hedges. Reader-oriented hedges have the goal of ensuring the acceptance of statements by the reader. The goals of content-oriented hedges are centered around two components of the content: the information being

presented, and the writer. Hedges commenting on the content have the purpose of ensuring accuracy of the information conveyed and expressing the degree of information reliability. Hedges that are writer-oriented are used to prevent criticism towards the writer. These hedging categories are important because they take into account the context of interaction between the reader, the content and the writer in order to have a better interpretation of how hedges are used in that context.

Rubin proposed a multidimensional annotation scheme for expressions that are categorised into various levels of certainty. The scheme comprises other contextual dimensions involved in the expression of certainty assessments such as the perspective or point of view being expressed, the focus or kind of subjective expression, and the time at which an event under assessment takes place.

Wiebe et. al. developed a more complex annotation scheme for subjectivity expressions including speculation and opinions. One of their main contributions is the concept of nested sources in subjective expressions, that is equivalent to the perspective dimension proposed by Rubin.

The Bioscope corpus, for which annotation was carried over texts from biomedical articles, addressed speculative expressions and negative polarity relations using a minimalist hedge categorization scheme compared with the one proposed by Rubin and Wiebe. This corpus has subsequently been widely explored in automatic identification of hedges.

Throughout this chapter, it is shown that the emphasis in empirical research of hedging expressions has been put into formal academic language style. Particularly in computational linguistics, most automatic methods for hedging identification analyse datasets from this style. Nonetheless, some of the main limitations emerged from research on automatic identification of hedges are related to: the quality and quantity of annotated text needed, the ambiguity of hedges, and the difficulty of porting some automatic method to slightly different domains which suggest that porting these methods to be applied to substantially different datasets might be even more challenging.

Although my research is mainly based on a corpus analysis approach, information about computational approaches was included as a means to give a sense of the requirements, strategies and limitations of Natural Language Processing systems dealing with hedging detection. As mentioned in the introductory chapter of this dissertation, one of the questions I intend to answer is which automatic methods would be most suitable for the detection of hedges in informal language style.

Chapter 3

Building an Annotation Scheme for Hedging

3.1 Introduction

In this chapter, I propose an annotation scheme for describing hedging expressions that occur in informal language domains. The fact that this style is markedly different from more formal language styles encompasses some issues which motivate creating an annotation scheme to be applied to informal language style. These issues and my motivation will be further explained in Section 3.2.

This distinction between dealing with hedges in informal and formal language style is one of the main aspects that influenced my research.

Requirements of the annotation scheme directly determined the choices in annotation elements; they will be described in Section 3.3. Previously, in Section 2.6, elements considered for the annotation of hedging were described in this dissertation according to how they were defined by different sources in the literature of hedging. Further details about these elements and the implementation of their annotation will be provided in Section 3.3. Elements of the proposed annotation scheme in this chapter retain some of these previously introduced concepts and naming conventions. I will often bring out discussions extracted from literature on hedging annotation and studies around epistemic modality to provide a rationale for my choices in annotation elements. In some cases, I will delve into deeper details of studies mentioned already in Section 2.6. In other cases, I will cite new sources as they provide more specific evidence that is relevant to the aspect of the annotation scheme being discussed. For instance, I show specific examples of lexical items used in studies already mentioned. I have to stress that all of the linguistic examples extracted from the dataset under study are pasted verbatim.¹ I will use terms such as ‘potential hedges’ or ‘potential speculation markers’ to refer to ambiguous expressions whose intention has not been determined yet to be either speculative or not. In contrast ‘actual marker’, ‘authentic

¹Linguistic examples from the web forum dataset are marked by a number preceded by the label **Post**, e.g. Post: 343.

hedges’, ‘authentic hedging expressions’, ‘hedging expressions’ or simply ‘hedges’ will be used when their speculative nature has been determined to be such.

Since creating an annotation scheme and studying the corpus was an iterative process, some initial and intermediate scheme versions were produced. However, only relevant annotation choices are described in this chapter. A preliminary corpus linguistic study was performed over a small dataset of forum posts to establish an outline of the resources necessary for annotation. A foreseen consequence of this iterative process is that fine-adjusting the annotation scheme is a task that could be continued indefinitely as every new document to be annotated brings out a new particular case that could be considered as a reason for extending the annotation scheme. Considerations about the steps and resources used in the annotation work of hedges will be described in Section 3.4. Annotation strategies to make the process more efficient and produce more congruent annotations will be described in Section 3.4.4. Finally, I will provide my conclusions regarding the annotation of hedges in Section 3.5.

3.2 Need for a hedging annotation scheme

For the purposes of this research, an annotation scheme of a linguistic phenomenon in text is defined as a formal encoding of the possible text chunks to be labelled, labels to be assigned and relations between chunks and properties attributed to chunks or relations, altogether with guidelines stating which kind of entities and relations should be included and excluded from manual annotation. The need to have well defined chunks of labelled text comes from constraints created by the computational task of automatically tagging linguistic expressions where hedging occurs. Since hedging or speculation expressions are likely to present ambiguous meanings, the need for reliable information about the hedging properties in text require manual annotation work. Annotation of hedges and computational approaches described in Section 2.5 comprise more or less complex annotation structures, as complexity increases proportionally to the number of entities intervening in the hedging phenomenon that are targeted for annotation.

Many of these hedging annotation works do not rely on complex annotation procedures as they are concerned only with one dimension of the hedging phenomenon: whether it is a text span of speculative nature or not. Particularly, previous attempts at speculation automatic detection such as [Light et al., 2004], [Medlock and Briscoe, 2007] and [Szarvas et al., 2008] (see Section 2.7) have focused on speculation in academic articles; however, they do not delve into the study of the Source of speculation and models for its identification. As shown in Section 2.6.2, the writer is not always the Source of the speculation and it is very common that uncertainty expressed in a proposition is the product of reporting other’s point of view. In sentences such as (59) and (60) taken from the Bioscope corpora [Szarvas et al., 2008], the hedges *suggested* and *suggest* express uncertainty experienced by the articles’ authors and both are marked as speculative since they do not reveal strong commitment. In contrast, the same lexical items used in (61) and (62) show the uncertainty

is originated in the cited source of reference’s point view i.e. the fact that sea urchin sequences represent remnants of transposable elements was not suggested by the sentence’s author but by whatever reference is indicated by [37]. Similarly, in (62), the existence of such an independent mechanism in mammals has been suggested by a secondary source not mentioned in the sentence, not by the author.

- (59) This **suggested** that there is insufficient data currently available to determine a reliable ratio for human.
- (60) These results **suggest** that SCOPE’s learning rule is highly effective, though it **may** certainly be improved further
- (61) Based on the presence of stop codons disrupting some of the RAG1-like sequences, it has been **suggested** [37] that the sea urchin sequences represent remnants of transposable elements.
- (62) The existence of such an independent mechanism has also been **suggested** in mammals.

Since the purpose of hedging annotation in these sentences is to recognize which ones present non-factual information, annotation disregarding the speculation experiencer is still valid according to this purpose: No matter who the experiencer is, the hedging phenomenon prevails. Therefore, the focus of methods for speculation identification around this kind of hedging expression is appropriate. This approach can be thought of being ‘content-centered’ as the interest on studying linguistic phenomena and proposing methods for automatization originate in an interest of studying pieces of language (e.g a sentence, a document, or a web forum post) as the main dimension to be taken into account. Any other property related to that fragment of language such as the writer, the time it was written, etc., is secondary in importance. The study of these secondary properties is treated as for supporting the main goal of finding insights about content in the appointed fragment of language.

A different approach emerges from a scenario where identifying who has a certain kind of knowledge is relevant. Potentially, identifying who is the experiencer of the hedging event could aid the building of the statements ‘Individual A knows X and has certainty about it’, ‘Individual B does not know whether X’ or ‘Individual C has not certainty about X’ and the like. This approach can be called ‘user-centered’ as we want to explore the qualities and properties of a particular writer’s utterances; in the same way, exploring any other properties related to the user revolves around the individual as an entity.

In the user-centered approach the primary interest lies in finding out whether a hedging expression reflects the writer’s perspective or not. If the hedging experiencer is explicit in the sentence, this is lexically translated into the usage of first person genitive and possessive expressions such as *I am not sure*, *My opinion*, *IMO* or *to me, it looks like*. In the web forum development dataset², sentences containing this kind of expression convey the forum post

²The development dataset is used for preliminary study of hedges and will be explained in detail in Section 3.4.3.

writer's direct involvement in the hedging phenomenon: in sentences (63) and (64) the use of *I* as the subject of the phrases *I am not sure* and *I'd suggest*; in sentence (65), *I* embedded in the acronym IMO (In My Opinion) and the pronoun *me* in the phrase *To me, it looks like* in (66).

- (63) **I am not sure** which SP is on here, or how to check. Post: 18706
- (64) **I'd suggest** the following additional steps: Post: 3655
- (65) **IMO** it is best to always leave tamper protection on to prevent threats ... Post: 16687
- (66) **To me, it looks like** the O/P wants to try out the 2011 beta for testing ... Post: 15134

At first sight, these hedging expressions are somewhat different than the ones found in other domains such as the news domain explored by Rubin [2006]³ in that normally they do not include phrasal expressions, acronyms or very informal expressions. Although some lexical items such as *don't understand* and *can't know* in Rubin's lexicon are similar to the ones found in the forum dataset, hedging expressions such as *not sure*, *IMO*, *AFAIK*, *dunno* are unlikely to occur in other domains where the register is more formal such as in research articles. Therefore, it makes sense to attempt the analysis of phrasal and other informal expressions as a separate case study in a corpus of informal language style in nature.

In the linguistics and computational linguistics literature, researchers differ in the conceptualisation of hedging markers in terms of granularity. Some examples correspond to the approach by Light et. al. [2004] who worked with a reduced set of hedging markers: {*suggest, potential, likely, may, at least, in part, possibl, potential, further investigation, unlikely, putative, insights, point toward, promise, propose*}, in contrast to the method proposed by Medlock & Briscoe [2007] and Bioscope [Szarvas et al., 2008], that consider up to 264 different hedging expressions of heterogeneous forms.⁴ Some of these forms include definite articles (*the* accompanied by a generic noun) and multi-word expressions (*can not exclude the possibility, can not rule out the possibility, raise the hypothesis*) alongside traditional hedging expressions as the ones used by Light et al.. Therefore, consequently with what happens with the definition of hedges, there is not consensus either in automatic methods research of which lexical hedges should or should not be considered as such. In the same way, they do not include more informal expressions of hedging, obviously because they do not conform to the kind of domain they are targeting.

As noted earlier in this section, most annotation of hedging endeavors and automatic speculation detection methods are appropriate in the same way that mainstream machine translation research is concerned with the content. The difference between content-centered and user-centered studies in linguistic speculation affects the way linguistic expressions are handled. My research aims to provide insights that contribute in both content and user-centered studies of hedging in computational linguistics. Therefore, the building of an annotation scheme and subsequent analysis take into account this perspective.

³Lexicons used in Rubin [2006] and works on hedging analysis can be found in Appendix C.

⁴These 264 forms correspond to those found in the Bioscope training set created for the CoNLL 2010 Shared task.

The designed annotation scheme considers the following important aspects:

- **Agency:** related to who is the person or entity experiencing the hedging event. That is, who is not making an assertion or committing to a proposition. This entity needs to be easily identifiable if found in the sentence.
- **Scope:** it is important that at least the core constituent of the sentential clause that has its factivity affected by the hedging expression be identified.
- **Extensibility:** it should be possible to extend the use of the annotation scheme to other domains of similar language style, e.g. Web forums covering other discussion topics. Therefore, items being annotated should not be domain-specific.
- **Coverage:** The annotated items should cover various uses of hedging expressions: from speculative to politeness expression.
- **Structure:** Introducing some taxonomy of hedging expressions. The aim is not only the identification of hedging expressions but also providing some categorization according to the language style being studied.
- **Flexibility:** it should include a way to aid in the detection of authentic hedging expressions without using rich grammatical information such as part-of-speech or syntactic information given the noisy features of text. The scheme should also contemplate there will be multiple entities in one sentence and it should support complex grammatical constructions even if there will not be grammatical annotation.
- **In-sentence scope:** all the elements related to a hedging event should be annotated in the scope of the sentence where the hedging expression is uttered. This can be used to study the potential utilization of a hedging event annotated in this way independently of other sentences in a document. This restriction aims to simplify the annotation process that otherwise would entail resolving anaphoric elements contained in distinct sentences.
- **Discontinuity:** discontinuous entity elements should be annotated making sure that the various components get identified as a single entity.

It has to be noted here that a further refinement on this annotation scheme on including degrees of uncertainty could be done. However, the judgement of uncertainty intensity in newly-crafted categories so created would imply a serious deviation from the main purposes of this research. Nonetheless, this does not mean studying hedges' degrees of uncertainty is not important in the study of linguistic hedging devices, they would instead contribute to a more complete depiction of the semantics of hedging in informal language.

3.3 Annotation Elements

The annotation elements described in this section are the final outcome of an iterative process that will be explained in Section 3.4. Initially, the annotation aimed at annotating two types of entities: Hedges and Non-hedges. Non-hedges would be the alternative label for entities that were potentially deemed as hedges but were not actually used in any hedging sense.

The final scheme for the annotation task was built around three elements: Entities, Relations and Attributes. Entities are used to represent: a) a hedge, b) its source c) its scope, d) non-hedge and e) other discourse markers. Hedges or hedging expressions are divided in four categories: *Single* hedges, *Not-claiming-knowledge (NCK)* epistemic phrases, *Syntactic* hedges and *Other* type of hedges.

So as to provide an overview of the hedging phenomenon building blocks considered in the annotation process, the entity categories *Single* hedges, *Source* and *Scope* will be described first in Sections 3.3.1, 3.3.2 and 3.3.3 respectively. The *Single* hedges category basically comprises the traditional definition of hedges and alongside the description of *Source* and *Scope* categories are needed to understand better the proposal of the *Not-claiming-knowledge* epistemic phrases category, to be described in Section 3.3.4. Subsequently, the other two proposed hedge categories *Syntactic* and *Other* will be described in Section 3.3.5 and 3.3.6 respectively.

Non-hedges as label keep the initial conception as non-actual hedging devices. Usually this label is assigned to entities that were initially pre-annotated as hedges and turned out they did not convey any hedging meaning in the end.

Relations are used to link the hedge entities with their source or scope when they are in the same sentence as the hedge. Attributes are additional information about hedge entities that can be filled in during the annotation process. Both relations and attributes will be better explained in the sections that describe the entities they are related to. The four categories of hedges will be explained in Sections 3.3.1 (*Single* hedges), 3.3.4 (*Non-Claiming-Knowledge* epistemic phrases), 3.3.5 (*Syntactic* hedges) and 3.3.6 (*Other* category of hedges) respectively.

3.3.1 Single hedges

This category of hedges corresponds to the traditional conception of hedges as single words conveying uncertainty, they usually belong to the grammatical categories described in Section 2.4.3 such as modal and lexical verbs expressing epistemic modality.

The initial lexicon of single hedge instances considered for manual annotation was extracted from the work of Rubin [2006] on categorizing levels of certainty. Hedges conveying moderate to complete uncertainty were the ones selected and are shown (Table 3.1) alongside their part-of-speech (POS) labels⁵. This information reveals the variety of grammatical

⁵POS information is given in the Penn Treebank annotation style.

realization of hedges in this category.

Table 3.1 – Words conveying uncertainty extracted from Rubin’s work. POS labels are shown as assigned by Rubin.

Keyword	POS	Keyword	POS
allegation	NN	allegations	NNS
alleged	V	appear	V
appeared	V	appears	V
arguably	RB	attempt	NN
beginning	V	ca n’t know	V
can	MD/V	can not know	V
chance	NN	chances	NNS
claim	V or NN	claimed	V
claiming	V	claims	NNS
coming	V	confused	V
confuses	V	confusion	NN
confusions	NNS	could	MD
cryptic	JJ	did n’t understand	V
do n’t understand	V	divided	V
doubt	NN	doubted	V
doubts	NNS	effort	NN
efforts	NNS	elucidated	V
far less likely	RB RBR RB	fifty-fifty	CD
for the most part	for the RBS NN	generally	RB
had not been decided	V	hope	V
hoped	V	hopes	V
in part	prep NN	in principle	NN
in theory	NN	intend	V
intended	V	intention	NN
intentions	NNS	largely	RB
may	MD	may be	MD V
may never	MD	may not	MD
may or may not	MD	might	MD
might not	MD	mostly	RB
no one seems to know	V	no telling	V
not clear	JJ	not for certain	RB
part	NN	partly	RB
perhaps	RB	plan	NN or V
planning	V	plans	NN or V
possibility	NN	possible	JJ
potentially	RB	pray	V
promise	V or NN	promised	V
promises	V or NN	question	NN or V
questions	NN or V	seem	V

Continued on Next Page...

Table 3.1 – Continued: Words conveying uncertainty extracted from Rubin’s work. POS labels are shown as assigned by Rubin.

Keyword	POS	Keyword	POS
seemed	V	seems	V
skeptical	JJ	some	DT
some degree	DT	somehow	RB
something	NN	sometimes	RB
speculation	NN	suggest	V
suggested	V	suggesting	V
suggests	V	supposed to decide	V
suspect	V	suspected	V
suspects	V	suspicion	NN
technically	RB	tend	V
tended	V	tends	V
tenuous	JJ	think	V
thinks	V	thought	V
tries	V	try	V
trying	V	unaware	JJ
uncertain	JJ	unclear	JJ
undecided	V	unlikely	RB
would	MD/VP		

Some lexical items such as *can not know* and *don’t understand* could overlap with the definition of Not-Claiming-Knowledge epistemic phrases (cf. Section 3.3.4), but they are only labelled as such if they are associated with a first person pronoun in the sentence.

Rubin’s lexicon was chosen because the corpus from where they are extracted has a more informal tone than other lexicons created out of corpora made out of academic research articles that follow strict stylistic conventions such as Bioscope. Wikipedia speculative lexicon was not considered suitable as the annotated keywords include some very ambiguous items such as *the*, *they* or *one* and Wikipedia articles have an encyclopedic style,⁶ similar to academic writing style. Although lexical items in this lexicon are categorised by degrees of uncertainty, this information is not used in the research described in the current dissertation.

Some of the items in Table 3.1 are certainly not single words, yet they do not include any pronoun that can indicate the source of an epistemic modality mental state.

Deciding on which Single hedge types would be annotated was an issue in the case of expressions such as *would likely*, *would suggest*, or *looks like* which have two hedging types are common in the dataset. There were two possible annotation strategies: a) annotating both types as a single hedge, or b) annotating both types individually so these expressions

⁶Encyclopedic style has similar features to academic writing style in that it needs to cite reliable secondary sources and it should neither contain personal points of view nor first-hand findings. A formal tone and avoidance of first or second person perspective style of writing is advised. [Wikipedia, 2014].

would render two hedge occurrences. The decision to choose either one or the other kind of annotation follows the method used by Kennedy [1987a]; the description of this method will be provided in Section 3.4.4 as it was applied not only to the annotation of Single hedges but in all the categories of hedges in general proposed in this research.

3.3.2 Source

Considering the distinctions observed in Section 2.6.2 regarding the Source or Perspective in epistemic modality, I have defined two categories of Source for a hedging expression: a) **Inner Epistemic Source** and b) **Outer Epistemic Source**. I use these denominations as a metaphor for the visibility of the hedging event: a proposition's writer always represents the outer layer of a reported event as the writer is normally the author of whatever document content. The use of *suggest* and *suggested* in examples (67) and (68) respectively illustrates when the writer matches the hedging Source and when he or she does not.

(67)

USER1: [...] I'd suggest the following additional steps: 1. ... 2. 3.

 Post: 3655

(68) USER2 User1 **suggested** following some steps, and you should consider [...]

The Outer Epistemic Source for a hedging expression in a sentence is always the post's writer. In (68), although it is the writer (USER2) who is uttering an uncertainty expression, USER1 is the one who originated this hedging event by asserting his/her own hedged point of view. In this case, the Inner Epistemic Source is attributed to USER1. In cases where the writer express his/her own point of view in the hedging event, the Outer Epistemic Source coincides with the Inner Epistemic Source.

In this way, the Inner Epistemic Source can take two values: {Writer, Other} in the annotation scheme implementation by setting the attribute `Inner Epistemic Source` that is linked to an entity that corresponds to a hedging expression, whatever its category: Single hedge, NCK epistemic phrase, Syntactic or Other.

Two issues emerge when annotating the Inner Epistemic Source of an epistemic expression:

- A) The Source can be explicit or non-explicit in the sentence. When it is explicit in the sentence, this is marked as an entity by means of the annotation tool. The entity Source is linked to the epistemic expression by the `SourceOf` relation. The case of non-explicit or covert perspective occurs when the Source is not found in the sentence, it can occur either implicitly as in (69) and (70) or as subject ellipsis as in (71).

(69) ... *Trying to install the McAfee software and then uninstalling it would likely just cause more problems due to it and [product_name] conflicting with one*

another. [...]

Post: 100943

(70) [...] *some sort of malware* **might** be preventing you from seeing the stock quotes. Post: 15134

(71) and **don't know** if this is the correct place to ask it so pls lemme know if i shud ask elsewhere on the forum [...]

Post: 114080

Particularly, web forum text is prone to subject ellipsis due to its informal style. Subjects in the form of the *I* pronoun have been reported to be dropped frequently in comparison to third person pronouns *he*, *she*, *they* in text extracted from family conversations and TV dramas [Nariyama, 2004]. This study also claims that the utterances where the subject has been omitted tend to be more dismissive and evasive in comparison to full utterances in e-mail communications. I will not explore the reasons behind subject ellipsis any further, but since the users of a web forum normally willingly participate and search for help, evasion does not look a plausible motive in this kind of domain. Later in Section 4.7, I show a comparison of subject ellipsis with overt subjects in some types of hedging expressions.

- B) It may be difficult to detect when the *Inner Epistemic Source* seems to be other than the writer, but a good criterion to take into account is that the source should be capable of agency and producing subjectivity, not just originating it. Although it is often the case, the grammatical subject does not always correspond to the *Inner Epistemic Source*. In examples (69) and (70), although implicit, the source of the epistemic judgement is the writer. It may be tempting to attribute the Inner Epistemic Source to the clause *Trying to install the McAfee software and then uninstalling it* and to *some sort of malware* respectively. Both phrases are syntactic subjects of the sentences in which they occur, however they are not the experiencer of the hedging event. The definition of modality for each type of modal value is a criterion that helps to decide on these kinds of cases (cf. Table 2.7).

The accurate distinction of the Inner Epistemic Source referring to the writer or not is relevant as experiencing uncertainty⁷ is the expression of the writer's mental state. As described in Section 2.4.3, epistemic verbs are used to express possibility in whatever the writer is trying to assert. Particularly in academic articles, authors often cite other researchers' work using epistemic verbs such as *suggested* in (61) described earlier. Varttala [1999] highlights the objection made by Crompton [1997, p. 283] of calling hedges to lexical verbs used in reporting structures. According to Crompton: "the use of any kind of reporting verb only counts as a hedge if authors have elected to use them to report their own proposition". To this, Varttala agrees to a point by noting

⁷Or any other attitude implied by a hedge, for that case.

the difficulty of identifying the origin of these verbs, and that in some cases such as in statements in passive form, the source of the proposition may be unknown.

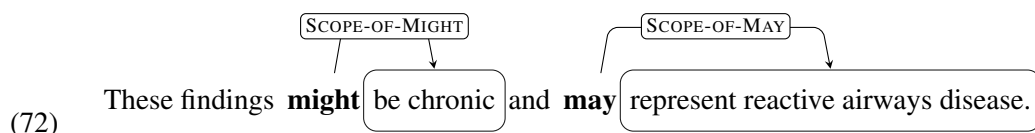
- (61) Based on the presence of stop codons disrupting some of the RAG1-like sequences, it has been **suggested** [37] that the sea urchin sequences represent remnants of transposable elements.

As mentioned earlier in this section, the annotation scheme proposed in the current dissertation provides the means for marking when the Source of a hedging expression is not the writer, even in the case the Inner Epistemic Source is not explicitly occurring in the sentence.

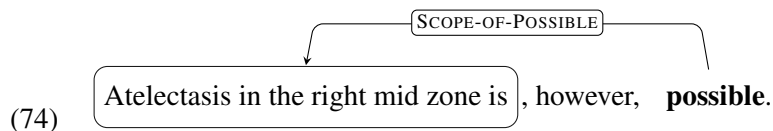
3.3.3 Scope

In this study, the scope of hedging refers to the sentential constituents that are affected by a hedge expression. While a strict identification of the scoped particles is important for an accurate interpretation of the sentence content, in this research the scope identification mainly targets the syntactic dependency head of the phrase or sentence constituent that is affected by the hedge. This means that my approach on annotating the hedging scope is not so strict: at least the syntactic head of the scope has to be annotated and linked to the entity representing the corresponding hedging expression. This is mainly due to the inherent complexity of identifying the boundaries of a particular clause within a sentence. I do not intend to elaborate in a major grammar of scope of hedges here, but I will provide some observations and discussion of singular cases.

I have designed the annotation scheme to have the scope entity separated from the hedging entity and linked to it by a SCOPE-OF relationship as in example (72), where the hedging expressions are in bold font (*might* and *may*), their scopes are enclosed by bubbles and both entities are linked by an edge labelled with SCOPE-OF-[HEDGING_EXPRESSION]. This has the advantage of avoiding the annotation of words which should not be comprehended in the scope boundaries. Particularly, Szarvas et al. [2008] annotated words that actually are not part of the hedging scope because of limitations of the annotation tools they used. For instance, in example (73) shown below, they include in the scope *however*, even though this word is not affected by the hedge *possible*.



(73) [SCOPE-of-POSSIBLE Atelectasis in the right mid zone is, however, **possible**].



In Section 4.4.2, additional observations about hedging scopes as found during the annotation work will be pointed out.

3.3.4 Epistemic phrases

In this section, I explain my rationale for proposing **first person epistemic phrases** as a distinctive category of hedges. Typically, epistemic phrases are grammatical forms of subjects and epistemic verbs introducing complement clauses used to express epistemic modality. I set out to explain how the concept of first person epistemic phrases moves away from the conceptualization of hedges coming from the epistemic modality tradition, which were covered in Section 3.3.1.

Expressions of epistemic modality in domain

Initially, this research was focused on the annotation of speculative expressions disregarding any categorisation that would make the task a complex endeavour. Decisions about the complexity of the annotation scheme were subordinated to the main goal of studying the automatic identification of hedges in an informal language domain.

Preliminary intuitions were drawn after observations of the dataset and specific exploration in the pilot dataset, and later reinforced by findings in the state of the art about phrases conveying epistemic modality. As postulated in Section 3.2, expressions of hedging such as *I am not sure* and *IMO* standing for *In My Opinion* [see (63) and (65)] emphasized the writer's involvement in the proposition. A potential identification of *not sure* as the expression that conveys speculation could roughly lead to think of it as a traditional hedge if deemed equivalent to *unsure*. On these grounds it could be decided that only *not sure* would be annotated. However, making a decision on what to annotate in (65) is less intuitive. Even in the case that *MO* could be identified as the particle conveying speculation, the sense of linguistic unit is lost and in other acronyms such as *AFAIK*. An attempt to isolate a particle resembling a traditional hedge would get artificial if trying to conform to the traditional conceptualization of hedges, noting that this concept as such was originated in the study of text in more formal registers. I reckon these expressions and the like constitute potential units of meaning with distinctive qualities and in this section theoretical support will be provided for proposing them as a new grammaticized category of hedges.

(63) **I am not sure** which SP is on here, or how to check.

(65) **IMO** it is best to always leave tamper protection on to prevent threats ...

Furthermore, I consider some illocutionary force is lost when annotating a speculative expression as a traditional hedge where an epistemic expression could be annotated instead. For instance, if uniquely the epistemic verb *think* in (75)⁸ is annotated, the subjective sense is lost. Modifiers contribute to a subjective reading in this case. A similar case is observed

⁸I am aware that *think* may have a non-speculative reading in this sentence.

in 76), where the source of the speculation is explicit. *look like* set apart is impersonal in comparison to the whole expression *to me, it looks like*. On trying a strict annotation of minimal units expressing speculation, the speculative particles to be annotated in (77) would be reduced to *not ... familiar* which again results in a non-natural annotation contravening the desired feature of flexibility. In (78), *if would* and *suggest* would be annotated separately that would break the compositional subjective meaning of *I would suggest*.

- (75) I for one **think** the best course of action to take when you believe you are infected is to first update your Post: 16687
- (76) To me, it **looks like** the O/P wants to try out the 2011 beta for testing
- (77) I am **not too familiar** with Vista yet.
- (78) I **would suggest** trying to clear out your temp files and also the temp files that you can find ... Post: 242544

As specified in Section 3.2, one of the requirements of the annotation scheme is that agency needs to be identified. While knowing specifically the experiencer of the hedging event is not compulsory, ensuring that the writer is the one experiencing the mental state expressed by a hedge is highly relevant and one of the goals of this research is finding ways the epistemic source of a hedging event can be easily identified. I propose first person singular and plural epistemic phrases as a hedge category when the subject of the epistemic experience is relevant to be identified. I have provided above some observations and issues in hedge's annotation that lead to proposing first person epistemic phrases as a hedging category.

In the following section, I will recourse to preliminary empirical observations and theoretical support that back up this proposal alongside a discussion of evidence presented and issues in proposing this category of hedges. In Chapters 4 and 5, I will provide further empirical evidence to support this hypothesis and explore its accuracy.

Subjective and objective epistemic modality

Early discussion about interpreting epistemic phrases as hedges originated in the different analyses of the phrase *I think*. These analyses discuss the contrast between subjective and objective uses of this phrase. Aijmer [2000, p. 280] states one of the uses is as belief evidential, but at the same time suggests that there is something else (subjective) that contributes to belief formation. The use of the complementizer *that* is considered as a feature that helps to discern between subjective and objective meaning [Simon-Vandenberg, 1998]. Nonetheless, Thompson and Mulac [1991], Aijmer [2000] and Simon-Vandenberg [1998] have agreed on the function of *I think* as a hedge.

Thompson and Mulac [1991] consider this epistemic phrase has achieved a hedging state through a process of grammaticalization. Particularly, they emphasize their first person quality is a means to express the speaker's personal attitude. Thompson and Mulac [1991] particular view is that the epistemic phrase *I think* is a grammaticized version of *I think* with

a *that* complementizer. For them, the subject and verb in *I think* in (79) is used to introduce a clause preceded by *that*, while in (80) and (81) *I think* functions as an epistemic phrase roughly similar to *maybe* in that it is used to express the degree of speaker commitment.

(79) I think that we're definitely moving towards being more technological.

(80) I think 0 exercise is really beneficial, to everybody.

(81) It's just your point of view you know what you like to do in your spare time I think.
[Thompson and Mulac, 1991, p.313]

Although they state that the grammatical status of epistemic phrases is not clear, they argue they may comprise a grammatical sub-category of adverbs. To support this, they show that the epistemic phrase formation process complies the five principles of grammaticization proposed by Hopper [1991].

Moreover, Joanne Scheibman [2001] provides a broad range of research in subjectivity that supports the intuition that epistemic phrases with subject in the first person singular differ at least pragmatically from the epistemic phrases with subjects in third person. Specifically:

“*I* with verbs such as *feel*, *believe* and *suppose* typically express the speaker's attitude with respect to a subsequent piece of discourse or an event in the current context, when this happens with the third person singular subject she or he, there is an impression that what is coming is conveyed is descriptive or informative”.

Joanne Scheibman [2001, p. 69] adds that first person singular is the prototypical site for expression of speaker point of view. In the corpus she studied, first person singular pronoun is the second most frequently occurring subject.

I emphasized the distinction between subjective and objective modality as for analysing epistemic phrases such as *I don't remember*, *I don't understand* or *I don't know* that do not have an explicit epistemic modal constituent conveying uncertainty as *I think* or *I am unsure* have (*think* and *unsure* respectively). Holmes [1988] lists *know* and *not know* as verbs expressing epistemic modality, however, she has a broader conception of epistemic modality as she extends this concept to comprehend expressions of certainty, beyond traditional modals than have been used to discuss epistemic modality in literature.

A point could be made stating that *I don't know* and *I don't understand* are categorical claims of lack of knowledge and lack of understanding. But in their quality as non-factive expressions, their utterance commits the speaker neither to the truth nor falsity of the proposition in which they are embedded. Therefore, I group phrases expressing this lack of commitment with phrases expressing weak commitment as a category of epistemic phrases that do not claim knowledge and as such I named them Not-Claiming-knowledge epistemic phrases.

There is still a distinction to describe between the categorical claim sense of phrases such as *I don't know* and *I don't understand*, and the use of other epistemic phrases such

as *I am unsure*. I intend to give an account of these two phrases by discussing the subjective and objective kinds of epistemic modality.

Considering the subjective/objective epistemic modality distinction made by John Lyons [1977, p.791] where he outlines the difference between (82) and (83) not in terms of possibility and necessity, but in terms of how objective these propositions are depending on whether the situation from where the proposition arises is appropriate or not. For instance, he stresses the case where (83) is asserted in a context where Alfred forms part of a community where 30% of the people are unmarried. Besides, nobody knows who is unmarried and who is not, therefore the possibility of Alfred being unmarried is an objective fact, not just a mere speculation.

(82) Alfred may be unmarried.

(83) Alfred must be unmarried.

Hyland [1998] has not thoroughly addressed the subjective/objective distinction. It may be due to the nature of scientific articles where the utilization of first person normally indicates subjectivity and although sentences such as (84) is clearly considered as objective epistemic modalised following Lyons' approach, is deemed as "explicitly subjective" by Hyland.

(84) I believe that major organisational principle of thylakoids is that of continuous unstacking and restacking of sections of the membranes...

[Hyland, 1998, p. 249]

In example (85), the underlined *I'm not sure* is an epistemic phrase, that does not entail a knowledge claim, the hedge *not sure* is the specific item that changes the certainty of the proposition. The span of text enclosed by square brackets *who's accepted solution to click on since all of you have given me perfect advice that solved my problems* is marked as the scope of the speculation because this is the clause being affected by the hedge. Also, *I* is marked as the source as it is explicitly observed here that is the writer who's point of view is expressed here.

(85) ... I'm not sure [who's accepted solution to click on] since all of you have given me perfect advice that solved my problems.

...

Post: 104689

The term epistemic expression was taken from Kärkkäinen [2010] as a short form of *explicitly personalized epistemic phrases*. This type of phrases is related to the first person singular pronoun *I* and it is reported as the most frequent type occurring in everyday American English [Kärkkäinen, 2003]. In this study, Kärkkäinen explores the concept of epistemic stance as a strategy for expressing knowledge states. She studied a set conversation transcriptions about miscellaneous topics (Part 1 of Santa Barbara Corpus of Spoken

Table 3.2 – Frequencies of most common epistemic markers found in Kärkkäinen [2003, p. 37] study.

Epistemic phrase	Freq.	Epistemic phrase	Freq.
I think	46	I can't believe	8
s/he said	34	looks like/to me	8
I don't know	28	of course	7
maybe	26	sure (adverb)	7
I said	26	I feel like	7
I don't know + compl.	23	seems like/to me	6
I guess	20	I don't think	5
I thought	18	I'm sure	5
probably	17	I figure	5
I'm thinking	11	true (adjective)	5
I remember(ed)	11	I know	5
would	11	s/he goes	5
must	10	I imagine	4
might	9	I know + compl.	4
could	9	I was thinking	4
will	9	should	4
may	8	(not) necessarily	4
apparently	8	definitely	4

American English - SBCSAE [du-, 2000]), some of them task-centered such as a conversation between an attorney and two witnesses in preparation for a trial; and other ones have a more casual nature such as conversations between family members and friends. She concludes that epistemic stance turns out to be highly regular and presents routinized discourse patterns in everyday English conversation in comparison to traditional epistemic markers (cf. Table 3.2). It should be noted that her study explored epistemic stance in general, in this way she considers items used to claim lack of knowledge such as *I think*, *I don't know*, *maybe* and *probably* alongside items used to claim knowledge such as *s/he said*, *I said* and *I know*, *of course* and *necessarily*. An emphasis is made in the analysis of *I think* in an interactional setting. It was shown that *I think* is used as a marker of uncertainty alongside other functions such as a boundary marker for turn-taking in conversation, as a speaker's perspective marker, and as a way to align the speaker's with the listener's stance. Kärkkäinen's study is seminal in the study of epistemic modality following a corpus linguistic methodology and offers a qualitative and quantitative view of how epistemic markers are used in an interactional corpus. Although I make an emphasis on epistemic phrases, my own study is not so ambitious on studying them, my research focuses on the annotation of hedging expressions in general with a view to be used in natural language processing systems.

Wierzbicka [2006] also highlights first person epistemic phrases as an emergent category in modern English used to clarify a speaker's stance in relation to what is being expressed. She lists phrases such as *I think* and *I guess* alongside epistemic phrases such as

I know, that according to Wierzbicka, are used to clarify that the writer is explicitly claiming knowledge in whatever he or she is saying. She suggests that this category of phrases would deserve to be established as a major grammatical and semantic class in modern English. Nonetheless, she acknowledges a rigorous semantic and cultural contextual analysis of each type would be needed to provide an accurate interpretation, which she does in part in Wierzbicka [2006].

Hyland does not consider epistemic phrases having verbs such as *know* as a particular category of lexicalized hedges, but as strategical hedging expressions, e.g. in (86)

- (86) **We do not know whether** the increase in intensity of illumination from 250 to 1000 E/m² per s causes induction of one specific ... [Hyland, 1998, p.142]

Following up the description of a subsequent study made by Kärkkäinen [2010], in this case only epistemic phrases are examined to gather insights about their position in sentence and with relation to their clausal and phrasal scope. In an extended dataset, she found out again that *I think* is the more frequent epistemic phrase, being uttered about every 3 minutes in conversation. Kärkkäinen highlights the difficulty of distinguishing between hedge reading and definitive meaning of epistemic phrases (consider (87) vs. (88)) and even that presence of a *that* complementizer cannot be relied upon to disambiguate to ascertain the correct meaning.

- (87) I believe there is a God. = “I assert the belief there is a God.”

- (88) There is a God, I believe. = “There may be a God.”
Quirk et al. [1985, p. 1113] cited by Kärkkäinen [2010, p. 206]

A preliminary examination in the dataset under study showed first person singular epistemic phrases exhibit similar behaviour to that described by Kärkkäinen [2010] and Wierzbicka [2006]. Epistemic phrases *I think* and *In my opinion* in (89) and (91) respectively are used to hedge the writers’ opinions expressed by the clauses following them. Compared to the bare clauses they scope over or even more to the committed propositions in (90) and (92) there is clearly a sense of uncertainty attached to the expressions. Moreover, looking at the online version of the dataset, *In my opinion* was visually highlighted by the writer as to emphasize his point of view.

- (89) **I think** what you will find is that some program you liked and put on both computers made some changes to the user part of the registry [...] Post: 3655

- (90) This is what you will see - a pretty clear warning that this is serious stuff and a last chance to chicken out. Post: 9574

- (91) **In my opinion** that is **almost** certainly going to touch on actions by individuals and be difficult to stop descending into ad hominem. Post: 2415

- (92) That is certainly the same for me, that the drives are not constantly attached. Post: 27747

Table 3.3 – Classification of epistemic phrases according to their main epistemic constituent.

	Verb centered	Non-verb centered
Lexical epistemic verbs	Primary	Lexically extended
Other epistemic items	Semantically extended	

First person epistemic phrases as a hedging category

The approaches for hedging annotation described in Section 2.5 and computational approaches for hedging detection in Section 2.7 are mostly built around single lexical epistemic items, such as modals, verbs, adverbs, etc. some syntactic categories as the conditional *if* and some relatively more complex expressions (particularly in Wikipedia text). It is to some extent surprising they have not taken into account epistemic phrases as annotation units for hedging detection considering the amount of work that has been done by Kärkkäinen and others in studying the relevance of this kind of expression. On the other hand is not surprising as most of these approaches and resources for hedging detection have addressed more formal domains. Only since relatively recently manual or automatic annotation of hedging has been done in other less formal domains, namely news-wires text [Rubin, 2006] and review articles [Konstantinova et al., 2012].

For the purposes of better describing the kind of items to be annotated, in this case epistemic phrases, I have classified them in three categories shown in Table 3.3: a) primary epistemic phrases, b) semantically extended epistemic phrases, and c) lexically extended epistemic phrases. The annotation scheme does not comprise this taxonomy, however these categories are relevant to give a characterization of epistemic phrases found in an informal domain style. The list of lexical items in each category are shown in Table 3.4.

The Primary type of epistemic phrases are expressions composed of a subject and an epistemic lexical verb conveying speculation as a main verb. This type of epistemic phrases is especially interesting because the main verb can be categorized into the Single-hedges type of hedges earlier described, therefore, an algorithm aiming to detect hedging based on traditional hedges could still identify if there would be a hedging phenomenon.

The Semantically Extended epistemic phrase category is composed of phrases equivalent in meaning to the primary type of epistemic phrases, where the main lexical verb does not necessarily convey uncertainty, but as a whole the phrase conveys uncertainty. To this category belong the objective epistemic phrases mentioned earlier also known as non-factives. For instance, *know* and *understand* are epistemic verbs, but on their own do not convey a sense of uncertainty or used for hedging in general. Their negated coun-

terpart could be deemed as a primary type of epistemic verb conveying uncertainty as in [Holmes, 1988] *not know* is categorized as an epistemic verb expressing epistemic modality. Nonetheless, we can easily see that negating *know* is quite versatile, e.g. *never/seldom/hardly/scarcely know*, *improbable (that) (personal pronoun) know(s)*, *hard to know*, etc. The same versatility can be thought of the case of *understand* and *remember*. Informal contractions such as *dunno* are included in this category.

The lexically extended epistemic phrase category comprehends phrases where the main epistemic component is not a verb, but the epistemicity is attached to another constituent such as a noun or adjective.

Table 3.4 – Extended list of epistemic phrases that convey uncertainty inspired by Karkkainen and Wierbizcka’s research jointly with items collected during the pilot annotation stage. Phrases are divided by the typology in Table 3.3.

Primary type		
I ’d suggest	I ’d think	I ’ve tried
I assume	I believe	I do n’t claim to be an expert
I do n’t claim to know	I do n’t expect	I doubt
I expect	I feel	I for one think
I gather	I guess	I hope
I imagine	I presume	I simply think
I suppose	I suspect	I take it
I think	I thought	I trust
I wonder	I would argue	I would suggest
I would think	little I know	Not that I think
seems to me	seems to us	to me , it looks like
We ’d suggest	We assume	We believe
We do n’t expect	We expect	We feel
We gather	We hope	We imagine
We presume	We suppose	We suspect
We think	We thought	We trust
We wonder	We would suggest	
Semantically extended		
I certainly do n’t know	I certainly do not know	I did n’t know
I do n’t even know	I do n’t know	I do n’t remember
I do n’t understand	I do not know	I do not remember
I dunno	I know little	I know nothing
I really don’t know	neither I know	Not that I know
Not that I understand	We certainly do n’t know	We certainly do not know
We did n’t know	We do n’t even know	We do n’t know
We do n’t understand		
Lexically extended		
AFAIK	dunno	I ’m inclined to think

Continued on Next Page...

Table 3.4 – Continued: Extended list of epistemic phrases that convey uncertainty inspired by Karkkainen and Wierbizcka’s research jointly with collected during the pilot annotation stage. Phrases are divided by the typology in Table 3.3.

Lexically extended		
I ’m not knowledgeable	I ’m not sure	I ’m unsure
I ’ve got no idea	I am inclined	I am not knowledgeable
I am not sure	I am not too familiar	I am unsure
I have no idea	I was n’t sure	I was not sure
IMHO	IMO	In my opinion
In our opinion	maybe I ’m incorrect	My guess is
my point of view	My understanding is	Our guess is
our point of view	Our understanding is	to my knowledge
to the best of my knowledge	We are inclined	We are not sure
We were n’t sure	We were not sure	

Holmes [1988, p.43] points out grammatical patterns for the use of lexical verbs.⁹ For the annotation work, our patterns of first person singular epistemic phrases correspond to the ‘personalized’ patterns in Holmes’ classification. Impersonalized and depersonalized patterns are not annotated, but only the lexical item contained within. In (93), *there is a good chance* corresponds to an impersonalized epistemic phrase proposed by [1988, p.43] that includes the epistemic noun *chance*. As this includes neither an explicit nor implicit subject, only *chance* is annotated as a Single hedge. I do not consider that annotation of impersonalized epistemic phrases would provide a significant benefit to the automatic detection of hedges.

(93) There is a good **chance** that a startup repair will make Win7 boot. Post: 184679

In this Section, I have provided my rationale for proposing Not-claiming-knowledge (NCK) epistemic phrases as a distinctive category of hedges that set them apart from more traditional types of hedges.

3.3.5 Syntactic hedges

I set aside a third category of hedging markers that for the purposes of this research were called ‘Syntactic’ for reasons that further will be evident.

In his study of surface features shown by hedges, Hyland [1998] places conditionals in a category of non-lexical hedges used for strategic purposes in academic texts to refer to ‘limitations of model, theory and method’. Hyland however, is not interested in a deeper analysis of scope of this kind of hedge or the interaction between *if* and main clauses in different types of conditionals when they are used as hedging devices.

Rubin [2006] mentions words in clauses of condition (*if*) and concession (*though*) as potential markers of some level of certainty, but does not provide an extended analysis

⁹Appendix C.2 shows the complete list of these grammatical patterns.

of either type of clauses. Neither do conditionals such as *if* make part of the final list of uncertainty keywords she provided and that were used as initial lexicon for my research.

Konstantinova et al. [2012], in their annotation of reviews, do not place conditionals in a distinctive category, conditionals are annotated as any other speculative expression. They point out *if* as the most frequent speculative word in the SFU Review Corpus (16.34 % of the cases of hedging).

Conditionals were not initially considered to be annotated as part of this research as conditionals have been said to present some issues regarding the scope they cover [Konstantinova et al., 2012]. However, due to their occurring frequency and perception as a particular hedging device in the dataset, they were considered for annotation in the end. Conditionals are, therefore, placed in a specific category of hedges named as Syntactic, I deemed proper to do this because of their frequency, variety of realisations regarding scope and to distinguish them from hedging devices that cannot be grouped together with Single hedges and Not-Claiming-Knowledge epistemic phrases.

In the remainder of this section, different types of conditionals and the situations where they carry a speculative function are described. Examples in domain are provided as empirical evidence for placing conditionals as a Syntactic category of hedges.

Subjunctive conditionals have a clear modal component in the main clause in contrast to indicative conditionals and conditional imperatives. Conditional imperatives belong to a complex category of imperatives that will be described in this section in what regards hedging and how they are used as hedging devices in web forums.

Following Portner [2007], imperatives display diverse meanings such as orders, invitations and suggestions. Although in the domain of web forums under study there is not a clear cut separation between meanings of imperatives, they are situated between invitations and suggestions/advice. Given the nature of the interaction between users (asking advice/-giving advice), even if the conditional imperative could be understood as an order, the reader is not likely to understand that as an order.

Syntactic hedges can play an interrogative role such as *if* (94). In this case, *if* would be annotated as a hedge.

- (94) I'm curious what your system profile is, and **if** there is a potential incompatibility here.

Iatridou [1991] divides conditionals into three major types: relevance, factual and hypothetical conditionals.

Relevance conditionals, also called speech act conditionals, have the function of specifying the conditions in which the main clause is relevant or appropriate given the circumstances, not the conditions in which it is true. Iatridou points out a paraphrase strategy where relevance conditionals cannot be paraphrased as *in any circumstance in which p, q*. Therefore the paraphrase in (95b) of the conditional in (95a) makes it lose its natural reading. Another strategy is suggested by Schwager [2007] that shows how relevance conditional

lose their speech act qualities when *only* is inserted before the *if*-clause or *then* is inserted before the main clause; see (95c) and (95d)

- (95) a. If you wish, you can send me a Private Message with the URL of your website and I'll see what suggestions I can post in this thread. Post: 107272
- b. In any circumstance in which you wish, you can send me a Private Message with [...]
- c. *Only if you wish, you can send me a Private Message with the URL of your website [...]
- d. If you wish, *then you can send me a Private Message with the URL of your website [...]

Other cases are shown in (97), (98) and (99). The writer is clearly not expressing uncertainty in (98) about the interlocutor wanting to ask questions, but the speech act taking place is for making him or her aware that he would answer in case further questions arise. Given the illocutionary act, it could be compared to the hedged *Hope it helps* in terms of being a speech act and a hedge, but I believe this is not the same case as when the writer utters *Hope ...*, since there is still uncertainty underlying. Therefore in this research, relevance conditionals are not marked as hedging occurrences. This decision could contrast with other annotation works, where this kind of conditional is deemed as conveying speculation, namely in (96) extracted from [Konstantinova and de Sousa, 2012, p.11], *if* is annotated as a hedge. I deem *if you prefer* is a speech conditional and therefore *if* is not used as a hedging expression.

- (96) This creative re-engineering draws the viewer or reader into a parallel universe where age-old lessons can be taught re-taught without the obstructions created in the minds or interferences or misconceptions **if you prefer**, or even pre-concepts that may probably lead to misunderstandings. [Konstantinova and de Sousa, 2012, p.11]
- (97) If you're thirsty, there is beer in the fridge. [Iatridou, 1991, p. 50]
- (98) If you have any questions, feel free to ask .
- (99) If you have any better ideas I'm willing to listen . Post: 107272

Nonetheless, I have to acknowledge that the amount of effort involved on deciding about the speech acts qualities of this type of conditional increases the time employed in manual annotation.

Factual conditionals, containing an *if*-clause that is believed to be true, not only specify the conditions in which the main clause is true, but assumes that these conditions exist for somebody. The speaker who utters the conditional does not necessarily have to agree with the proposition expressed in the *if*-clause, however, he or she has to know that it is true for somebody. Again, paraphrasing using *In any circumstance in which p,q* should help

to check if there the conditional is factual, as in Iatridou's example (100). B's utterance cannot be paraphrased as in (100b) as there is not acknowledgement of the if-clause truth. A more accurate paraphrase is given in (100c). Schwager [2007] shows also that modifying the if clause with *only* is not allowed (100d), whereas *then* is acceptable preceding the main clause (100e).

- (100) a. A: Bill is very unhappy here.
 B: If he is so unhappy he should leave.
- b. In any circumstance in which Bill is so unhappy, Bill should leave.
- c. In these circumstances, in which according to somebody's belief Bill is so unhappy, the belief that Bill is unhappy implies (the belief) that Bill should leave.
- d. (*Only) if he is so unhappy he should leave.
- e. If he is so unhappy (then) he should leave.

Observing (101) in an isolated way can render doubts about whether it is a factual conditional or not. Looking at the dialogue circumstances where this conditional is uttered makes it easier to determine the if-clause content truth (103). User1 made a request presupposed by the content of User2's if-clause, consequently uttering (102) would be clearly strange.

- (101) If you want every file to back-up , you must set evey filetype to find in the back-up settings good.
- (102) *Only if you want every file to back-up , you must set evey filetype to find in the back-up settings good.

It has to be remembered that in the annotation task at hand, it is not possible to have access to complete dialogue that takes place starting with a question, comment or announcement. Nonetheless, having access to the full text written where the conditional occurs might enable to determine if this corresponds to the factual type. For instance, the utterance of *[Product-name] is not capable to back-up and restore full partitions* makes possible to determine the addressee has requested information related to the content of USER2's if-clause, moreover there is the presupposition that the addressee will perform the advised action contained in the main clause, evidenced by the utterance of *Let us know the reslts*.

- (103) USER1: Hi. I'm trying to back up my hard drive to an external one. [...]
 Post: 125653
- USER2: [Product-name] is not capable to back-up and restore full partitions
 , only the files from it.
 **If you want every file to back-up , you must set evey filetype to find in
 the back-up settings . [...]**
 Let us know the reslts . Post: 125750

Other examples of factual conditional are shown by (104a) and (105a) where the context needs to be checked to confirm this status. Another way to test for factuality is paraphrasing by preceding the construction by *In the hypothetical case in which*. Since the propositions contained in the if-clauses are already known to be true, the paraphrases in (104b) and (105b) are not sensible.

- (104) a. So if you have [Product-name1] already just purchase the [Product-name2] for \$49.99 for the 3 PC version ! Post: 216510
 b. So *(in the hypothetical case in which) you have [Product-name1] already just purchase the [Product-name2] for \$49.99
- (105) a. And if it has to search for zero files why doesn't it do it when the system is inactive as it claims it will do? Post: 9056
 b. And *(in the hypothetical case in which) it has to search for zero files why doesn't it do it when the system is inactive as it claims it will do?

In Hypothetical conditionals, there is a clear speculative component. Example (106a) can be paraphrased as (106b), the meaning conveyed by both propositions is equivalent, so *if* may be marked as a hedge in this case. Also in hypothetical conditionals both the if- and main clause make part of an assertion, while in a factual conditional as in (104a), it is not the case; only the main clause is part of the assertion there.

- (106) a. For example **if** you buy [product_name] today will be entitled to a free upgrade to the 2009 version , as long as the subscription is active. Post: 11959
 b. For example (in the hypothetical case) you buy [product_name] today will be entitled to a free upgrade to the 2009 version , as long as the subscription is active .

About the scope of conditionals, Kratzer [2012] gives an account of modals quantifying over possible worlds and states that the traditional analysis of conditionals as a two parts structure is a wrong one. She shows how the if-clause acts as a restrictor for the main clause. This approach has been useful to given an account of why the scope of the conditionals as hedging expressions spans over the if- and main clauses. However Kratzer has not mentioned how this applies to main clauses in imperative form.

I have explained in this section my rationale for including conditionals as a category of hedging devices. In the process of annotation other hedging types were attributed to this category when their characteristics make them align to a syntactic construction. The whole set of types comprehended by this category is described in Section 4.2.3.

3.3.6 Other hedges

This category of hedges was created to contain expressions not clearly identified as belonging to either Single-hedge, NCK epistemic phrase or Syntactic categories. This arises from

a difficulty in identifying a potential hedge into one of these categories, as they have mostly well defined structures, either because they come from the traditional conception of epistemic modality (Single hedges) or because they have a clear syntactic-semantic structure (NCK and Syntactic hedges).¹⁰ For instance, *it is hard to tell* in sentence (107) is one of the hedges included in this category.

Due to the requirement of extensibility pointed out in Section 3.2, subsequently during the annotation process, a particular type of NCK epistemic phrases was included in this category. These epistemic phrases are specifically circumscribed to the knowledge domain where forum conversations are carried out (i.e. knowledge of the world) such as the expression *I am a computer illiterate*. This category of hedges resembles the strategic hedges category proposed by Hyland [1998]. One of these hedges annotated under this category is *Just to make sure* in (108), found in the development dataset.¹¹

(107) Without more information **it is hard to tell**. Post: 13182

(108) **Just to make sure**: You created new user accounts on both the computers you were having problems with; and both are now working correctly? Post: 3655

Further details about this type of hedges and a more exhaustive set of examples will be provided in Section 4.2.4 as they were found during the annotation task.

3.3.7 Relations

Defining how the different entities involved in hedging would be annotated was a necessary step in the creation of the annotation scheme since in a single sentence two or more hedging events may occur. Two different hedging events in the same sentence may have a Source and a Scope each one making two different hedging sets, so a mechanism for discriminating both sets is compulsory.

`SourceOf` relation binds the source with the hedging element. `ScopeIs` relation binds the scope entity with the hedge entity.

3.3.8 Summary

This section has described the main elements that compose the annotation scheme for hedging in a domain where informal language register is used. The main elements are entities that represent three elements: the hedging expression, its scope, and its source. At the same time the hedging expression belongs to one of these categories: Single-hedges, Not-Claiming-knowledge epistemic phrases, Syntactic hedges and Other hedges. Additionally, the Inner Epistemic Source is defined as an attribute of the hedging expression since the Source is not always linguistically explicit in the sentence where the hedge is contained. A

¹⁰Nonetheless, in few cases, the difficulty of identifying hedges may be due to limitations of the expertise of the annotator.

¹¹It was eventually not taken into account in profiling hedge occurrences as it is contained within an interrogative sentence.

provided in Section 4.1.

3.4.2 Annotation tools

I chose **brat rapid annotation tool**¹² [Stenetorp et al., 2012], because it can be used to develop annotation using a web graphical interface. Annotations carried on through this GUI are reflected in textual annotations in a stand-off format. Stand-off annotations are produced in parallel in a way the original text is not modified. Another annotation tool that was considered is GATE [Cunningham et al., 2011], however despite being a more advanced and complete tool, the annotation format is obfuscating in comparison to `brat`. GATE bases its annotation in a language graph annotation scheme where annotations are edges between two nodes: beginning and end of annotation are recorded in a XML format. Also, GATE's dynamic of use is more complicated in the case of multiple annotators intervening in an annotation task.

`brat` on the other hand produces simpler annotations and the tool is relatively easy to configure, and if this is installed on a web server it can be accessed by many annotators independently of platform. GATE offers a web environment also (Teamware <http://gate.ac.uk/teamware/>), but as this is a more sophisticated and complex tool, the difficulty of configuration is proportional.

The sentence in (109) was graphically annotated using the `brat` web user interface, which produced the visual annotation shown in Figure 3.2. The parallel stand-off annotation of this example is shown in Figure 3.3. The boundaries of elements in the stand-off format annotations are represented by numerical parameters according to the position of a token in a file being annotated. The names attributed to annotated textual expressions can be chosen graphically according to an encoded annotation scheme as Figure 3.4 shows. This coding reflects the elements in the annotation scheme proposed in this research.

(109) I'm not 100% sure I have the right settings fro the Virtual Memory . Post: 52058

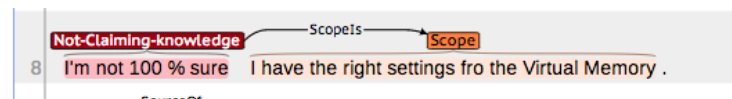


Figure 3.2 – Visual interface of the annotation of example (109).

```
T11    Not-Claiming-knowledge 357 375  I'm not 100 % sure
T12    Scope 376 424    I have the right settings fro the Virtual Memory
R6     ScopeIs Arg1:T11 Arg2:T12
```

Figure 3.3 – Stand-off annotations of the sentence in example (109).

¹²<http://brat.nlplab.org/index.html>

```

[entities]

Scope
Source
Focus
Epistemic-phrase
    Claiming-knowledge
    Not-Claiming-knowledge

[relations]

[events]

Hedging Exp:Focus
Not-Hedging    Exp:Focus

[attributes]
Politeness    Arg:<EVENT>
Inner_Epistemic_Source Arg:<EVENT>, Value:Writer|Other

```

Figure 3.4 – Coding on the annotation scheme for hedging.

One advantage of the stand-off annotation that `brat` uses to store representation of annotations is that is more readable than other formats such as complex annotations stored in XML format. This format is the one used in the BioNLP Shared Tasks.¹³

While `brat` provides a visual interface for manual annotation, its simplicity and flexibility relies in that it is not an analytical tool for corpus linguistics. This means this tool does not automatically produce concordances, or calculates statistics in corpus. Therefore, as pre-annotation and post-annotation procedures were needed in the iterative annotation, I built my own set of tools to carry on these automatic procedures. These procedure comprises conversion of annotations to/from the `brat` stand-off format from/to other representations that could be analysed in terms of occurrence in sentences and in posts overall. For instance when pre-annotation (Section 3.4.3) is carried out, matched strings have to be converted to the stand-off format so manual annotation can be done using `brat` graphical interface.

3.4.3 Annotation procedure

Data collection and pre-processing

The original dataset of posts extracted from the web forum environment contained 308,274 files. From this total number of posts, only 230,570 were considered for processing as the rest contained confidential information not publicly available. A post is encoded as a file composed by a subject and body of textual elements corresponding to the message format they have in the web forum. Additionally, there are metadata related to posts and users in

¹³<http://2011.bionlp-st.org/> Last accessed on 20-03-2012.

the web forum. Section 4.1 describes in more detail the nature of this dataset.

From a technical point of view, they were extracted using a REST API service¹⁴ from the platform used by the web forum,¹⁵ and it was provided as a set of files corresponding to posts containing text mixed with XML and HTML elements which contributed to the amount of textual noise in the content.

In this research, three main datasets are used for analysis, profiling and building of models. These ones are the development dataset, annotation dataset and reduced training dataset (RT dataset or RTD for short).

From the set of 230,570 posts, to create the RTD dataset 172,920 posts were randomly selected following a stratified sampling based on 25 categories of users who contribute to the forum by writing posts. The number of posts across these 25 categories is detailed in Appendix B. The remainder of posts constitute a test set that may be used in future research.

For the first round $\sim 1\%$ of the posts to be annotated were selected (56 posts) to compose a small sample which from now onwards I will call the development dataset. The first round of annotation is described in Section 3.4.3. The annotation dataset of 3,000 files was randomly selected according to this distribution of posts per user category. Further description of the annotation dataset nature in terms of function and related to findings in the annotation stage is given in 4.1.

Figure 3.5 shows procedures the text was submitted to prior to manual annotation. The different procedures are enclosed in two main building blocks: Text Extraction and Pre-processing. The text extraction procedure provides text in a format that can be used for diverse purposes in text analytics. The Cleaning procedure has the purpose of dropping out hyper-textual elements (XML, HTML) and turning significant ones into wildcards that represent them in a textual way as they are potentially important sub-textual element. For instance graphical emoticons such as 😊 and 😞 were replaced by wildcard SSSSSS and the polarity of emotion conveyed by them was recorded as it was deemed emotions and hedges in the same proposition could be explored as to study their interaction.¹⁶ More details about the Cleaning procedure are presented in Appendix E.

The next procedure, Sentence Splitting, was necessary as I wanted to carry on observations at a sentence level. The procedure was performed using an adapted version of the sentence splitter from MSBP tool.¹⁷ However, in many cases, the sentences were not correctly split, specially when noisy text was present in text to be split

The normalisation procedure processed text with the purpose of turning it into a version with fewer unintelligible elements so text analytic techniques could be more successfully applied. Again, because of the heterogeneity of non linguistic elements in the text, the heuristics employed for normalisation did not cover 100% of the cases where it should

¹⁴REST API allows metadata querying through HTTP methods. These metadata are stored in a REST web architectural style.

¹⁵While we are not able to share the corpus, the data is equivalent to what could be obtained by using a web robot to scrape data from the publicly visible content.

¹⁶Further details about emoticon processing and results from profiling them are given in Section 5.5.4.

¹⁷<http://www.clips.ua.ac.be/pages/MSBP>. Last accessed on 02/12/2011.

have proceeded. Details about the normalisation heuristics are explained in Appendix E alongside the cleaning procedure details.

Once the posts were in a more tractable format, they were submitted to the Pre-processing procedures, where the posts were anonymized in text¹⁸, as I wanted to prevent the possibility of associating names to particular language style. Alongside the forum post files, meta-data information about forum posts and users was provided (cf. §1.1.2 for more details), and a list of login names extracted. This list was used to perform a replacement of login names by wildcards, however, in many cases login names matched other kind of named entities in the text. For instance, login names such as : *June*, *Dlink*, *Robert*, *Netherlands*, *Smackdown* could be found in the text not referring to users such as in (110). Because of this, the login name list was manually reduced to entries that are not likely to cause mismatch.

- (110) a. I have not seen a solution or a new post since mid *June*.
- b. ... of other machines to backup to an *Dlink* DNS323 or some other Network.
- c. ... said *Robert* L. Carothers, president of the University of Rhode.
- d. ... sites here in the *Netherlands* ...
- e. The very last *Smackdown* game where ...

The tokenization procedure is used to provide standard forms for contractions and separate punctuation from words, that will allow to perform the keyword matching in a more efficient way. Some measures were taken to preserve the originality of text, for instance contractions such as *I'm*, *shouldn't*, *I've* were tokenized to *I 'm*, *should n't* and *I 've* correspondingly and not to their normal forms *I am*, *should not* and *I have*.

The matching of keywords was performed with a modified version of the Aho-Corasick string matching algorithm [Aho and Corasick, 1975]¹⁹

For annotation with brat the files need to have *.txt extension and annotations are stored in a file with the same name and *.ann extension. Depending on the nature of annotation there are two options:

a) Annotating raw data, no pre-annotation of speculation markers, hedges or negation particles is done. In this case the annotator has to annotate data from scratch only relying on guidelines. So the full text of every document (post) is flushed into a *.txt file and create a parallel *.ann empty file.

b) Annotating text with pre-annotations. Some data preprocessing would need to be done to convert current XML mark-ups into brat stand-off format (<http://brat.nlplab.org/standoff.html>).

¹⁸Posts are frequently ended by a signature where the author writes his login name down.

¹⁹The original implementation for this modified version can be found in <http://search.cpan.org/~vbar/Algorithm-AhoCorasick-0.03/lib/Algorithm/AhoCorasick.pm> Last accessed: 12/12/2012

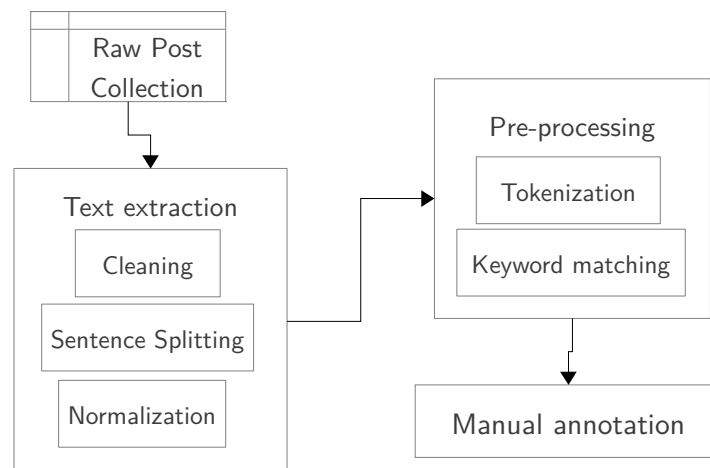


Figure 3.5 – Pre-processing pipeline.

Pilot study (first round)

The pilot study has the purpose of empirically building and refining an annotation scheme and carrying out the initial steps of a corpus study.

The corpus study allows the collection of insights about the kind of hedging expressions plausible of being found in this dataset, it exposes some difficult cases and potential problems of linguistic or technical nature that could be faced in the annotation process.

Lastly, this annotation round allows the definition of boundaries and limitations of the annotation task and consequently of a potential assessment of certainty in texts from this dataset.

Insights from the pilot annotation study are summarized in Chapter 4.

Pre-annotation

Pre-annotation towards a training dataset annotation of hedging expressions was done based on a lexicon collected from various sources. The single hedges lexicon was mainly extracted from [Rubin, 2006], they are words conveying at least some degree of uncertainty.

The lexicon of epistemic phrases was taken initially from [Kärkkäinen, 2010] and [Wierzbicka, 2006]. This lexicon was then expanded with similar plausible epistemic phrases and the ones found in the pilot annotation phase.

As was observed earlier, the annotation tool has its own format for showing marked entities, the pre-annotations were automatically generated following the stand-off format conventions.

Manual annotation

For this round, a stratified random sampling procedure was employed to select those posts to be annotated. Also, care was taken to avoid selecting posts of users whose posts were explored in the pilot annotation phase, which restricted the pool of posts to be selected. Based

on the amount of text and document selected for annotation in other similar projects, 3,000 posts were selected for manual annotation in a way that reflects the main user categories and that does not conflict with the training and testing subsets created out of the complete dataset.²⁰

Interrogative and quotative statements and non linguistic content in posts were disregarded for annotation. As not all interrogative sentences were marked with an interrogation symbol, a manual checking was performed. Quotative statements that could not be identified by pattern matching methods were also checked manually. These statements were not dropped from the documents, but the sentences corresponding to them were skipped when annotations were collected automatically.

There is an obvious limitation to the findings reported on here due to the fact that manual annotation was not carried out by multiple annotators. While hedging is an inherently subjective phenomenon and therefore disagreement between different annotators is to be expected, the amount of disagreement would be better measured if individual-independent annotations could be compared to produce inter-annotator agreement statistics. Multiple annotation was considered but not implemented in this research, since I decided to put more emphasis on the discovery of new forms of hedging, rather than only requiring annotation decisions on top of pre-annotated categories. Another impediment to implementation of independent annotations was the demanding requirements of suitable annotators: a good understanding of the concept of hedging, and of the annotation corpus topic. Additional annotators' involvement is considered as a future research described in Chapter 6.

3.4.4 Annotation strategies

Manual annotation bootstrapping

The pre-annotation based on the lexicons of Rubin's items, epistemic phrases and additional hedging expressions found in the pilot task ease the annotation task by helping hedge spotting so there is less probability of overlooking lexical items. Since these lexical items are mostly out of domain, it is still likely some items may be dismissed. A procedure of bootstrapping based on a percentage of annotations already done helps to decrease the dismissal of hedges in-domain.

Annotation of consecutive hedges

When multiple hedges appear consecutively in a sentence in what would seem a hedging expression in the form of a collocation, the criterion of substitutability was chosen to decide if there was a single hedging expression or multiple expressions. This criterion was followed by Holmes [1988] as proposed by Kennedy [1987a] in his study about quantifiers in English.

²⁰The principle that examples in the training dataset should not appear in the testing dataset is kept.

Delimitation of hedging expression constituents

For Other category of hedges, it was important to identify the boundaries of hedging expressions. One question to be asked was if a hedge candidate could be replaced by any other hedging device.

For instance *Without more information it is hard to tell* can be rephrased to *Without more information I probably won't be able to tell* to check if the resulting proposition resembles the original one and to be able to determine which are the items to be marked as part of the hedging expression.

Uniformization

Hedging expressions marked by the automatic matching procedure had their spans modified in order to provide more standard matching of hedging according to their morphology. For instances, a hedge such as *does n't always*²¹ was changed to *n't always* in favour of making expressions more standard as to match other phrases such as *do n't always*.

Forms similar to the standard ones defined in Section 3.3.4 for epistemic phrases were preferred to semantically equivalent but not lexically equivalent forms. For instance, if *I am hesitant* is found in a proposition the whole phrase was annotated and not just *am hesitant*. The same happens with *I hope* and *hope*, as *hope* is likely to be used in informal setting, but if there is a match for the subjected phrase *I hope* was annotated.

As the annotation process followed an iterative strategy to ensure cohesion in annotation, this uniformization was continually revised and where discordances were found, also deletion of previous annotations took place.

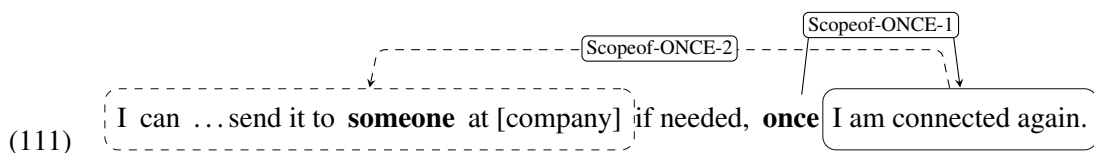
Checking items out of context

This is useful to detect potential mistaken hedging expressions as they could only be given an apparent hedging sense because of the words surrounding it. This strategy was also used to correct miss-selected expressions caused by visually selecting the wrong text span when dealing with the annotation tool.

3.4.5 Overlapping issues

Given the entities defined for the annotation scheme, there will be cases where the spans covered by the entities overlap. For instance, the second scope of *once* covers the Single hedge *someone* in (111). The fact the *someone* falls within the scope of a hedging expression may require further semantic interpretation as the segment in the sentence contained by the scope do not represent factual information. The annotation scheme has to allow this kind of overlap in order to comply with the requirement of flexibility. In Section 4.4.2 I will point out more cases where the scope overlaps with other entities, however, semantic interpretation of the interaction of overlapping entities will not be addressed in more detail.

²¹Extracted from Rubin's lexicon.



3.5 Conclusions

This chapter has described an annotation scheme for hedging in informal language style. This kind of work is relevant, apart from the study of linguistic expressions, because automatic identification of hedges creates the need to have well delimited annotated chunks of text that represent entities related to the phenomenon of hedging.

I have identified two kinds of scenarios in addressing hedges. In one, despite being clear that the Source of hedging is not always the writer, existing automatic methods are not concerned with identifying this element, because they focus on the propositional content (content-centered). Nonetheless, this approach is valid for analysis of hedges in academic prose and that has been used in most of the automatic methods for identification of hedges so far.

There are, however, other scenarios where discerning who is the experiencer of the hedging event is relevant, such as when it is imperative to determine whether or not a user A in an online web community has certainty about information X; this gives rise to analyses that are ‘user-centered’ and this situation is frequent in informal language style, where expressions such as *I’m not sure* and *IMHO* are used. At the same time these expressions are unlikely to occur in more formal domains such as in research articles.

As seen in the previous chapter, there is not consensus about the concept of hedges in literature and in the same way, there is not consensus in studies for automatic identification of hedges about which the adequate lexical realisations of hedges are, moreover so far, they have not fully included more informal expressions of hedging.

This study aims to contribute to both content and user-centered research on hedging expressions in informal language style. Therefore, I have defined a set of requirements, tailored to the domain and language style under study, the designed annotation scheme should comply to so these objectives are achieved.

I defined four categories of hedges found in informal language that were described in this chapter: Single hedges, Not-Claiming-knowledge epistemic phrases, Syntactic hedges, and Other hedges. Guidelines on how to proceed on particular and exceptional cases were described along with each category.

Furthermore, I defined the category of Single hedges out of a set of lexical hedges extracted from Rubin that mostly conform to the concept of epistemic modals or traditional hedges such as *may*, *probably* and *likely*.

Additionally, there are two types of Source: Inner Epistemic Source and Outer Epistemic Source. The Inner Epistemic Source refers to the individual or individual experiencing a mental state that translates into a hedging expression, while the Outer Epistemic

Source is the individual who wrote a proposition (Writer). Some criteria to distinguish these two types of Source have been mainly proposed according to how they occur in the language style under study. The annotation scheme provides the means to annotate distinctively when the Inner epistemic source is not the writer.

I proposed the annotation of the hedging scope in minimal fashion: at least the dependent head of the scope has to be annotated in contrast to other studies, where the exact scope has to be annotated.

In the proposed annotation of the scope, its constituents are separated from the hedging expressions in contrast to earlier studies where the hedge was annotated within the scope boundaries. Choosing the right annotation tools helps to transcend this notional choice as the benefit is that lexical constituents that do not actually form part of the scope can be left out from being annotated.

Based on the idea that first person epistemic phrases such as *I think*, *I don't know*, and *I would suggest* acquire a semantic and pragmatic interpretation in comparison to other kinds of epistemic phrases, I have proposed a category named Not-claiming-knowledge (NCK) epistemic phrases or hedges. I have provided linguistic support for this distinction and as initially a set of these phrases was taken from Karkannen's work, I have described which kind of NCK phrases could be expected to be found and therefore, extended types, such as *My guess is* and *IMO* are also considered for annotation in this category. I have also underlined some descriptions in terms of subjective and objective distinctions in this category to account for some cases where it seems that categorical assertions are made in contrast as what a hedge is supposed to convey: modalised assertions.

I have also presented a strong linguistic foundation of the subjective role of first person epistemic phrases and distinctions between subjective and objective uses of epistemic modality. These distinctions show that the group of not-claiming-knowledge hedge comprises epistemic phrases expressing weak commitment and epistemic phrases expressing lack of commitment to the claim of knowledge.

Particularly, Kärkkäinen and Wierzbicka were the first to have studied epistemic phrases, not only those used to hedge but also those that convey claiming knowledge such as 'I know'. However, they have studied them with respect to the way they convey a speaker's stance.

One of the potential advantages is that in NCK the source is enclosed within the hedging expression.

The third category of hedges I proposed is Syntactic hedges, given their structural differences from Single-hedges and NCK hedges at the sentence level. I have considered in this study the classification of conditionals made by Iatridou: relevance, factual and hypothetical conditionals, and I have discussed representative examples of each kind to discuss in which cases they can be deemed as signals of hedging.

The fourth category of hedges that I have proposed called Other hedges comprises hedges that could not be either classified into any of previous categories or that have the structure of a Non-Claiming-knowledge hedge but are domain-specific such as *in a com-*

puter illiterate and therefore would not be compliant to be used in other domains.

From the four categories proposed, I consider Single-hedges and NCK phrases as the main categories that I addressed in my study and are typical representatives of hedging expressions. Nonetheless, Syntactic and Other hedges are relevant for a complete analysis of hedges in this particular language style.

The set of entities and relations proposed were organized into an annotation scheme template so manual annotation could be carried out.

I described the main steps of this manual annotation, that included some semi-automatic procedures since it was designed as an iterative procedure where annotations and annotation template could be refined and tailored to the language style being addressed. I organized the annotation procedure into these main steps: pre-processing, pilot annotation, pre-annotation and manual annotation.

The pre-processing is mostly an automatic procedure, although a previous analysis of the dataset is required. It comprises common pre-processing steps in language processing, such as sentence splitting and tokenization. Other steps are normalization of extra-linguistic (eg images embedded in posts) and pseudo-linguistic textual elements e.g. emoticons and smilies and identification of meta-tags e.g. tags for quotations.

The pilot annotation had the purpose of performing a preliminary corpus linguistic study of the dataset. This step went alongside study of the state of the art around hedging. This step provided an initial annotation template that encoded conceptual representations from the designed annotation scheme. Additionally, hedge types found in this step helped to shape up the initial lexicons described in Section 3.4.3.

Pre-annotation included an automatic marking in the required annotation tool format of entities according to the initial lexicons of hedges.

The manual annotation itself comprised checking over pre-annotated entities representing hedge occurrences, finding new ones and marking other elements such as the source and scope of the hedging expression. I deemed this step as a point of feedback for the iterative annotation, since new occurrences could be taken to enrich the lexicon, perform a new pre-annotation and improve the annotation scheme and in some cases produced insights to be used in pre-processing, for instance to improve ill sentence splitting.

I also described some strategies for manual annotation that may help to improve the quality of annotation to some extent, since in this study the annotation task was carried out by a single annotator.

Some overlapping issues had to be addressed practically and, they conceptually became cases where annotation entities overlap such as in the cases where a hedge lies within the boundaries of another hedge scope.

Chapter 4

Theoretical and empirical issues about hedging in-domain

The process involved in the creation of a categorization for hedges in an informal domain was described in Section 3. Each element playing a role in in-sentence hedging was outlined separately to provide a rough depiction of the features that emerge from the use of hedging. In this chapter, I set out to describe in-depth empirical findings and discuss theoretical issues that emerged as a result of the annotation work. In previous sections the nature of the dataset has been outlined and in Section 4.1, more detailed dataset descriptions will be provided. An overall description of the distinct kinds of hedges distribution will be described in Section 4.2 and each hedge category will be discussed separately in Sections 4.2.1 to 4.2.4.

Various possible interactions between hedging categories and elements are found to be occurring in the domain under study, however I only choose a minimal important set of interactions with the hope they will representatively illustrate the various realisations of hedging in the domain of a particular web forum corpus. These interactions are presented in Section 4.3.

In Section 4.5, I will provide insights on how hedges are used in the web forum dataset under study according to categories from a pragmatic taxonomy of hedges proposed by Hyland [1998].

Finally, conclusions about hedges in the forum dataset will be provided in Section 4.6.

4.1 Forum dataset

In this section, I describe the nature of the dataset processed in the annotation task described in Chapter 3. As hedges were annotated at a sentential level, this description is given in terms of distinctive types of sentences relevant to the analysis of hedges. Categorizing sentences by their function as declarative, imperative, interrogative and exclamative aided in obtaining insights about the distribution of hedges and in defining the research boundaries of this study. Given the noisy nature of text from web forums, these functions were extended to make them suitable to the less traditional content style that is characteristic of this

particular dataset. Three variables are taken into account to devise a categorization of sentences that have interactions with the occurrence of hedges, namely whether the sentence is: a) interrogative, b) processable and c) a quotation.

Sentences with an interrogative function in the forum dataset do not necessarily end with a question mark as the language used is not strictly grammatical. Nonetheless, the interrogative intention can be understood due to the syntax, as in (113) and (114). Usually, question marks are employed for formulating questions, nonetheless they are also used to accomplish various goals as: for making petitions (114), suggestions (115), used in a rhetorical fashion as in (116), or even used when there is not a clear interrogative pursuit (117). The absence of question marks in interrogative sentences is a drawback when considering devising an automatic system for language processing, a single method relying on the occurrence of a trailing question mark would certainly fail in many cases. More complex techniques could be used for this purpose, but elaborating on them is beyond this research's objectives. To ensure a consistent study of hedges, questions with an absent question mark were manually marked as interrogative in the annotation dataset. Out of 1,595 interrogative sentences, 136 were manually marked as such because they had an interrogative function even if they did not have a trailing question mark.

- (112) How long should this require? Post: 44026
- (113) Does the two out of seven not detect anything at all or just some. Post: 51953
- (114) Could someone try to help me with this problem ? Post: 649
- (115) Could it be Ghost is just not compatible with Vista? Post: 6222
- (116) # \$ % ^ why must new computers have a preinstalled AV product ! @ # ! \$ @and i must do lots of things to get rid of them ? ARGH ... Post: 41133
- (117) Doing a search on Google there's hardly any info on this? Post: 61734

Processable sentences are those that correspond to declarative sentences in the traditional functional classification of sentences and they can also include or be exclusively one of the non-linguistic elements described in Appendix E, such as links, emoticons and timestamps. Examples (118) to (118) show some cases of processable sentences, these were authored by the post's writer and are remarkably different from quotative and usually from non-processable sentences.

- (118) I do manual backups so I have always disabled the backup option anyway.
- (119) Hi UUUUUU,
- (120) EDIT: or check this: LLLLLL
- (121) Perhaps I should check if they would let me buy the CD [company_name] downloaded 2010 to my desktop when I wasn't looking OOOOOO with instructions to click one button for an automatic install, which worked perfectly, this is the first time I do not have a CD.

(122) Thanks

Non-Processable items on the other hand, are not proper sentences but strings that are not fully grammatical natural language expressions. These are often product of software applications processing such as (123) or technical information provided by users (123b). No automatic mechanism was devised for the identification of complex non-processable sentences;¹ in the annotation dataset they were marked manually resulting in a total of 1,184 sentences.

- (123) a. Process 1700 (\Device\HarddiskVolume3\Program Files (x86)\
[product_name]\Engine\IPIPIP\ccSvcHst.exe) has opened key \REGISTRY\
USER\S-TTTTTT-163695203-2985681545-29013369-1001
- b. Win32 Version: IPIPIP (RTM.050727-4200)

Quotative sentences represent quotations of other users' talk, content copied from software applications' output and pasted by the user, or information from heterogeneous sources that the user deemed important enough to include in the body of the post. The quote could be explicitly formatted using the forum support system capabilities; in this case the identification of a quotation is straightforward and appears in the standardized text as the wildcard *QQQQQQ*. Quotative sentences without a standard format are not automatically detectable, except for cues that the author provides or some other non-standard format. Therefore, 1,832 quotative sentences in the annotation dataset were manually identified.

Considering these variables, two levels of sentence categories were defined to examine them in terms of function. The first one divides sentences into Interrogative and Non-interrogative. The second level of sentence categorisation divides sentences into Processable, Non-processable and Quotations. These categories are mutually exclusive within each level, but there are intersections between categories from both levels.

Overall, the whole set of posts' contents are divided into 21,552 sentences² in the annotation dataset; the distribution of sentences across the two levels of categorisations is shown in Table 4.1. Out of this, 18,538 are processable sentences that are authored by the user independently of whether they are interrogative or not. There was an emphasis on identifying this subset of processable sentences as it is used to leverage the number of words that are actually produced by a user and that therefore can be used to determine the distribution of hedges in posts' contents. Otherwise, the analysis would include posts whose main content is built out of non-processable strings as in (123) for instance. Within the set of processable sentences, the analysis of hedges regarding individual sentences and across the

¹Complex non-processable sentences are not easy to spot as they are a mix of natural language and non-processable string. For instance, in (123), the word *Process* is enclosed in a non-processable sentence as this statement was generated automatically by a software tool, not by the post's writer. On the other hand, it could also be used in a processable sentence such as *I left the process running on my pc..* The mix of words and non-processable strings makes this kind of sentence difficult to identify without having mismatches and a prior knowledge of which kind of sentence was not produced by an individual.

²The sentence splitting procedure is described in section 3.4.3.

annotation dataset overall will focus on the set produced by intersecting processable and non-interrogative sentences (16,720 items).

Table 4.1 – Distribution of sentences per type in the annotation dataset.

	Processable	Quotations	NotProcessable
Interrogative	98.68% (1,575)	1% (16)	0.31% (5)
Non-Interrogative	83.78% (16,720)	10.3% (2,056)	5.91% (1,180)
Subtotals	18,295	2,072	1,185

In the annotation dataset consisting of 3,000 posts, 2,981 of them are processable as 19 of them contain quotations only. From the remainder, 73 posts only contain questions and 1 post contains exclusively both questions and quotations, therefore, when analyzing non-interrogative processable sentences, only 2920 posts are analysed. An adequate identification of processable sentences and posts is fundamental in a text analytic system, where author characterisation is required. This issue is related to the remarks in Section 3.2 about the source of hedging identification. Incorrect attribution assessment could be done in for instance the post in (124). The only sentence written by the post’s author is *ok - but stick with me ... it says:*, the shaded area highlights content quoted by the author, and it contains many potential hedging words such as *should*, *attempt*, *someone* and *may*. From these ones, only *should* does not convey uncertainty, however, the shaded text is more likely extracted from instructions for the use of a specific software application.

(124) ok - but stick with me ... it says:
 You **should** not run ComboFix unless you are specifically asked to by a helper.
 Also, due to the power of this tool it is strongly advised that you do not **attempt** to act upon any of the information displayed by ComboFix without supervision from **someone** who has been properly trained.
If you do so, it **may** lead to problems with the normal functionality of your computer .

Post: 91941

4.2 Profiling hedges in the forum dataset

This section provides a detailed description of hedge types and frequencies according to the categorization proposed in Chapter 3 as product of manual annotation. I start by giving an overview of hedge types’ distribution across posts and later go into detail presenting lexical types for each hedge category. In Sections 4.2.1 to 4.2.4 hedge types belonging

to the four categories: SINGLE-HEDGE, NOT-CLAIMING-KNOWLEDGE (NCK) epistemic phrases, OTHER and SYNTACTIC are shown and described.

The dataset analysed comprises 2,981 posts that contained at least one sentence that was not a quote or non-processable sentence. The other 20 were not used in profiling because they had no content that was useful or that could be attributed to the post’s author eg. they only contain quotations, as shown in previous Section 4.1.

As a whole, 790 distinct types of hedges were found in the annotation dataset: 272 SINGLE-hedges, 300 NOT-CLAIMING-KNOWLEDGE epistemic phrases, 209 OTHER hedges and 8 SYNTACTIC hedges. Some original types categorised as SINGLE-HEDGES or NCK epistemic phrases were normalised, so as to provide a better leverage of the quantity and quality of hedging expressions for these two categories. The lexical normalisation procedure for each category of hedge will be described in Sections 4.2.1 and 4.2.2 respectively. Normalised types sum up to 189 for SINGLE-hedges and 138 for NCK phrases. Normalisation widened the gap between number of types of SINGLE and NCK hedges types, showing that NCK variations (reduction of 54% in NCK types) conveying the same meaning of a particular NCK type are more frequent than SINGLE-hedge variations (reduction of 30.5%). These lexical variants will be discussed in subsequent sections.

Overall, 69.51% of posts in the annotation dataset have at least one hedge occurrence. For each hedge category, Table 4.2 shows the proportions of posts containing hedges from that particular category in both annotation and reduced training datasets. Nonetheless, counts in the RTD are made only by following a string matching algorithm, therefore some cases may be false positive cases of hedges. Single hedges occur in the highest proportion of posts in both datasets, while the difference in proportions varies significantly in the case of Syntactic hedges.

Table 4.2 – Percentage of posts containing hedges from each hedge category in the annotation dataset (Annset) and reduced training dataset (RTD).

Hedge categories	Posts containing hedges			
	Annset		RTD	
	Count	Percentage	Count	Percentage
Single-hedges	1,803	60.48	118,618	76.39
NCK	746	25.03	36,241	23.34
Syntactic	738	24.76	83,558	53.81
Other	276	9.26	9,328	6.01

Across individual posts, Figure 4.1 shows the distribution of hedge tokens found in each category. The largest number of hedge occurrences in posts concentrates in the SINGLE-hedge category with 5,651 hedge occurrences overall. The highest frequency of individual hedge occurrences in posts are SINGLE-hedges, they occur 1.87 times on average in each post, but these hedges are quite spread as Table 4.3 shows. However compared to the dispersion in other categories revealed by the coefficient of variations, single-hedges are the less

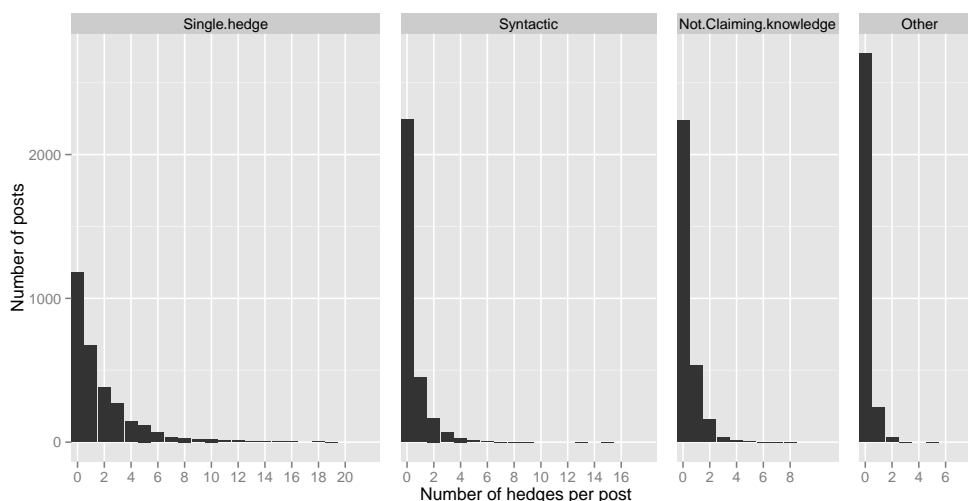


Figure 4.1 – Distribution of posts by frequency of hedges across different hedge categories in the Annset. The upper limit in the number of hedges per post shown in the distributions for SINGLE-hedges and SYNTACTIC hedges was set to 20 for reasons of space. There are 7 posts that have a number SINGLE-hedges greater than 20, each one with 22, 24, 25, 29, 30, 49 and 96 hedges correspondingly. Also, there is one post with 29 SYNTACTIC hedges.

disperse. Nonetheless, all hedge occurrences in the annotation dataset are sparse. A group of 131 posts are statistical outliers with respect to occurrence of SINGLE-hedges since they have more than 7 hedges per post. The maximum number of hedges in a single post is 96 SINGLE-hedges. Next highest hedges in token frequency are SYNTACTIC hedges that make up to 1,257 occurrences, NOT-CLAIMING-KNOWLEDGE hedges have 1,050 occurrences while OTHER hedges sum up to 313. Considering the distributions of the number of hedges per post in each hedge category shown in Figure 4.1 have a mass point at zero, the statistical outliers are posts with one or more hedges.³ This happens because most posts do not include any kind of hedge.

Table 4.3 – Mean, standard deviation and coefficient of variation of hedges from each category across posts in the annotation dataset (Annset).

	mean	sd	cv
Single-hedges	1.87	3.33	1.78
NCK	0.35	0.73	2.07
Syntactic	0.43	1.12	2.57
Other	0.11	0.36	3.36

Some methods were applied for outlier identification for each category of hedging individually and for hedge occurrences as a whole. A standard method for outlier detection based on lower and upper quartiles was applied to hedge frequencies per category and to hedge frequencies averaged to the post's size. However these methods did not prove effective since the application of this method to rule out outlier posts lead into dismissing posts

³The number of hedges per posts will be addressed as a zero-inflated distribution in Chapter 5.

that are not real outliers. For instance, posts such as (125) and (126) would be outliers as they have a high average of hedges in relation to number of words, but this happens often when there are multiple hedges per sentence or when most of the sentences in a single post have at least one hedge.

(125) **Perhaps** this **might** help

LLLLLL

Post: 157919

The next thing I **would try** is to delete manually the virus def files after Tamper protection is off.

But I THINK you should **try** it only after Tim gave **some other** solution ,

(126) **maybe** he **would** need **some** files **or** logs from there

Or do a backup of those files , and from the log files as well , and then **try** to delete these.

After it run LU manually , and it will redownload the whole definition files.

Post: 244871

An outlier analysis based on multiple features was tested as alternative to individual hedge category-based outlier identification as outliers for each hedge category do not always converge to the same set of posts. Multivariate methods for outlier detection were tested such as methods based on the Mahalanobis distance of the observations to a hypothetical normal distribution [Filzmoser et al., 2005], however the sparseness of hedge occurrence across the annotation made these methods unsuitable. Outlier detection is not further addressed in this research and therefore the profiling and relations found in this dataset are performed over the entire set of annotated posts without other restriction than dropping posts with non-processable content. Particularly, for the findings described in this chapter no outlier posts were dropped.

Figure 4.2 shows the frequency per hedge type in total. This chart shows that a high frequency of hedge types have only one or two occurrences in the annotation dataset. A more detailed view per hedge category in Figure 4.3 shows that types from the NCK epistemic phrases and Other hedge categories are the ones that have more sparse occurrences of hedge types, followed by hedge types from the Single-hedge category. The issue of sparseness was relatively accounted by normalising lexical occurrences. This was done by looking at the various types that are lexical/syntactic variations of more general hedge types.

In the annotation dataset, 2,042 posts have at least one hedge type. The distribution of posts containing at least one hedge from a particular category is shown in Fig. 4.4. Intersections express the occurrence of hedges from distinct category in a single post. Posts only containing Single-hedges constitute the largest group in the annotation dataset (26.97%) after posts with no hedges (30.49%).

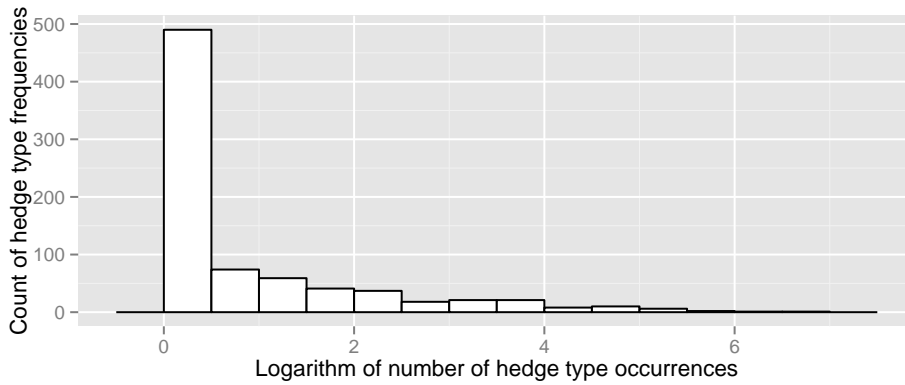


Figure 4.2 – Count of non-normalised hedge type frequencies (log scale).

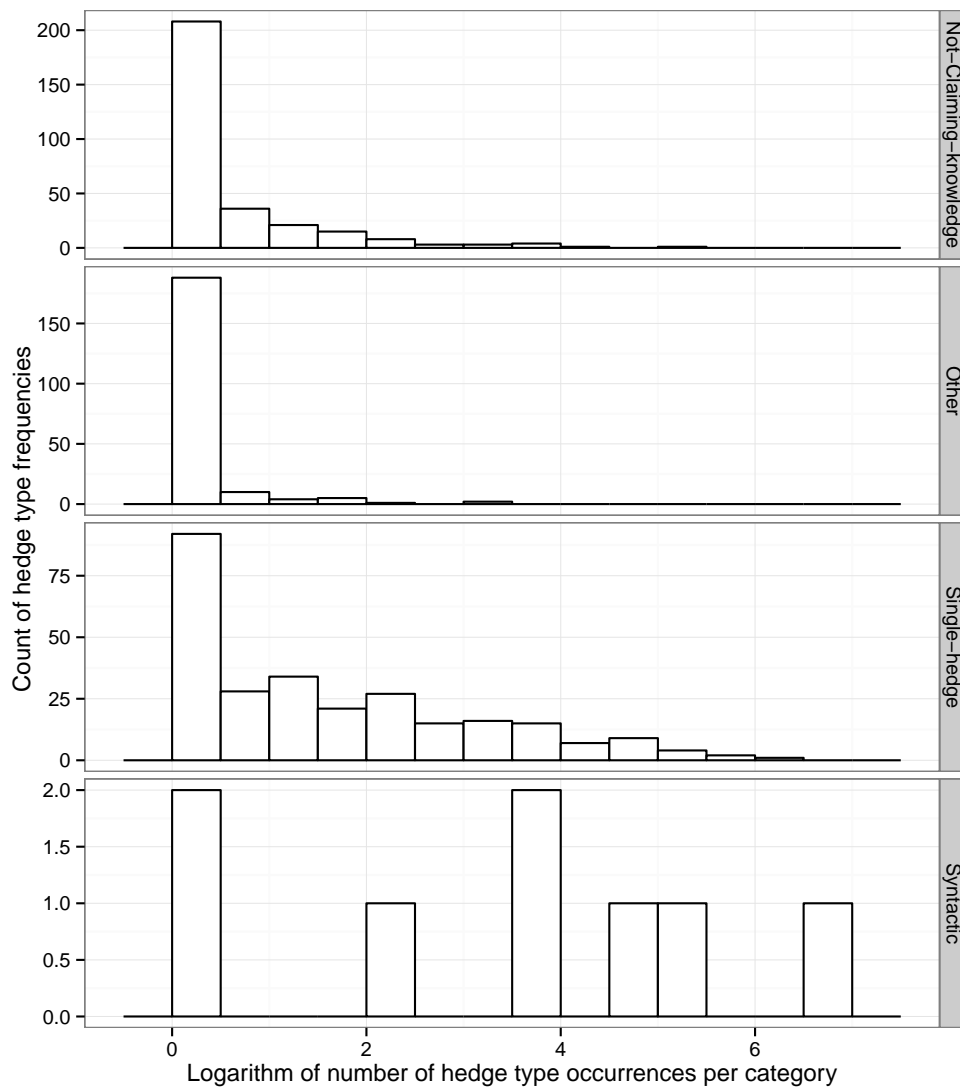
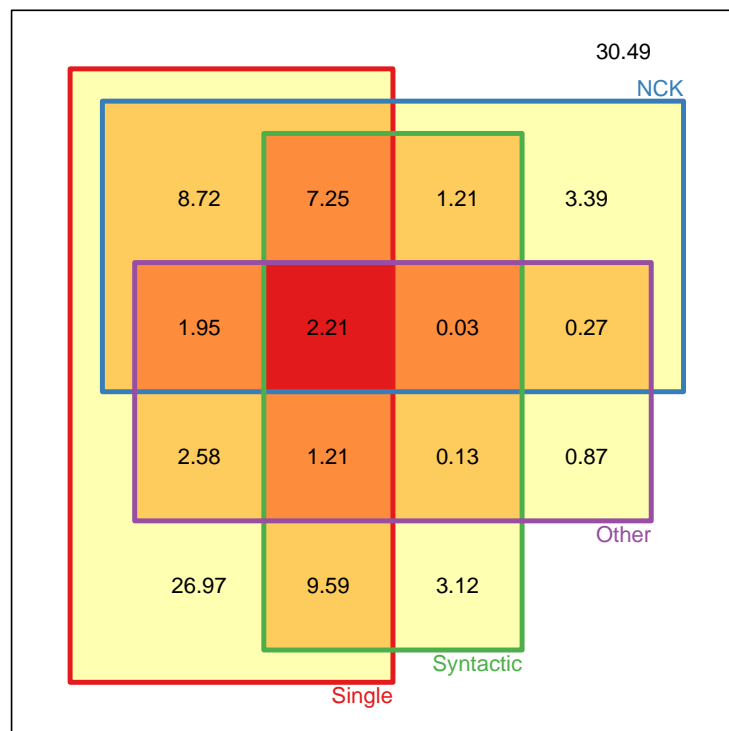


Figure 4.3 – Count of non-normalised hedge type frequencies (log scale) in the four hedge categories.

Table 4.4 – Percentage of posts containing hedge types.

		has.NOT.Syntactic		has.Syntactic	
		has.NOT.Other	has.Other	has.NOT.Other	has.Other
has.NOT.Single	has.NOT.NCK	30.49	0.87	3.12	0.13
	has.NCK	3.39	0.27	1.21	0.03
has.Single	has.NOT.NCK	26.97	2.58	9.59	1.21
	has.NCK	8.72	1.95	7.25	2.21

Figure 4.4 – Percentage of hedge per type occurrence in posts from the annotated dataset.

4.2.1 Single hedges

The hedging items in this category were found following the principles underline in Section 3.3.1. Raw frequencies of Single hedges in the annotation set are shown in Table 4.5, grouped by a normalised form of them in case of the token not being in a standard form. Subtotals for normalised groups are also provided.

The normalisation procedure was carried out in a post-annotation stage with the purpose of grouping expressions with non-significant morphological divergence. The normalisation strategies explore these morphological variations of items corresponding to the same word type and they were undertaken depending on the original type form as found during annotation. These strategies are:

- **Abbreviations and non-standard abbreviations.** For instance, a standard abbreviation for *approximately* is *approx.*. Abbreviations such as *approx* were also found in the annotation dataset. All of these items were normalised to *approximately*.
- **Contractions.** For instance, the contracted form *'d* was normalised to *would*.
- **Typographical errors and misspellings.** For instance, expressions such as *world* and *wuuld* were found to stand for *would*.
- **Tense and number variations.** Items such as *claim*, *claims* and *claimed* were normalised to *claim*.
- **Colloquial forms.** Hedge types such as *kinda* were normalised to *kind of*.

Following this procedure 270 original types are condensed into 189 normalised types.

Table 4.5 – Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
1	would	would	441	491
		'd	48	
		world	1	
		wuuld	1	
2	try	try	175	450
		tried	147	
		trying	111	
		tries	17	
3	some	some	396	396
4	other	other	305	357
		others	52	

Continued on Next Page...

Table 4.5 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset (items with minimum frequency of 27, for complete table cf. Section D.1).

	Normalized keyword	Original keyword	Freq.	Subtotal
5	may	may	155	319
		maybe	93	
		may be	71	
6	can	can	226	226
7	seem	seems	113	196
		seem	57	
		seemed	23	
		seemed like	1	
		seem like	1	
		seems like	1	
8	could	could	169	169
9	something	something	144	162
		something like	9	
		something else	7	
		somethink	1	
		somethinge	1	
10	might	might	133	133
11	question	question	75	125
		questions	50	
12	many	many	101	101
13	a few	a few	98	99
		a few others	1	
14	several	several	98	98
15	about	about	88	88
16	should	should	83	83
17	suggestion	suggestions	48	80
		suggestion	30	
		sergestion	1	
		sergestions	1	
18	probably	probably	80	80
19	appear	appears	51	79
		appear	19	
		appeared	9	
20	most	most	45	64
		most of	19	

Continued on Next Page...

Table 4.5 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset (items with minimum frequency of 27, for complete table cf. Section D.1).

	Normalized keyword	Original keyword	Freq.	Subtotal
21	attempt	attempt	27	63
		attempted	15	
		attempting	11	
		attempts	8	
		attemting	1	
		attempteing	1	
22	sometimes	sometimes	48	61
		sometime	10	
		some times	3	
23	suggest	suggested	49	60
		suggests	5	
		suggest	3	
		suggesting	3	
24	perhaps	perhaps	57	57
25	someone	someone	53	53
26	similar	similar	46	48
		similiar	2	
27	possible	possible	45	45
28	another	another	43	43
29	like	like	40	40
30	a bit	a bit	35	39
		a bit of	4	
31	look like	looks like	26	35
		look like	6	
		looked like	3	
32	one of	one of	34	35
		one of those	1	
33	a lot	a lot of	25	34
		a lot	9	
34	think	think	16	33
		thought	13	
		thinks	4	
35	a little	a little	32	32
36	anyone	anyone	29	29

Continued on Next Page...

Table 4.5 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset (items with minimum frequency of 27, for complete table cf. Section D.1).

	Normalized keyword	Original keyword	Freq.	Subtotal
37	hopefully	hopefully	29	29
38	certain	certain	27	28
		cerain	1	
39	almost	almost	27	27
40	strange	strange	24	27
		very strange	3	
41	likely	likely	27	27

I discuss below some instances of single hedges that are not traditionally considered as epistemic modals.

Adverbials of time can be used as hedging expressions when they convey that a state of affairs does not occur in a constant basis (127).

- (127) Full diagnostic, specifically video card and ram ... **sometimes** [_{SCOPE} it's just the way the software is using parts of your hardware that causes it to freeze rather than the software itself].

Post: 100139

Hedging expressions equivalent to adjectives in negated form are also included in this category:

- (128) Moving on, there are **no specific** [_{SCOPE} rules], I make them up as I go along

Post: 12475

Items such as *don't know* were included in this category when the source of the speculation is not the writer, even when the subject is not explicit in the text. A comparison of the case when the implicit subject is the writer and therefore the expression tagged as NCK-phrase is given in Section 4.2.5.

4.2.2 Not-Claiming-Knowledge Epistemic Phrases

As mentioned in Section 3.4.3, using epistemic phrases as annotation units was inspired by insights in [Kärkkäinen, 2010], a study on spoken American English conversations that analyse epistemic phrases and their scope. The set of epistemic phrases composed of: {*I think, I don't know, I know, I thought, I guess, I don't think, I remember, I'm sure, I hope* } and some additional phrases from Wierzbicka's [2006] work: {*I expect, I believe, I suppose, I assume, I imagine, I gather, I presume, I guess, I suspect, I take it, I understand,*

I trust, I wonder, I feel} was refined by dropping out phrases that do not convey hedging,⁴ and expanded with new types as found during the annotation phase. An abbreviated list of Not-Claiming-Knowledge (NCK) first person epistemic phrases is shown in Table 4.6. Normalised types are shown in the second column, accompanied by the original type followed by its raw frequency, and a subtotal number representing the sum of all original types frequencies is shown in the last column, this value is also used to sort the occurrences being shown there. The full set of NCK first person epistemic phrases is presented in Table D.2. For instance, *I think* is the most frequent item across the whole annotation dataset as *I thought* is also considered a morphological variation. The second most frequent hedge type is *I hope*, the frequency for the original type as *I hope* does not differ significantly from its elliptical form frequency. The phrases *I do not know* and *I am not sure* follow in frequency. The lexical variations for *I do not know* and *I wonder* constitute the largest groups in the NCK phrases category.

The asterisk (*') in the collected lexical forms (as *I * think* in the first group at Table D.2 and *I * do n't know* in the third one) represents a non-contiguous phrase such as the one in example (129). In this example, the constituents of the epistemic phrase *I think* are marked in bold and they are linked by an edge labelled with the complete epistemic phrase. The non-contiguous marking highlights the case where the grammatical dependency constituents in the epistemic phrase are not consecutive in the sentence. This case should be distinguished from the case of elliptical NCK phrases as *Think* in (130). Frequencies of some elliptical epistemic phrases found in the annotation stage are presented later in this section.

- (129) ... **I** found this one (sorry) and **think** it may be related . Post: 20579
- (130) **Think** you're talking about the [product_name] .. Post: 85261

Epistemic phrases are often constructed around a lexical hedge as those belonging to the Single hedge category and include a personalized article. For this research as the main motive was capturing where the author's stance is conveyed, the first person article *I, my* and *we* are targeted to create a set of NCK (first person) epistemic phrases in contrast to the super set of possible epistemic phrases, whose use may differ in intentionality.

The normalisation done for NCK phrases was similar to Single hedges in terms of the strategies for abbreviations, contractions, typographical errors, tense and number forms and colloquial forms. Nonetheless, there are some particular cases more prevalent in NCK phrases such as the inclusion of modifiers that extend the normalisation procedure. Modifiers are also constituents of epistemic phrases as in (131). While it is acknowledged that *I'm not 100% sure* is different to *I'm not sure* in terms of gradability, *I'm not 100% sure* is considered as a syntactic variation of *I'm not sure* to the same extent *I'm not completely sure* is. Elliptical forms of phrases where the pronoun is suppressed is also normalised to

⁴A study using the whole range of epistemic phrases in text from the same domain is addressed in [Vogel and Mamani Sanchez, 2012].

the epistemic phrase with overt subject. Finally, automatic normalisation procedures would need a conglomeration of the aforesaid techniques as combination of non-standard realisations for a single phrase are common, e.g. to map *I still dont understant* to the normalised form *I do not understand*, three issues have to be addressed: (a) correct the misspelling in *understant*, (b) correct the typo in *dont* to *don't* and (c) normalise the constituent *don't* to *do not*. Grammar and spelling correctors can address some of the issues here, but still de-contraction to a standard form would be needed.

(131) **I'm not 100% sure** I've got a nasty bug.

Post: 201297

The manual normalisation procedure applied to original NCK types rendered 138 types, out of 303 original types. Compared to the set of Single hedges that was reduced to a 70% set of the original set of types, there is a larger reduction in the NCK types set to 45%. This difference reveals a higher variation in the group of NCK hedges, be it due to grammatical morphological variations or ungrammatical forms.

Table 4.6 – Raw frequencies for Non-Claiming-Knowledge phrasal types in the annotation dataset (phrases whose frequency is higher than 11, for full set of lexical items cf. Section D.2.).

	Normalized keyword	Original keyword	Freq.	Subtotal
1	i think	i think	157	210
		i thought	35	
		i 'm thinking	6	
		i * think	5	
		i thing	1	
		i was thinking	1	
		i still think	1	
		i am thinking	1	
		i now think	1	
		think	1	
		i thinh	1	
2	i hope	i hope	59	125
		hope	49	
		i was hoping	4	
		i 'm hoping	3	
		i sure hope	2	
		i do hope	2	
		i just hope	2	
		i hoped	1	
		i am hoping	1	
		i had hoped	1	
		i am hopeful	1	

Continued on Next Page...

Table 4.6 – Continued: Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset (phrases which frequency is higher than 11, for full set of lexical items cf. Section D.2.).

	Normalized keyword	Original keyword	Freq.	Subtotal
3	i do not know	i do n't know	43	89
		i dont know	8	
		i do not know	8	
		do n't know	6	
		i did n't know	5	
		did n't know	3	
		i really do n't know	2	
		do not know	2	
		i idd not konw	1	
		dont know	1	
		i dont even know	1	
		dunno	1	
		i do n't know for sure	1	
		i did not know	1	
		i do n't have a way of knowing	1	
		i know nothing	1	
		i really do not know	1	
donno	1			
i do notknow	1			
i * do n't know	1			
4	i am not sure	i 'm not sure	30	75
		not sure	19	
		i am not sure	16	
		i am just not sure	2	
		im not sure	2	
		never been sure	1	
		i was not sure	1	
		i 'm not 100 % sure	1	
		i * am not sure	1	
		i am not quite sure	1	
		i 'm not quite sure	1	
5	i believe	i believe	46	48
		i beleive	1	
		i believed	1	
6	i wonder	i wonder	12	43
		i was wondering	6	
		wonder	4	
		i am wondering	4	
		i 'm wondering	3	

Continued on Next Page...

Table 4.6 – Continued: Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset (phrases which frequency is higher than 11, for full set of lexical items cf. Section D.2.).

Normalized keyword	Original keyword	Freq.	Subtotal
	i wondered	2	
	i 'm just wondering	2	
	i did wonder	1	
	i * wonder	1	
	i am starting to wonder	1	
	wondering	1	
	i 'm really wondering	1	
	i also wonder	1	
	wonders	1	
	i 'm still wondering	1	
	i am just a tad wondering	1	
	i * am wondering	1	
7	i guess	30	36
	guess	2	
	i 'm guessing	2	
	i am just making educated guesses	1	
	i guest	1	
8	i do not think	21	29
	i do n't think	3	
	i dont think	2	
	i do not think	1	
	i did not think	1	
	i also do n't think	1	
	i don think	1	
9	i assume	11	21
	i 'm assuming	6	
	i assumed	3	
	i am assuming	1	
10	i do not understand	5	20
	i do n't understand	3	
	i can not understand	2	
	i ca n't understand	1	
	i dont understand	1	
	i do n't really understand	1	
	i do not understand	1	
	do not undrstand	1	
	still ca n't understand	1	
	i personally do n't understand	1	
	i still do n't understand	1	
	dont understand	1	
	i * do not truly understand	1	

Continued on Next Page...

Table 4.6 – Continued: Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset (phrases which frequency is higher than 11, for full set of lexical items cf. Section D.2.).

	Normalized keyword	Original keyword	Freq.	Subtotal
		i still dont understant	1	
11	i suspect	i suspect	14	17
		i suspected	1	
		i 'm suspicious	1	
		i now suspect	1	
12	i have no idea	i have no idea	10	14
		no idea	2	
		i had no idea	1	
		i have not the faintest idea	1	
13	i suggest	i suggest	8	13
		i also suggest	2	
		i respectfully suggest	1	
		i suggested	1	
		i do suggest	1	
14	i am confused	i 'm a bit confused	3	13
		i am confused	1	
		i 'm a little confused	1	
		i am so confused	1	
		i am a little confused	1	
		i am a bit confused	1	
		i am totally confused	1	
		i was totally confused	1	
		i 'm confused	1	
		i am also somewhat confused	1	
		i am really confused	1	
15	imo	in my opinion	7	12
		imo	5	

In the remainder of this section I continue the discussion outlined in Section 3.3.4 about lexical semantic forms of NCK epistemic phrases, emphasizing the subjective/objective epistemic modality distinction and how elliptical forms are realised in the annotation dataset.

The certainty in Not Claiming Knowledge epistemic phrases

It is still not very clear to me how to interpret a proposition such as (132) in terms of certainty vs claiming knowledge values. So, I will start by presenting axioms (A), (B) and (C) that will lead my discussion about this particular case.

(132) I don't know how to post the log file.

- (A). The referent of *I* is not claiming knowledge on how to post the log file.
- (B). The referent of *I* has certainty about the fact of not knowing how to post the log file.
- (C). The referent of *I* has uncertainty about how to post the log file.

I don't know is an epistemic phrase that has taken the primary function of hedging, specifically as an epistemic downtoner or politeness marker [Joanne Scheibman, 2001, Tsui, 2009, Kärkkäinen, 2010]. In sentence (132), it is specifically used as an epistemic downtoner since it conveys doubt in the embedded clause *how to post the log file* in contrast to (133) where the referent of *I* claims knowledge and conveys certainty.

(133) I know how to post the log file.

To me it looks like the focus of the certainty evaluation in (B) is different from the one in (C). The uncertainty in axiom (B) seems to comment about the uncertainty about the world at glance and the one in (C) comments on the certainty about the referent of *I* stance. This stance is in relation to the world, therefore is a comment about the referent's view. Axiom (C) is a comment equivalent *I am not knowledgeable about a specific world* where the focus of hedging is on the specific world whose referent is enclosed by the hedging Scope *how to post the log file*. Axiom (B) is straightforward and certain: I'm certain that I don't know. In this interpretation, the focus is on the referent of *I*, that is the Source of hedging.

This element of certainty underlying NCK epistemic phrases seems to be caused by the presence of the first person articles in them. Moreover, I would say this element of implicit certainty does not seem to underlie propositions such as (134) that have a Single-hedge.

(134) This **may** not be the way to post the log file.

I can think of propositions such as (135a) and (135b) where expressions of certainty and claiming knowledge or not-claiming knowledge are intertwined. Also (135c) and (135d) are perfectly plausible in some scenarios, however the element of certainty does not seem to underlie them as in (132) since they have lost their quality of categorical assertions.

- (135) a. I'm sure I don't know how to post the log file.
- b. I'm unsure about how to post the log file.
- c. **Maybe** I don't know how to post the log file.
- d. I **probably** don't know how to post the log file.

Categorical assertions and modalised propositions

Another issue is the implicit knowledge implied in non-modalised propositions: does any non-modalised proposition claim knowledge?

Taking into account that the propositions (136) and (137) do not have an explicit Source, a rephrasing of (136) as (139) becomes infelicitous, while (140) is perfectly plausible. Also an explicit I-know element accompanying a NCK epistemic phrase is perfectly plausible as in (141).⁵ This shifts the focus onto the question of whether every proposition contains an underlying I-know element. One could make the assertion that only non-modalised (by a Single-hedge) propositions have a covert I-know element equivalent to an explicit I-know. However, propositions such as (138) leads to consider that issues of unqualified assertions and embedded epistemics need further study. However, they will not be addressed in this dissertation.

(136) Not sure what that means, if anything.

Hope it helps!

Post: 179494

(137) This **will** help later in case you want to reinstall [product_name]. . .

Post: 10386

(138) I know I will likely sound like an illiterate flop . . .

Post: 192591

(139) Not sure what that means, if anything.

* **I** know I **Hope** it helps!

Post: 179494

(140) I know this **will** help later in case you want to reinstall [product_name]. . .

Post: 10386

(141) I'm a little scared to start downloading malwarebytes and going at it on my own mostly because I know I don't know what I'm doing. Post: 219850

To conclude this discussion about categorical and modalised assertions, it is worth mentioning that in the literature there are two views about the underlying I-know element in non-modalised propositions. In one, John Lyons [1977] refers to categorical assertion as “straightforward statements of the fact” and are epistemically non-modal. John Lyons introduces two components to give an interpretation of modals functions: An *I-say-so* component and an *it-is-so* component. According to [John Lyons, 1977, p. 797], when a speaker is uttering an unqualified assertion, he or she is committing him/herself to the truth of what he or she asserts, but he is not asserting the epistemically modalised proposition *I know that p*. The speaker is saying, without qualifying either the *I-say-so* or the *it-is-so* component that *p* is true.

The other view is the position adopted by Furmaniak [2011], that holds the idea of an implicit epistemic judgement indicated by *I know that* embeds any categorical assertion, e.g. (142) implies (143).⁶

⁵Sentences such as (141) have been approached in the literature as negative introspection [Egré and Bonnay, 2010, Egre, 2008].

⁶Example by Furmaniak [2011, p. 59]

(142) Peter is here.

(143) I know that Peter is here.

From the examples cited above, it seems that one or other view holds depending on the interaction of the I-know element with the stance expressed in the proposition. Nonetheless, this discussion about the underlying *I know* in sentences where a NCK phrase occurs compared to sentences where a Single-hedge occurs, made evident that different from what happens in Single-hedged propositions, in NCK-hedged propositions the focus of hedging is shared between the Source and the Scope of the hedging phenomenon.

Ellipsis in epistemic phrases

Subject ellipsis is considered an important issue in determining the source of hedging, as introduced in Section 3.3.2. Although subject ellipsis cases were not frequent in the dataset sampled for annotation, it is nonetheless an important morphological variation of NCK first person epistemic phrases. Because of the missing subject, elliptical NCK phrases can be confused in automatic matching with occurrences that do not correspond to this type of hedge. *Wonder* in (144) has the pronoun *anyone* as subject while in (145), *I* is tacit. The same distinction prevails comparing (146) with (130). Some features that may aid in the detections of subject elliptical epistemic phrases are the absence of subject in-sentence and the position at the beginning of the clause, however this model is not guaranteed to be the rule in social media where a subject not referring to the post's writer may be in any other sentence in the document. Darling et al. [2012] point out that in social media platforms such as Facebook and Twitter, first person subject is often assumed e.g. *Went out* instead of *I went out*. Also in informal spoken English and text extracted from personal diaries subject is often dropped out as analysed by Weir [2012].

(144) ...anyone who might find it in the future and **wonder** what happened ... Post: 155852

(145) Just remembered that this is a [product-name] forum ... **wonder** if I will get a rap on the knuckles for mentioning other products and asking for info on them. Post: 180482

(146) The easiest thing to do is to run a custom file scan on the file you **think** is causing this . Post: 12332

(130) **Think** you're talking about the [product_name] ..

Scope of epistemic phrases

In some cases, first person epistemic phrases do not have a scope in-sentence. These cases happen when the epistemic phrase is an expression of uncertainty whose scope does not appear in the same sentence or does not appear at all. In (147), *We only know* is used to express reserve that the total aspect that describe a situation are not known to the writer. It does not have a particular scope attached to it, although the scope of *I don't know* in the subsequent sentence, could be slightly related.

Table 4.7 – Comparison of the frequency of hedges with an explicit pronoun subject and with an elliptical subject.

Keyword with overt pronoun	Freq.	Keyword with subject ellipsis	Freq.
i think	157	think	1
i hope	60	hope	49
i wonder	13	wonder	4
i do n't know	43	do not know	2
i do not know	8	dont know	1
i dont know	8	donno	1
i * do n't know	1	dunno	1
i do notknow	1		
	61		5
i did n't know	5	didn't konw	1
i did not know	1	did n't know	1
i idd not konw	1		
	7		2
i 'm not sure	30	not sure	19
i am not sure	16		
im not sure	2		
i * am not sure	1		
	49		
i have no idea	10	no idea	2
i had no idea	1		
	11		
i am not familiar	1	am not familiar	1
i 'm not familiar	1		
	2		
i am hesitant	2	am hesitant	1
i do n't understand	5	dont understand	1
i do not understand	1		
i dont understand	1		
	3		
-		never been sure	1

- (147) **We only know** that they are working on the patch , but when it will be released
I do n't know any exact date ... Post: 122879

While in the majority of cases the scope is found adjacent to the hedging expressions, there are some cases where this appears after some tokens such as in (148).

- (148) ... the **suggestion** is to uncheck Music and entertainment and technology.
Post: 100139

A particular case is where two or more hedging expressions can be related to the same single scope. This shows an emphasis in the uncertainty of that conditions that a proposition would hold otherwise. In 149

- (149) My guess (and it is only a guess) is that it is something to do with how ACPI [...] is implemented in the BIOS and hardware . Post: 13016

Acronyms

Some hedges in form of acronyms commonly used in web communications such as *IMO*, *IMHO* and *AFAIK* were detected in the forum corpus. *IMO* stands for *In My Opinion*, *IMHO* for *In My Humble Opinion* and *AFAIK* for *As Far As I Know*. *In my opinion* is defined as equivalent to “the way I think about it” [imo, 2002] while *In my humble opinion* has a stronger shift to conveying speaker’s opinion: “a phrase introducing the speaker’s opinion”, but not further comments about their function as speculative markers.

In the annotation dataset, *AFAIK* is found 4 times, *IMO* 5 times and *IMHO* 4 times (out of 9,193 hedge occurrences in the annotation dataset). *AFAIK* and *IMO* occur also in their lowercase versions (*afaik*, *imo*).

Profiling *I don't know*

The annotation dataset (Annset) comprises 20 lexical variations for *I don't know* as Table D.2 shows. In the Annset, 42 sentences contain at least one of these items, out of 18,491⁷ sentences and out of 5,153 containing at least one hedging marker. In the RTD there are 2,424 sentences with at least one *I don't know*.⁸

In Table 4.8 there is a list of the most frequent patterns where *I don't know* items occur in the RTD as it contains more occurrences than in the Annset. These are strict patterns, for instance, *I do n't know, if* counts only those sentences where exactly there is a *I do n't know* preceding a *if* anywhere in the sentence, but no more hedges occur in the sentence.

Examples (150) to (154) illustrate some uses of *I don't know*. Particularly, (153) and (154) show it occurring with other hedges such as *can* and *if* respectively.

- (150) Perhaps more would not have worked? **I don't know**. Post: 85103

⁷We should remember this dataset, while “clean” to a certain point is still raw in the sense it could contain non natural language sentences.

⁸No variations are considered.

Table 4.8 – Frequencies of co-occurrences of *I don't know* with other hedges in the reduced training dataset.

Co-occurrent pattern	Frequency
I do n't know, if	153
I do n't know, about	64
I do n't know, if, or	84
I do n't know, can	24
I do n't know, would	19
if, I do n't know	13
I do n't know, might	10
I do n't know, could	9
I do n't know, other	7
I do n't know, try	6
I do n't know, some	8
some, I do n't know	4
i do n't know, would n't	3
I do n't know, if, could	3
I do not know, can	3
I do n't know, something	2
I do n't know, may	2
could, I do n't know	2

- (151) There is a Synchronisation Tool but **I don't know** whether it **would** help. Post: 71532
- (152) There are many things **I don't know**, this html stuff is mostly UUUUUU to me, I don't understand what invalid html is. Post: 104565
- (153) **I don't know** what can be done. Post: 231233
- (154) **I don't know** if it 's the same, but I can make disposable ISP addresses . Post: 162637

In sentences such as in (155), the subject of *I don't know* is several constituents to the right (it is not a case of elliptical use though). *I* is subject for the clause *I'm computer challenged* as well. This was annotated as a unique keyword with separate spans.

- (155) **I'm computer challenged** and **don't know** how to deal with these things. Post: 3379

4.2.3 Syntactic hedges

Frequencies of syntactic hedging types are shown in Table 4.9.

Hedging markers such as *if* work generally as interrogatives. If it is the case they signal a hypothetical condition that was verified as explained in Section 3.3.5, in these cases, the consequent part of the conditional was marked as part of the scope of the hedging expression.

- (156) Also, please check **if** [_{SCOPE} you are still noticing the [product name] error].

Post: 148148

Keyword	Frequency
if	886
or	211
when	108
whether	48
once	45
whether or not	8
either * or	1

Table 4.9 – Frequency of Syntactic hedges in the Annotation dataset.

Or in its function as coordination conjunction for offering alternatives does not convey a speculative meaning in (157), as it does in (158). Nonetheless in the later, *maybe* is used in the same sentence, accentuating its speculative meaning.

(157) Uninstalling [product_name] v1 **or** V2, with **or** without reinstallation of either version . Post: 17941

(158) [...] maybe he would need some [_{SCOPE1} files] **or** [_{SCOPE2} logs] from there
Post: 244871

4.2.4 OTHER category of hedges

This rather miscellaneous category of hedges comprises hedging types that do not fulfill all the characteristics required by any of the other categories of Single-hedges, NCK epistemic phrases and Syntactic hedges categories. I stated some of the reasons for defining this category in Section 3.3.6. In the annotation process some candidate expressions to be annotated as hedges were categorised as Other hedges for the following reasons: a) because they are rather domain-specific, b) they are conceptually similar to hedges from the previous categories but its interpretation as hedges within these categories is not straightforward, or c) they did not morphologically or conceptually look similar to hedges from the previous categories.

The kind of hedge described in Section 4.2.2, in particular, was created to be topic-neutral as these hedges are likely to occur in any other web forum discussing a different topic. The NCK category design evolved into tuning a more standard category of hedges whose items could be spotted across various domains. In the web posts dataset, NCK epistemic phrases-like expressions that are particularly tailored to the domain of the web forum (ie. software products) are quite frequent. Expressions such as *I'm not really techie enough* in (159), *From a techie's point of view* in (160) and *I am technically challenged* in (161) are very specific to a software-related and technical domain. The same expressions are less likely to occur in other domains (e.g. fashion or politics), therefore they are annotated as OTHER hedges.

- (159) I'm afraid that **I'm not really techie enough** any more to know why this would be the case , but If you are using ... Post: 570
- (160) **From a techie's point of view**, the problem is, therefore, not [product_name]'s fault. Post: 83790
- (161) **I am technically challenged**, so be gentle, in other words very slow. Post: 7778

In general, the decision a potential hedge has actually a hedging meaning depends on the context where the hedge is inserted. In part, this context is determined by the source and hedging scope. Hedges such as *I don't get* and *it would appear* could be deemed as NCK epistemic phrases. However, they are either likely to be frequently used as a non-hedge expression such as in (162a) in comparison to (162b), or when the expression does not completely comply with the definition of NCK phrases as *it would appear* used in (163), which would be an actual NCK hedge in the form *it would appear to me*. Some epistemic phrases such as *does not know* where also included in this category as they do not seem always to be linked to a first person pronoun.

- (162) a. So **I don't get** why nothing even shows up in my account even clicking on order history shows blank! Post: 50249
- b. When I don't have this checked **I don't get** a second instance in Task Manager. Post: 257077
- (163) So, although I never heard back from [user_name] **it would appear** they fixed the problem ... Post: 232954

Hedges such as *cross my fingers* or *wish me luck* are quite colloquial but reveal an on-going uncertain situation, I deem them as emergent hedge types frequently used in informal language, however they do not fit in any of the three previous categories.

Hedges in this category have not been normalised the way NCK epistemic phrases or Single-hedges were, however they were separated in two groups noting those that resemble similar lexical syntactic structure to NCK-phrases. Table 4.10 shows a partial list including the most frequent Other-hedge types found in the annotation dataset. The full set of types and frequencies are detailed in Table D.3 from Appendix D. I have divided them into two groups which have labels: 'nck.like' for the types that morphologically resemble NCK hedges, and 'other' and for the remaining types respectively. Thus, from 209 types of Other hedges, 84 have the form of Not-Claiming-Knowledge epistemic phrase.

Table 4.10 – Raw frequencies for hedge types in the *Other* category found in the annotation dataset.

N	Tokenized expression	Freq.	Subcat. label
1	so far	30	other
2	or so	22	other

Continued on Next Page...

Table 4.10 – Continued: Raw frequencies for hedge types in the *Other* category found in the annotation dataset and their sub-categorization by phrase style.

N	Tokenized expression	Freq.	Subcat. label
3	or something	9	other
4	to be sure	7	other
5	more or less	6	other
6	or whatever	6	other
7	as * as possible	5	other
8	fingers crossed	5	other
9	at your own risk	4	other
10	it would appear	3	nck.like
11	you can find out	3	other
12	i do n't get	3	nck.like
13	from time to time	2	other
14	i am looking for a way	2	nck.like
15	in theory	2	other
16	i want to make sure	2	nck.like
17	just a thought	2	other
18	my question is	2	nck.like
19	im new	2	nck.like
20	there are times	2	other
21	10-15	2	other
22	do not know	2	nck.like
23	does not know	1	other
24	i have never done this kind of thing before	1	nck.like
25	being a newbie	1	nck.like
26	i know i sound like	1	nck.like
27	i have n't found an answer	1	nck.like
28	does n't lend itself to a quick or easy diagnosis	1	other
29	you do not know	1	other
30	can not trust it 100 % of the time	1	other
31	i do n't have a specific eta	1	nck.like
32	i need to research	1	nck.like
33	does not provide any references	1	other
34	n't * that i 'm aware of	1	nck.like
35	they do n't understand	1	other
36	without knowing for sure	1	other
37	does not increase confidence	1	other
38	nobody knows	1	other
39	their engineers are researching	1	other
40	we are working with the powerdesk engineers to figure out	1	nck.like
41	can't believe	1	nck.like
42	it is only a guess	1	other

Continued on Next Page...

Table 4.10 – Continued: Raw frequencies for hedge types in the *Other* category found in the annotation dataset and their sub-categorization by phrase style.

N	Tokenized expression	Freq.	Subcat. label
43	its hard to know	1	other
44	it is hard to tell	1	other
45	no longer has trustworthy	1	other
46	so far today	1	other
47	keeping fingers crossed	1	other
48	you did not make clear	1	other
49	by this reasoning	1	other
50	of the like	1	other
51	i finally did a google search	1	nck.like
52	i am not a techie	1	nck.like
53	i need guidance	1	nck.like
54	we 're investigating	1	nck.like
55	we are stuck on how to	1	nck.like
56	i am not that brilliant in technical side	1	nck.like
57	i have n't been able to find any solution	1	nck.like
58	keep my fingers crossed	1	other
59	you do n't say	1	other
60	doesn't give me a lot of confidence	1	other
61	i haven't tried this personally	1	nck.like
62	need to explain	1	other
63	nothing however is guaranteed	1	other
64	it's hard to say	1	other
65	no other (known)	1	other
66	cross my fingers	1	other
67	you do n't know	1	other
68	(as a second opinion)	1	other
69	it does n't know	1	other
70	from what you have said	1	other

4.2.5 Ambiguous expressions

A main problem of automatic speculation detection is the difficulty in determining when a potential hedge is not an actual hedge. I present ambiguous expressions in the annotation dataset in Table 4.11.

Keyword frequencies are compared when they occur in speculative and non-speculative senses to calculate a precision measure P_{amb} by dividing the keyword frequency when occurring as hedge by the overall keyword frequency.

In the dataset, I found 106 non-hedging keywords from which 95 have a hedging counterpart, 12 of them have their NCK phrase counterparts and 76 have Single-hedge coun-

terparts. All Syntactic hedges types are in this list, while no counterparts are found in the category of *Other* hedges. Four *NCK* keywords out of 12 *I forgot*, *we believe*, *I have tried* and *I never knew* have precision lower than the overall precision (0.63), while 28 Single-hedge types out of 76 and 4 out of 7 Syntactic hedges types have lower than the average precision. The local mean precision P_{loc_i} results from averaging the individual precisions P_{amb} within each hedge category i , for *NCK* phrases is 0.68, for Single hedges 0.63 and Syntactic hedges is 0.51. It is safe to conclude that *NCK* hedges are the less ambiguous and that Syntactic hedges are the most ambiguous items in this dataset. Expressions mostly occurring as hedges such as *I don't know* appear in non-speculative contexts when used in figurative language as in (164a) in contrast to an expression acknowledging lack of knowledge (164b) about the domain world or situation⁹. Also a keyword appears as non-hedge when it has a generic reading such as in (165a) where *sometimes* does not raise a question of determining how many times the user looks at the Event Viewer in contrast to when it is used in (165c) where it has a speculative reading. In this example *sometimes* does not change the certainty of the proposition in comparison to the unqualified proposition in (165b).¹⁰ In this proposition, the main word signaling uncertainty is *seem*.

- (164) a. **I don't know** what language one has to speak to get people in support to understand what is happening. Post: 22839
 b. **I don't know** what a LiveCD is, so I can't help you there. Post: 90198
- (165) a. **I sometimes** look at the Event Viewer, but this doesn't tell me enough. Post: 3132
 b. I look at the Event Viewer, but this doesn't tell me enough.
 c. The scan stopped and crashed at a different place this time, it seems to be different **sometimes** but then others it is the same. Post: 20693

Table 4.11 – Comparison of frequencies for hedge keywords that occur in speculative and non-speculative contexts and their precision.

N	Keyword	Nonhedge Freq.	Hedge occurrence		Precision P_{amb}
			Category	Freq.	
1	coming	31	Single-hedge	1	0.03
2	test	69	Single-hedge	4	0.05
3	part	68	Single-hedge	6	0.08
4	i forgot	7	Not-Claiming-knowledge	1	0.12
5	when	591	Syntactic	95	0.14
6	like	230	Single-hedge	40	0.15
7	beginning	11	Single-hedge	2	0.15

Continued on Next Page...

⁹See Section 1.1.2 for a description of types of knowledge in this domain.

¹⁰In this case, the “looking at the Event Viewer” is an activity that can be done constantly but done in a recurrent manner does not change the observed properties in comparison to when is done in a continuous manner.

Table 4.11 – Continued: Comparison of frequencies for hedge keywords that occur in speculative and non-speculative contexts and their precision.

N	Keyword	Nonhedge Freq.	Hedge occurrence		Precision P_{amb}
			Category	Freq.	
8	liable	5	Single-hedge	1	0.17
9	or	863	Syntactic	180	0.17
10	we believe	4	Not-Claiming-knowledge	1	0.20
11	can	747	Single-hedge	226	0.23
12	must	61	Single-hedge	19	0.24
13	about	280	Single-hedge	88	0.24
14	should	215	Single-hedge	83	0.28
15	another	100	Single-hedge	43	0.30
16	hidden	2	Single-hedge	1	0.33
17	intention	2	Single-hedge	1	0.33
18	multiple	18	Single-hedge	9	0.33
19	once	79	Syntactic	44	0.36
20	around	42	Single-hedge	24	0.36
21	would not	12	Single-hedge	7	0.37
22	doubt	3	Single-hedge	2	0.40
23	guess	4	Single-hedge	3	0.43
24	one of	46	Single-hedge	35	0.43
25	a long	8	Single-hedge	8	0.50
26	amount of	7	Single-hedge	7	0.50
27	either * or	1	Syntactic	1	0.50
28	i have tried	1	Not-Claiming-knowledge	1	0.50
29	i never knew	1	Not-Claiming-knowledge	1	0.50
30	not appear	3	Single-hedge	3	0.50
31	technically	1	Single-hedge	1	0.50
32	a lot	33	Single-hedge	34	0.51
33	could	145	Single-hedge	169	0.54
34	intend	6	Single-hedge	7	0.54
35	many	82	Single-hedge	101	0.55
36	i feel	8	Not-Claiming-knowledge	10	0.56
37	effort	6	Single-hedge	8	0.57
38	think	21	Single-hedge	33	0.61
39	other	224	Single-hedge	357	0.61
40	claim	10	Single-hedge	16	0.62
41	anyone	18	Single-hedge	29	0.62
42	appear	47	Single-hedge	79	0.63
43	most	38	Single-hedge	64	0.63
44	try	253	Single-hedge	450	0.64
45	possible	25	Single-hedge	45	0.64
46	a couple	14	Single-hedge	26	0.65
47	every other	1	Single-hedge	2	0.67

Continued on Next Page...

Table 4.11 – Continued: Comparison of frequencies for hedge keywords that occur in speculative and non-speculative contexts and their precision.

N	Keyword	Nonhedge Freq.	Hedge occurrence		Precision P_{amb}
			Category	Freq.	
48	hope	1	Single-hedge	2	0.67
49	largely	1	Single-hedge	2	0.67
50	may not	12	Single-hedge	24	0.67
51	optional	1	Single-hedge	2	0.67
52	questionable	1	Single-hedge	2	0.67
53	i do not think	14	Not-Claiming-knowledge	31	0.69
54	chance	7	Single-hedge	16	0.70
55	various	7	Single-hedge	16	0.70
56	wonder	3	Single-hedge	7	0.70
57	sort of	4	Single-hedge	13	0.76
58	if	262	Syntactic	862	0.77
59	sound	7	Single-hedge	24	0.77
60	temporarily	4	Single-hedge	14	0.78
61	not sure	1	Single-hedge	4	0.80
62	whether or not	2	Syntactic	8	0.80
63	whether	9	Syntactic	43	0.83
64	would	99	Single-hedge	491	0.83
65	based on	3	Single-hedge	15	0.83
66	i expect	1	Not-Claiming-knowledge	6	0.86
67	mostly	1	Single-hedge	6	0.86
68	a few	15	Single-hedge	99	0.87
69	question	18	Single-hedge	125	0.87
70	certain	4	Single-hedge	28	0.88
71	generally	1	Single-hedge	7	0.88
72	someone else	1	Single-hedge	7	0.88
73	i think	30	Not-Claiming-knowledge	212	0.88
74	some	52	Single-hedge	396	0.88
75	a little	4	Single-hedge	32	0.89
76	attempt	7	Single-hedge	63	0.90
77	unknown	2	Single-hedge	18	0.90
78	i believe	5	Not-Claiming-knowledge	48	0.91
79	curious	1	Single-hedge	10	0.91
80	plan	1	Single-hedge	10	0.91
81	typically	1	Single-hedge	10	0.91
82	someone	5	Single-hedge	53	0.91
83	confused	1	Single-hedge	11	0.92
84	suggestion	5	Single-hedge	80	0.94
85	something	10	Single-hedge	162	0.94
86	few	1	Single-hedge	17	0.94
87	i hope	6	Not-Claiming-knowledge	125	0.95

Continued on Next Page...

Table 4.11 – Continued: Comparison of frequencies for hedge keywords that occur in speculative and non-speculative contexts and their precision.

N	Keyword	Nonhedge Freq.	Hedge occurrence		Precision P_{amb}
			Category	Freq.	
88	may	14	Single-hedge	319	0.96
89	similar	2	Single-hedge	48	0.96
90	several	3	Single-hedge	98	0.97
91	look like	1	Single-hedge	35	0.97
92	i am not sure	2	Not-Claiming-knowledge	76	0.97
93	a bit	1	Single-hedge	39	0.97
94	sometimes	1	Single-hedge	61	0.98
95	i do not know	1	Not-Claiming-knowledge	89	0.99

Another set of candidate keywords for hedging: {*cannot believe*(2), *divided*(1), *I do not believe*(1), *promise*(5), *so far*(10), *temporary*(4), *there are times*(1), *We forgot*(1), *why*(1), *wish*(17)} were found in the annotation dataset in a non-speculative sense only¹¹.

4.2.6 Conflicting cases of hedging

Some potential instances of hedging were not marked as such because they do not have an inherent speculative or uncertainty component in their meaning, nonetheless they entail less than full pertinence to a typical category in a way similar to the one described by Hyland or their meaning reveals indirectly an uncertain scenario. *Workaround* in (166) does not refer exactly to the writer's stance about *to just go in every month and edit the backup job to change the destination after I swap the drives*, but express uncertainty related to temporary quality of the referred situation, ie. it may not always work. The proposition in (167) reveals uncertainty but it is not expressed by one word or phrase that could be used in a different domain, I reckon that *outdated version with known security issue* entails a kind of uncertainty but the subjectivity shift towards a negative evaluation of the situation. Moreover, it seems here the uncertainty emerges from the interaction of words such as *security* and *issue* that have some uncertainty-carrying meaning on their own.

(166) By the way, the **workaround** that I use now is to just go in every month and edit the backup job to change the destination after I swap the drives. Post: 2211

(167) Java is **outdated version with known security issue** ... the current ver IPIPIP also has a known security issue ... Post: 221385

These particular uses may require further study but it lies outside the boundaries of this research.

¹¹Particularly *promise* (singular form as *promises* has been found as a hedge) was found as a part of a branding denomination.

4.3 Interactions between hedging elements

4.3.1 Interaction of negation and speculation phenomena

The relation with negation is shallowly explored in this section in the sense that no deep semantic analysis is done but lexical observations about how particular hedges are affected by negation. Nonetheless, this interaction is worth exploring as it allows observing limitations of pattern matching or bag-of-words based methods for hedging detection.

Not all hedging expressions are affected by negation in the same way. For instance, while *possible* stereo-typically realizes epistemic possibility the negative polarity item *not* cancels out the potential happening sense making the phrase to carry an absolute categorical assertion. The same happens with *doubt* and *not doubt*, (168) conveys uncertainty when compared to (169). On the other hand, *not* affects in different way the hedge *may*, the negated hedge still conserves its epistemic possibility quality; the proposition in (170) express uncertainty about *need to run a full system scan* and in (171) there is still uncertainty related to this clause.

(168) I doubt this is a firewall issue as such, . . .

Post: 148148

(169) I don't doubt your experience but it surprises me nevertheless.

Post: 93789

(170) On the other side of this issue, you may need to run a full system scan

(171) On the other side of this issue, you may **not** need to run a full system scan

4.3.2 Co-occurrence of multiple hedging markers

In the Annset, the count of hedges per sentence follows a negative binomial distribution as most sentences have zero hedges as Table 4.12 shows. Restricting to sentences that have at least one hedge occurrence, 28.45% (5,261) of the total of sentences, most of them have only one hedging occurrence (18.18% of the total or 3,362). Nonetheless, 10.27% have two or more hedges which represents occurrence of multiple hedging markers in one sentence.

In the dimension of hedge categories, out of 5,261 sentences that have at least one hedge, most of them contain hedges belonging to a single category (4,194 sentences), 950 sentences have hedges belonging to two categories, 113 to three and 4 sentences have hedges belonging to the four categories.

For sentences that have more than one hedge (1,899), 50.03% have hedges from two different categories, 43.91% have hedges that belong to a single category, 5.95% belong to three categories, and only 0.21% have at least one hedge of each of the four categories. Table 4.13 shows the frequency of hedge co-occurrence types. The first column contains labels for hedge types co-occurring in one sentence, the second column shows the number of sentences that have more than two hedges with those labels, and the last column shows the frequency of sentences that have exactly two hedges with categories corresponding to

Table 4.12 – Distribution of count of hedges per sentence in the annotation dataset (Annset).

Number of hedges	Sentences	
	Count	Proportion
0	13,230	71.55
1	3,362	18.18
2	1,241	6.71
3	420	2.27
4	149	0.81
5	54	0.29
6	17	0.09
7	5	0.03
8	7	0.04
9	4	0.02
12	1	0.01
15	1	0.01

those labels. For instance the first row in this table shows that there are 771 sentences that have co-occurrence of Single-hedges, this means one sentence can have two, three or more hedges and all of them are Single-hedges, and 583 of these sentences have exactly two Single-hedges. The most frequent types of co-occurrences in sentences are the ones that include Single-hedges.

Table 4.13 – Frequencies of multiple hedges per sentence.

Categories involved	Sentences	
	Frequency	Number of hedges is 2
Sing	771	583
Sing, Synt	447	244
Sing, NCK	299	208
Sing, Oth	103	72
Synt, NCK	82	61
Sing, Synt, NCK	79	-
Synt	43	39
Sing, Synt, Oth	19	-
Sing, NCK, Oth	14	-
NCK, Oth	14	13
NCK	13	13
Oth	5	5
Synt, Oth	5	3
Sing, Synt, NCK, Oth	10	-
Synt, NCK, Oth	14	-

Table 4.14 shows that in some cases there is more than one hedging marker in a single sentence. This table also shows the seven most frequent patterns per sentence.

Some sentences that contain hedges belonging to the four types can be questioned as

Table 4.14 – Frequency of co-occurrence patterns of hedging in the Annset.

Pattern	Frequency
Single-hedge	2,247
Single-hedge, Single-hedge	562
Not-Claiming-knowledge	492
Syntactic	421
Not-Claiming-knowledge, Single-hedge	158
Single-hedge, Single-hedge, Single-hedge	143
Other	141
Syntactic, Single-hedge	137
Single-hedge, Syntactic	99
Single-hedge, Not-Claiming-knowledge	49
Not-Claiming-knowledge, Syntactic	49
Single-hedge, Other	46
Syntactic, Single-hedge, Single-hedge	43
Syntactic, Syntactic	36
Not-Claiming-knowledge, Single-hedge, Single-hedge	35

being product of ill-sentence splitting such as in (172) and (173); each one should have being split into multiple sentences. This is the same case with sentences that have 12 and 15 hedges (Table 4.12).

- (172) Bottom line, i don't know how to remove this thing and would appreciate any advice in doing so - I'm not TOO techy so try and keep the instructions a little on the basic side if possible - THANKS! Post: 19586
- (173) I suspect that Ghost 15 might be creating a conflict between the two boot sectors or a conflict between the drive signatures, and that conflict might be a potential cause of the source drive losing something in its boot sector, but this is only a theory on my part. Post: 190566

The occurrence of multiple hedges in a single sentence can be explored from the point of view of embedding. For instance, *would* and *think* are speculation markers on their own, however we can find that both markers are likely to appear contiguous in a sentence as in example (174).

- (174) I would think that they are not good friends for a few reasons.
Post: 181653

4.4 Profiling Other elements of hedging

4.4.1 Source

In Section 3.3.2 some evidence about the Inner Epistemic Source not appearing in sentence was observed. This can happen because this is attributed implicitly to the writer or because

the referent for the Source is in some other sentence in the document.

The overall distribution of explicit source across the four types of hedges is shown in Table 4.15.

Table 4.15 – Overall distribution of Non-Explicit and Explicit Source mentions per hedge category.

	Non-explicit	Explicit	Explicit per category
Not-Claiming-knowledge	1.25% (100)	11.66%(933)	89.80%
Other	2.46% (197)	1.41% (113)	36.62%
Single-hedge	58.72% (4,697)	9.09%	(727) 13.14%
Syntactic	15.38% (1,230)	0.03% (2)	0.16%

The cases of Syntactic hedges with explicit source are very few, amongst them cases where is a reported speech as in (175). The proposition in example (176a) which was not considered as a quotation because it is the writer who describes a hypothetical reported speech, it can be paraphrased as (176a).

(175) One further note: [_{SOURCE} Somebody else] said *that the scan will not work if nobody is logged in*. Post: 2505

(176) a. It creates a situation for [_{SOURCE} the user] of, *I wonder if my files are being backed up when I have the file set created correctly?* and that's not a good thing with a backup solution. Post: 296614

b. It creates a situation where [_{SOURCE} the user] wonders **if** his files are being backed up and when he has the file set create correctly. And that's not a good thing with a backup solution.

NCK phrases represent the largest group of hedges having an explicit source or perspective. The cases where it is omitted is due to elliptical subject use; Table 4.16 shows two highly frequent NCK phrases with overt source: *hope* and *not sure*, the next items following in frequency higher than two are *don't know*, *wonder*, and *didn't know*. The remaining hedges have either 1 or 2 occurrences.

When the Source of hedging is not the Writer

'Other' as a value for Inner-Epistemic-Source is given in cases where the writer comments about the mental state of other individual or entity. Some cases where this occurs is for instance when the writer reports a hedged speech as in (177). In this proposition, *probably* does not reflect the writer assessment about the issue, but he or she communicates their (tech support) assessment which is speculation about the issue in which the writer is involved.

(177) I spent 4 hours with [Org-name] tech support and [_{SOURCE} they] said I **probably** had a virus . Post: 107577

Table 4.16 – Not-Claiming-Knowledge epistemic phrases with non-explicit Source and their frequencies in the annotation dataset.

NCK phrase	Raw frequency	NCK phrase	Raw frequency
hope	46	do not remember	1
not sure	18	do not understand	1
do n't know	6	don't know	1
wonder	4	don't recall	1
did n't know	3	don't understand	1
do not know	2	dunno	1
guess	2	forgot	1
no idea	2	never been sure	1
am hesitant	1	still can't understand	1
am unsure	1	think	1
can't figure out	1	wondering	1
can't seem	1	wonders	1
donno	1		

Table 4.17 – Distribution of Inner Source attribution in the annotation dataset showing percentages in-group per type of Source.

Source	Hedge type	Inner Source		
		Other	Writer	Other per category
Non-explicit	Not-Claiming-knowledge	0% (0)	1.27% (102)	0%
	Other	0.11% (9)	2.36% (189)	4.55%
	Single-hedge	1.49% (119)	57.16% (4577)	2.53%
	Syntactic	0.2% (16)	15.18% (1216)	1.3%
Explicit	Not-Claiming-knowledge	0.01% (1)	11.7% (937)	0.11%
	Other	0.24% (19)	1.17% (94)	16.81%
	Single-hedge	1.5% (120)	7.58% (607)	16.51%
	Syntactic	0.02% (2)	0% (0)	100%
		286	7,722	

In most cases, every time the writer uses an uncertainty expression it is he or she who comments about his or her own certainty. Nonetheless, although in a smaller proportion, hedges have a Source that is different from the writer. Proportions per hedge type where this is the case are shown in Table 4.17, which shows the distribution of hedges that have their Inner Source attributed either to Other or Writer.

Comparing the proportions between hedges whose Inner Source is Other or the Writer, it follows that in all the cases there is a larger percentage of hedges whose Inner Source is attributed to the writer; only for NCK phrases and Other hedges this difference is not significant ($p > 0.05$). The largest group of hedges that has Other as Inner Source is the category Single-hedges where the source is non-explicit. The only case where a NCK phrase has Other as inner source and explicit source was shown in example (175), it happens in the

context of a hypothetical reported speech. In this case, the source corresponds to *the user* since it is not clear that the uncertainty attributed to the NCK phrase *I wonder* reflects the writer's mental state.

This example also shows a Syntactic hedge (*if*) that has *the user* as Inner Epistemic Source. On the other hand, (178) shows *if* as a Syntactic hedge which has *Other* assigned as its non-explicit Inner Epistemic Source.

- (175) It creates a situation for [_{SOURCE} the user] of, ***I wonder*** *if my files are being backed up when I have the file set created correctly?* and that's not a good thing with a backup solution. Post: 296614
- (178) I received this message : **Warning: If** you currently have the Add-on Pack installed, you need to save your settings before installing [product_name1] 2008.
Post: 12045

Table 4.18 – Frequency of single hedge occurrence that have 'Other' as Inner Source.

Hedge	Frequency	Hedge	Frequency	Hedge	Frequency
suggestions	31	attempt	1	intended	1
suggested	15	claim	1	may be	1
suggestion	15	confused	1	may not	1
question	11	confusion	1	might	1
questions	9	confusion	1	might not	1
trying	4	could	1	other	1
can	2	'd	1	possible	1
suggests	2	doubt	1	proposed	1
temporarily	2	effort	1	should	1
tried	2	efforts	1	some	1
another	1	guesses	1	someone	1
appear	1	intended	1		

- (179) and welcome any **suggestions** for even for further performance improvements.
Post: 222674

- (180) Under General Issues it says When configuring a backup job to start a new recovery point set , be sure that you do not schedule the new set to start when the backup job is scheduled to create an incremental recovery point. If you do, the backup job scheduled to create the new recovery point set **might not** run as expected

- (181) Research on the Internet **suggests** this is quite a commonly known piece of spyware known as indt2.sys

4.4.2 Scope

Most of the literature about hedging does not put emphasis on explaining the nature of a hedge's scope. Some studies provide loose definitions or put forward the importance of an accurate detection of in-sentence scope of hedges for more sophisticated natural language processing tasks such as machine translation.

Marking of speculation's scope in academic articles as in Bioscope is said to be determined by the grammatical structure of the sentence and that scope of verbs, auxiliaries, adjectives and adverbs usually extends to the right of the hedge expression [Vincze et al., 2011]. However in the forum annotation dataset there were some borderline cases where the real hedge scope was not explicitly evident in the sentence. One of these cases is the verbal form *suggested*, it could be argued that the left clause *I tried the [product_name]* corresponds to the scope of this hedge in (182), by finding a similarity in the rephrased proposition (183). But this is not straightforward to notice in (182). In this case *suggested* was addressed either to this user or to another one that requested help in a similar event.

(182) I tried the [product_name] that some **suggested** already to no avail , but found mention of [website_name]. Post: 6256

(183) someone **suggested** to try the the [product_name]. Post: 96873

I do not intend to fill that void but highlight some cases that may differ or characterise various hedge realisations. Scope for Syntactic hedges comprises two segments determined by the structure of conditional. As elaborated in section 3.3.5, syntactic markers are considered hedges because often they introduce hypothetical situations that highlights one aspect of uncertainty about the world.

Out of 8,005 hedge occurrences,¹² 81.75% of hedge occurrences have scopes corresponding to them. This means only 18.26% (1,496) do not have a scope associated. Single hedges are the most frequent category that do not have a related scope. In Table (4.19), Single hedges without scope whose frequency is greater than 3 are shown. The types are shown sorted decreasingly by relative frequency, therefore the top elements read from left to right show those types whose ratio of occurrence without scope to the total number of occurrences is 1. These types do not appear to be modifying in terms of uncertainty any particular clause or constituent within the sentence scope. Within this subgroup, particularly it could be argued that *strangely* should take as a scope the forward clause in sentence, e.g. in (184), however this clause conveys a fact and does not transcend at the same level as a similar proposition in (185). Therefore, there is no scope associated with the occurrence of *strangely*. The scope for *I am intrigued* in (186) could be argued to have the same nature as the scope for *strangely*, however content of the forward clause in the sentence helped to decide that *by your reports* was affected by NCK phrase.

¹²These hedge occurrences are not in the category of sub-hedges, that is hedging expressions within the boundaries of another hedging expressions.

Table 4.19 – Single hedge types without related scope, their frequency as such, their overall and relative frequencies. Shown hedge types are for those occurrences whose frequency is greater than 3 and sorted decreasingly by relative frequency (to be read from left to right).

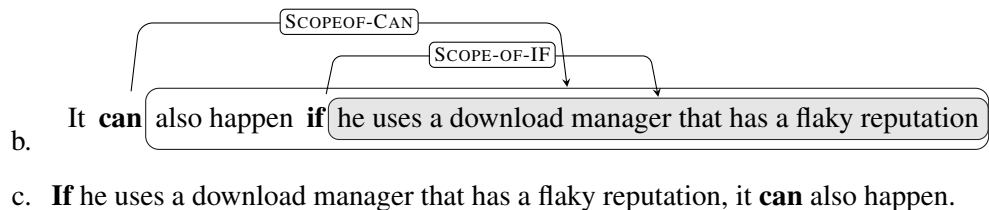
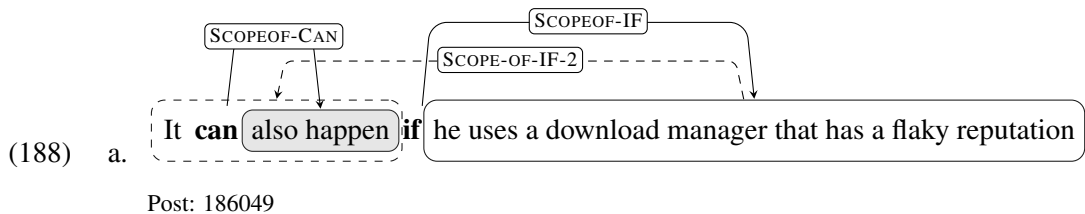
Single hedge	Freq.	Total Freq.	Rel. Freq.	Single hedge	Freq.	Total Freq.	Rel. Freq.
based on	15	15	1.000	confusion	9	9	1.000
somebody	8	8	1.000	someone	53	53	1.000
someone else	7	7	1.000	something else	7	7	1.000
strangely	4	4	1.000	something	143	144	0.993
question	74	75	0.987	anyone	28	29	0.966
others	49	52	0.942	questions	47	50	0.940
suggestions	45	48	0.938	somewhere	10	11	0.909
confused	9	11	0.818	suggestion	23	30	0.767
a while	14	19	0.737	chance	8	13	0.615
curious	6	10	0.600	confusing	6	12	0.500
odd	5	10	0.500	somehow	6	12	0.500
something like	4	9	0.444	suggested	21	49	0.429
strange	10	24	0.417	sometime	4	10	0.400
attempt	10	27	0.370	possible	15	45	0.333
think	5	16	0.312	try	53	175	0.303
tried	35	147	0.238	unknown	4	18	0.222
like	8	40	0.200	can	34	226	0.150
most	6	45	0.133	trying	13	111	0.117
perhaps	6	57	0.105	another	4	43	0.093
may be	6	71	0.085	might	11	133	0.083
some	30	396	0.076	should	6	83	0.072
could	12	169	0.071	many	7	101	0.069
about	6	88	0.068	may	10	155	0.065
other	18	305	0.059	would	26	441	0.059
a few	4	98	0.041				

- (184) **Strangely** though it still triggers the mscowrkd.dll can't be loaded/is missing message. Post: 251167
- (185) **Perhaps**, [_{SCOPE} it still triggers the mscowrkd.dll can't be loaded/is missing message].
- (186) **I am intrigued** [_{SCOPE} by your reports], although I am not clear from your responses to questions as to what you have exactly identified at this point. Post: 155909.

Regarding cases of hedges that occur within the scope of another hedge in the sentence, 16% of hedges co-occur in this way such as the cases shown by (187) and (188a). For instance in example (187) *could* is within the scope for *perhaps*. In (188a) the main clause of a conditional encloses the hedge *can*. Example (188b) shows an alternative tagging for the scope of *can*, however it can be observed that its correct scope is *also happen* if this proposition is rephrased as in (188c).

While in this research a deep analysis of co-dependence of hedging markers is not done, it is an important feature to analyse as further hypothesis could be done about the meaning of hedges that are modified by other hedges. In the sentence presented both hedges preserve their quality of conveying uncertainty, nonetheless presence of multiple hedging marker in

a sole proposition could increase the degree of uncertainty. .



Dissecting in-scope hedges by category, 1.35% of NCK phrases fall into other hedge's scope, 5.79% of Other hedges, 17.21% of Single hedges and 23.76 % of Syntactic hedges that have the largest percentage of in-scope occurrence. This is explained in part by the interrogative functions of conditionals such as *if* and *whether*. Another obvious reason is that Syntactic and Single hedges do not appear often in a phrasal form as compared to NCK and Other hedges, they are single parts of speech such as verbs, nouns, adjectives and adverbials that are more likely to be embedded within another clause in order to perform their hedging function in contrast to phrasal hedges. The low percentage of NCK phrases falling in another hedge's scope also reveals their use as stance conveyors, as they are not often embedded into other clause, they accomplish the function of introducing an opinion, in this case to assert tentativeness and lack of enough knowledge. On the other hand, from all hedges in-scope, Single hedges is the largest group (74.17%), followed by Syntactic ones (23.29%) while the shares for NCK and Other hedges is marginal (1.11% and 1.43% respectively).

4.5 Hyland's pragmatic model in web forum dataset

Following the discussion about Hyland's [1998] pragmatic categories introduced in Section 2.4.4, this taxonomy was created to reflect the purposes of hedging in the settings of academic texts as a form of communication within a research community. This academic community shares some characteristics with a web forum community in that it refers to interests when communicating by means of written text. This form of communication entails the presence of a writer, readers, the written text itself and a topic. The intentions of

web forum writers that frequently contribute to the forum can be compared to some extent with academic writers' ones: being acknowledged as valuable individuals in the community which subjects them to its assessment. This is what motivates a continuous participation of writers in a web forum in contrast to occasional visitors or lurkers. The concepts of Advice Seeker, Advice Giver, Commenter, and Facilitator described in the introductory chapter of this research are brought up to attention in this analysis of how hedges in a web forum are used according to this pragmatic model of hedging. The similarity in intentions and communication elements suggested an application of the most specific pragmatic categories of hedging proposed by Hyland: Attribute, Reliability, Writer-oriented and Reader-oriented hedges to texts extracted from web forums.

Attribute hedges are realized in web forums by their use for making assertions with accuracy, shifting from absolute categorical assertions. Users include attribute hedges in their text with the intent of describing accurately the problem they are seeking advice for, or to limit the extent a given solution would apply. For instance, in posts starting a thread in the forum a user stating (189) describes a problem where an expected normal behaviour (*The site builds up*) is hedged to depict the context the problem comes up in. In (190), *some* and *a few* are used to describe in detail various behaviors occurring instead the ideal single expected behaviour where the *startup* should happen without incidents. Although it may be thought these hedges are directed to readers in that they help to understand the reported circumstances more straightforwardly, they are not solely used with that main concern but they are concerned with a precise explanation, the content orientation here is emphasized. Moreover, in giving advice, *about* is used in (191) to precisely inform of a numerical estimation as it might be unknown to the writer, *about* is commonly used to describe this kind of approximation.

A common mechanism used by individuals for providing solutions different from straight advice is by recounting their own experiences in dealing with a similar situation. In this kind of statement, hedges are used in a similar way to when describing a problem in that they hedge to accurately depict an event circumstances, e.g. in (192), *more or less* can be thought of as being used with the same purpose as *about* in (191) to approximate a numeric and in this case a temporal feature.

- (189) The site builds up **partly** only, when searching for videos I get a blank screen.
Post: 1369
- (190) ... **some** startups will return to normal after **a few** minutes , **some** just freeze the OS, requiring a hard boot. Post: 8566
- (191) ... you could have bought the 5 PC version from [company_name] itself for, here, **about** \$20 per seat. Post: 216510
- (192) I have 4 packs of these numbered 6,7,8 and 9 and use them in sequential order. One pack every 3 months unless I'm doing something special and take a copy **more or less** frequently. Post: 68240

- (193) We have developed a fix for the issues that **some** users experienced installing the [product_name2] beta plug-in after receiving the latest [product_name] patch . Post: 46453

Forum threads are also initiated with the purpose of announcing news, in this case not only a problem might be described but also a solution that does not address a named issue (i.e. answering to a post that requests advice), but that refers to long standing issues or improvements for a specific situation (e.g. upgrading products' characteristics). In (193), a writer making an announcement, uses a hedge to limit the extent that the solution applies (in this case it is directed to users who experienced a specific problem in contrast to all users), making his or her statements more accurate.

Reliability hedges emphasize in providing an accurate depiction of a situation by assessing how reliable is a statement's truth. Often these hedging devices correspond to those in the category of Single-hedges as in (194) and (195). A user suggests a seeming cause for an issue by hedging his or her statement as in (194). Then, a solution can be suggested under the stated assumption. Users also describe a problem and point to their intuitions on how the problem could be solved. However, as they do not possess and the circumstances do not allow them to gather enough knowledge to assert a categorical statement they use hedges to express this limitation while they ensure their claim is accurate. Example (195) shows the use of *possible* and *could* to express perceptions of this sort.

Expressions of lack of knowledge related to the circumstances (knowledge about the situation) are reliability hedges such as *We will investigate* in (196). It does not assert that the user does not know because lack-of-knowledge about the world but due to the circumstances (they just got aware of the situation at the point) and any possible temporary suggested solution has limited validity until the real cause has been cleared out. Therefore, the user's intention is to give reliable information while he or she cannot yet give a definite solution. This kind of NCK epistemic phrase seems to belong to the category of reliability hedges but they would have to be assessed on a case-by-case basis to find out the whole range of NCK phrase types that show similar characteristics.

- (194) It is **possible** that you got a bad download of the latest update that just came out recently. Post: 201667
- (195) How do I temporarily disable [product_name] so I can play? I don't want to uninstall. Another **possible** solution **could** be if [company_name] ever released a patch that removed the conflict with [videogame_name]. Post: 9550
- (196) **We will investigate** a fix. Post: 227866

Reliability hedges in the sense they reflect "I do not speak from secure knowledge" might look they have a correspondence with NCK epistemic phrases of the type *I don't know* or equivalent, however, these often do not emerge from limitations of the conditions around a phenomenon that prevent an individual from categorically asserting a statement, but from limited knowledge about the domain world in the fact they do not possess enough expertise to point to a reliable solution. Advice-seekers submit their questions to the forum

because themselves lack the knowledge to address their issues, should they have figured out a solution it would be less likely they post details of the event¹³. In statement (197), the writer is quite sure about what he or she is saying, however by using “From what limited knowledge I have” is not trying to make his or her assertion more accurate but looks for acceptance by likely readers. The expression is therefore a reader-oriented hedge. Also, it may look like the hedge is a writer-oriented one but the writer’s involvement reflect his or her main motivation is not face-saving.

(197) **From what limited knowledge I have** about these things, there isn’t going to be possible to actually have a program set up that can actually clean the machine and at the same time fix the machine so it can continue to be working. Post: 182067

It is quite hard to map the category of writer-oriented hedges to types according to the hedge categorisation proposed in my study. Writer-oriented hedges are often impersonal as writers want to limit their personal commitment. Particularly in academic writing from where writer-oriented hedges category originated,¹⁴ style guidelines often advise avoiding personal stance using first person subjected statements in favour of using third person or passive voice constructions, for instance:

[...] most academic analysis focuses on the subject matter rather than on you as you respond to it. If you use the third person, you keep the attention where it belongs. [Rosenwasser and Stephen, 2011]

In web forum posts’ writing, these stylistic limitations do not take place. Most of the times, NCK epistemic phrases include a subject, so they ensure writer’s commitment. Some expressions in Other hedge category such as *its hard to know* in (198) can be deemed as writer-oriented hedges. While it conveys uncertainty, the hedging expressions limits writer’s involvement.

(198) And only after they were trying to log into the wireless network...
... and when you see those [00:13:BC:21:X:X], **its hard to know** who is who!

Hyland frequently acknowledges the difficulty of discerning to which of these categories a particular hedging type may belong. Particularly, this occurs in the distinction between writer and reader oriented hedges, which he maps to Fraser’s terminology of “self-serving acts” for writer-oriented hedges in contrast to reader-oriented hedges deemed as “altruistic acts” [Fraser,1980]. Despite the obvious distinction between two types of acts, in web forum statements they co-occur and the extent one device is affected by the other one is not easily established. For instance, the commitment provided by a hedge such as *My guess* in (199) and therefore acting as reader-oriented is weakened by *it is only a guess* that reveals a writer-oriented hedge. Besides the difficulty in differentiating to which kind of individual is the

¹³Although is not uncommon that users tell about their experiences in solving a problem with the hope other user can benefit from that.

¹⁴Hyland provides more insights on the category of writer-oriented hedges and how state of the art research in academic writing is related to them [Hyland, 1998, p. 170].

hedging intention directed, there is the difficulty to see whether hedging usage is focused towards enhancing the content or towards protecting a writer's reputation. Although at a different level than sentential one (discourse level), *makes no warranty that* in (200) reveals an uncertain situation; its use in the statement beyond trying to avoid writer's commitment by referring to the company as the subject of this assertion can be thought of coming closer to the category of reliability hedges as the truth of aforesaid statements is subjected to an uncertain situation. This hedged statement makes the whole set of statements a more accurate account of what should be expected by forum users.

- (199) **My guess** (and **it is only a guess**) is that it is something to do with how ACPI (Advanced Configuration and Power Interface) is implemented in the BIOS and hardware.
- (200) Without limiting the foregoing , [company_name] **makes no warranty that** (i) the community website or the services will meet your requirements ; (ii) the [product_name forum_name] website or services will be uninterrupted ...

I mostly deem NCK epistemic phrases as reader-oriented hedges as they are subjective, they mostly include first person pronouns. Thus, they express writer's personal alignment to the statement they are contained in. They try to prevent criticism by softening assertions with the expectation of getting acceptance from readers. Users coming to the forum as Advice Seekers express their lack of knowledge about a particular issue, what spares them from criticism from possible blunders when describing their issues, as a person with less experience is expected to be. So this admission of limited knowledge is an assurance of "I have no knowledge about this and whatever I might say is bound to be wrong, so do not be harsh on me", however this does not have the purpose of protecting his or her reputation as with writer-oriented hedges, nor this admission of lack of knowledge is related to limitations of the phenomenon being described which would make them reliability-hedges. In this sense NCK phrases in web forum documents dissent from reader-oriented hedges in the way they were conceived in research articles writing: there is no danger for users who just seek a solution in admitting non-knowledge either partially or completely about a particular issue.

Sentences (201) and (202a) are used by Advice-Seekers in posts starting a thread. *I am totally lost* and *I'm not sure* in those propositions are committed assertions of uncertainty, they mark the hint for other user to make suggestions providing advice or potential solutions. In this sense, these NCK phrases take the role of interrogative sentences even in the case of appearing on their own in a post. Hyland draws attention to questions as devices for involving the reader on the matter being discussed as they convey tentativeness while asking for an answer. However, the contribution of questions to the expression of speculation or uncertainty has not been further explored in this research.

In answers to advice-seeking posts, hedges such as *I'm not sure* are also used to express perplexity before unforeseen circumstances while maintaining personal stance (202b). If conversation between Advice-Seekers and Advice-givers is seen as a cooperative process

of knowledge discovery, hedges act to promote cooperation in contrast to the usage of categorical assertions that do not invite to further contributions or that can initiate flames when disagreement occurs (202c).

- (201) **I am totally lost** when it comes to back up.
- (202) a. **I'm not sure** where to post this. Post: 7170
- b. **I'm not sure** why you can't see unallocated space from your [product_name] as that's how people successfully clone in this forum. Post: 95447
- c. ... as you can see I have number of issues here and at \$90 each **I'm not sure** I can afford the pay additional service ... Post: 239436

It was mentioned earlier in this section, the case when users take the role of Commenters to recount anecdotal situations of solving a problem in the hope other users can benefit from this information. Reader oriented hedges are used in this type of post to improve chances of acceptance as a hedge such as *My guess would be* in (203) precedes a suggestion to a question while keeping personal attribution. The question is raised by himself or herself as communicative device and while he or she provides a seeming solution, the question keeps open to contributions on finding the real cause for the stated problem. Similar hedges are used in a reader-oriented style in (197) and (204) to make statements either by user posing as Commenters or Advice-givers.

Acronyms as NCK phrasal types are also used to show judgement according to personal beliefs. *afaik* in (205) embeds a declaration of knowledge within boundaries of personal responsibility for what it is being said. These boundaries contrast with a situation where absolute knowledge could be declared that is more likely to be rejected by a reader.

Agglomeration of NCK phrases as in (206) reinforce the goal of regarding readers' opinion. Personal involvement takes place where impersonal constructions could be deemed as outright criticism. Particularly *I wonder* softens a request as compared to the impression an imperative could make on an Advice-Seeker.

- (203) Why didn't the Boot disk find the problem ?
My guess would be that there are files that it does n't check that would be checked if the anti-virus software was run from Windows. Post: 283636
- (197) **From what limited knowledge I have** about these things, there isn't going to be possible to actually have a program set up that can actually clean the machine and at the same time fix the machine so it can continue to be working.
- (204) **My suspicion is** that [product_name1] needs to ensure all its services are started at the earliest possible opportunity ... Post: 177599
- (205) The Explorer is however, **afaik** , unrelated to QQ , and I cant see why Explorer should be attempting to access [product_name2]. Post: 226057
- (206) **I am intrigued** by your reports , although **I am not clear** from your responses to questions as to what you have exactly identified at this point . **I wonder** if you would be prepared to check out something . Post: 155909

Because web forum's language style is more informal than in academic writing, there is no pressure on describing accurately a given situation as in (207). Nonetheless, in this kind of situation accuracy in descriptions is desirable for the user's advantage. Thus, this scenario prompts to further questions to get a better understanding of the problem underneath in the search for an answer that can address properly the issue being communicated.

(207) It stops at **one of** the windows splash screens and just sits there. Post: 3099

One of the main *raisons d'être* of web forums is providing a means of users' communication of uncertainty. This is communicated through web forum posts where questions are put forward and situations causing uncertainty about the world are described. In contrast, research articles mainly serve the purpose of answering (specific) research questions. Although in research articles subsequent questions can be formulated as research outcomes, their main purpose is the communication of conclusions which authors have certainty about.

4.6 Conclusion

The main intent of this chapter was the empirical description of hedges found in the annotated dataset extracted from the web forum under study. This description was restricted to empirical and theoretical findings about the linguistic elements involved in a hedging event.

Since the annotation of hedges is done according to their occurrence in each sentence, I have described the dataset in terms of which types of sentences are relevant for the analysis of hedges. Only one kind of declarative sentence that I named 'processable' is to be considered for analysis of hedges. Other types such as interrogations, quotations and non-processable sentences were not deemed apt for different reasons. Particularly, I showed that the case of quotations was important to identify since they represent content that was written by a user that is not the original post's writer and therefore in those cases, the Source of the hedge used in the quotation would not reflect the writer's point of view. Interrogative sentences have an inherent representation of uncertainty but I did not include them in this study as it may need further analysis that is outside the boundaries of my research.

In terms of frequent categories of hedges, Single-hedges are the most frequently occurring in the annotation dataset and approximately the most frequent also in the RTD although it has to be considered that counts of hedges in the RTD are tentative as not all occurrences may be actual hedges. The next most frequent types of hedge is Syntactic, followed by NCK and finally Other-hedges. Looking at the whole set of hedge occurrences in the Annset, it can be observed that they are quite spread since the overall and per category distribution of hedges reflects a negative binomial, filled with zeros distribution.

Apart from the above, I have described a lexicon of hedges comprising words and phrases used for speculation and other hedging functions. Lexical hedging types belonging to four categories of Single hedges, Not-Claiming-Knowledge first person epistemic phrases, Syntactic and Other hedges were presented and described separately.

Overall I found 790 unique types of hedges, 272 of them belong to the Single hedge category, 300 to NCK phrases, 8 to Syntactic and 209 to Other hedges.

I described some normalization techniques mainly for Single hedges and NCK epistemic phrases which could be explored in future endeavours for formulating strategies for detecting hedges in user generated content particularly. Further discussion about potential uses of this lexicon and its features is outlined in Chapter 6.

Some normalization techniques reduce the number of lexically extended epistemic phrases to primary type in the case of NCK and in both, Single and NCK categories normalization causes grouping of equivalent types that were lexically different because of typos, tense and number variations, abbreviations, non-standard forms and colloquialisms. Single hedges were normalized from 270 to 189 types and NCK were normalized from 303 to 138.

Single hedge types reflect mostly what is found in literature and in previous hedging studies, Table D.1 shows the complete list of these types.

I put emphasis on discussing the relevance and conception of NCK phrases as lexical units which exhibit substantial differences with hedges originated in the epistemic modality tradition.

The variety of realisations of NCK epistemic phrases has been shown (Table D.2) and some particular cases such as ellipsis in NCK hedges (eg. *hope, not sure*) have been discussed as the findings in the dataset suggest to be a trending feature in web media and particularly in user generated content. Other underlined types of hedges are acronyms and variants of *I don't know*.

I have continued the discussion started in Section 3.3.4 about the distinction between subjective and objective epistemic modality, by showing with lexical findings of what seems to be a claiming-knowledge component in Not-Claiming-knowledge phrases such as *I don't know*, and comparing them overall to distinctions between categorical and hedged assertions. With respect to this point, I conclude that in NCK phrases, the focus of the interpretation of hedging is divided between the source and what is being hedged, in contrast to Single-hedges where only what is being hedged is under scrutiny. This particular feature of NCK phrases emphasizes what has been suggested in the literature about their difference from other types of epistemic expressions. Moreover, empirical findings reinforce the idea that first person epistemic phrases is a distinctive semantic category of hedges.

The two remaining categories of hedges were less extensively addressed in this chapter since either their types are quite regular (Syntactic) or quite heterogeneous (Other). However, I have found that *if* is frequently used as a speculative marker.

The group of Other-hedges is mainly composed of NCK-like epistemic phrases but whose content is tailored to the domain of the dataset under study, for instance *I'm not really techie enough* and other miscellaneous types. However, I have suggested they could be built into patterns taking into account terminology from the domain where they would be analysed.

Lexical types that would potentially convey a hedging meaning but were not actually being used as such (ie. non-hedges) were also analysed as ambiguous types of hedges. I have

shown that NCK phrases have fewer ambiguous occurrences compared to Single-hedges which suggests they can be used as improved types of hedges that convey less ambiguity and can be used in datasets from other non-explored domains since they would require less automatic natural language processing resources such as parsers. Parsing and similar linguistic tasks are quite accurate in formal styles of language but they still have challenges to overcome in language styles that are noisy.

Co-occurrence of hedges in each sentence was measured: 10,27% of the sentences have at least two hedges within their boundaries. The most frequent co-occurrence is of two Single-hedges per sentence and the most frequent combination of two hedging categories is where one Single-hedge and one NCK appear in one sentence.

In accord with their designed lexical types, it was found that most NCK hedges have an explicit Source (89.9% of occurrences) in comparison to Single-hedges (13.14%). As expected, the source for 99.99% of all NCK occurrences is the writer, while for Single hedges, 2.53% of explicit Inner epistemic Sources is attributed to another individual that is not the writer, and in 16.51% of the cases when this source is explicit in the post. The most frequent hedge types whose source is not the writer are variations of *suggest*: [*suggestions*, *suggested* and *suggestion*]

Further, I have pointed out some possible caveats in the manual identification of the scope of hedges, such as when the scope is not evident in the sentence or when there is the possibility of attributing a hedge a scope that actually does not correspond to it. I have found that 18.26% of hedge occurrences do not have a scope in-sentence, being Single-hedges the ones that have the highest frequency, for instance *based on*, *somebody*, and *strangely*. I have also identified cases where a hedge scope comprises another hedge. These findings could be further explored in the sense of studying interactions between hedges subordinated other hedge types.

I have provided numerical descriptions of source and scope that illustrate the variety of hedge realisations in this informal style of language. Regarding the scope of hedges, it was shown and discussed how this is not solely determined by syntactic features in-sentence but by the semantics of certain hedge types.

Finally, I have provided linguistic examples to discuss how the pragmatic categories proposed by Hyland match the function of hedges in the domain under study. I compared the intentions of academic writers with the ones from forum contributors, emphasizing these are different from occasional visitors.

The main pragmatic categories of hedges analysed are: Attribute, Reliability, Writer-oriented and Reader-oriented hedges. I have described frequent specific situations where hedges are used and could be matched to these categories. For instance, attribute hedges are frequently used to make accurate descriptions of problems that make users seek answers in the forum. Some types of NCK phrases are used in ways that could match reliability and reader-oriented hedges, particularly in the latter when they look for reader's acceptance. One striking difference that leads hedges into particular categorization is that in research writing, authors are discouraged of overusing first person with the intent that focus remains

on the research topics; such limitation do not exist in web forums, so it could not be said that in web forum writer-oriented hedges are used as often as in academic articles. I believe that individuals in the web forum use NCK hedges to prevent criticism, and therefore these are representative of reader-oriented hedges. However, in some senses they are not equivalent since many users seeking advice in the forum are not afraid of admitting lack of knowledge.

I have also emphasized the user of hedges in comparison to categorical assertions in cooperative tasks between Advice-seekers and Advice-givers. I noted as well that individual using hedges in some cases make imprecise descriptions as they do not think the situation needs to be accurately described.

Nonetheless, I have pointed out the difficulty connected to matching hedges from specific categories proposed in my study with those proposed by Hyland. Hyland himself discusses the difficulty related to placing specific hedge types into any of his pragmatic categories as many would change their function according to context.

Considerations about lexical findings in each hedging category are useful for the building of automatic models for identification in this kind of domain. However, I believe this discussion only scratches the surface of the nature of limitations of the identification of hedges in informal language.

Chapter 5

Hedges and forum features interaction

Analysis of hedges based on their presence in fragments of text, on its own, will hardly serve to get an integral overview of how hedges are used in a particular domain. The domain of the dataset used in this research is represented by other longitudinal features built around individual users or around individual posts. A small set of features was selected as minimal units for domain characterisation in order to aid finding answers to the research questions: how is uncertainty expressed in web forum texts? and, is the expression of uncertainty in this domain related to other post categories? Reputedly, web communities such as web forums aim at creation and sharing of knowledge where trust in terms of users' competence and user's benevolence has been shown to be the main determinant of desired participation in these communities [Abrams et al., 2003, Paroutis and Saleh, 2009]. Hedges in part account for determining user's benevolence, therefore it is interesting to explore other features that seemingly represent benevolence and competence in the web forum under study in particular.

In this chapter I will describe findings about the interaction of hedges and other dimensions of characterisation in forum posts. The main assessed dimensions are a) forum structure in terms of user's categories hierarchy, b) ratings given to posts by users and c) emotions conveyed by users. Posts are categorised by concurrence with each of these dimensions and their co-occurrence with hedges is analysed to discover correlations. In the first instance, these analyses are done over the annotation dataset and for the sake of comparison and obtaining meaningful insights some analyses are extended to the reduced training dataset (RTD).

I will describe web forum posts in terms of these three features: user categories, ratings and signals of emotion. In Section 5.1 it will be described how user categories are organized according to the web forum hierarchy and how they were abstracted into a more simplified categorization. In Section 5.2, the concept of kudos as representation of ratings given to posts will be explored. Simple features that account for the expression of sentiment polarity in posts are described in Section 5.3. Various strategies for post categorization based

on hedge occurrence are described in Section 5.4. Post categories labels thus allowed for exploring correlations to the aforementioned posts features. Results, analysis and discussion of statistical models resulting from these correlations are presented in Section 5.5. To conclude, I summarise the main findings in Section 5.6.

5.1 User categories

User categories are pertinent to characterise users in terms of a particular or various desirable features depending on the web community's nature. In general web-based communities promote and reward various kinds of user participation. Reactive kinds of participation such as answering questions made by other users are highly regarded in web forums [Nam et al., 2009, Jain et al., 2009, Sinha et al., 2013]; on the other hand, proactive kinds are rewarded in web communities with broader range of activities such as in CodeProject, a web community for software and design development, where there are special rewards for users who publish articles. Characterisation of users categories enables bestowing users who show these desirable features with a prominent position within the web forum community. In most web forums, users are categorised according to a single dimension of criteria, but they also can be categorised according to two or more dimensions. For instance, in CodeProject, users are categorised according to 7 dimensions. Each dimension comprises many levels where a designation is accorded based on points given following a system of points that varies depending on the dimension.¹

The web forum under study considers 2 dimensions of user categories, both aim to describe the degree of expertise of users. The first dimension is arranged as a hierarchy of ranks given to users on the basis of a set of *de facto* and dynamic qualities. *De facto* qualities refer to the functions users accomplish in the forum, i.e individuals appointed to be forum moderators, company experts helping customers, volunteers, etc. Dynamic qualities refer to those that emerge from a user's interaction and contribution to the forum, such as frequency and recency of visits, answers to questions the other users request, participating in discussions, ratings given to their contributions, etc. The set of original ranks that compose this user hierarchy was extracted from metadata corresponding to the web forum under study. This set of ranks is shown in Figure 5.1, where ranks on the far left side of the spectrum represent less expertise than the ranks closer to the far right side. The number of users in each category and posts produced by them as a whole are detailed in Appendix B. Ranks towards the right side of spectrum shown in Figure 5.1 can be achieved by reaching quantitative thresholds determined by *employees* (forum moderators in their meta-role of Facilitators). Four other ranks (*[Company name] Employee*, *Volunteer*, *Moderator* and *Administrator*) are not attainable based on merit but on *de facto* qualities. The second dimension of user categories is based on the Guru role which originates two user categories: *Gurus* and *Non-gurus*. *Gurus* are users who were given this "badge" name on grounds of

¹<http://www.codeproject.com/script/Membership/Reputation.aspx>. Last visited on 31/03/2014.

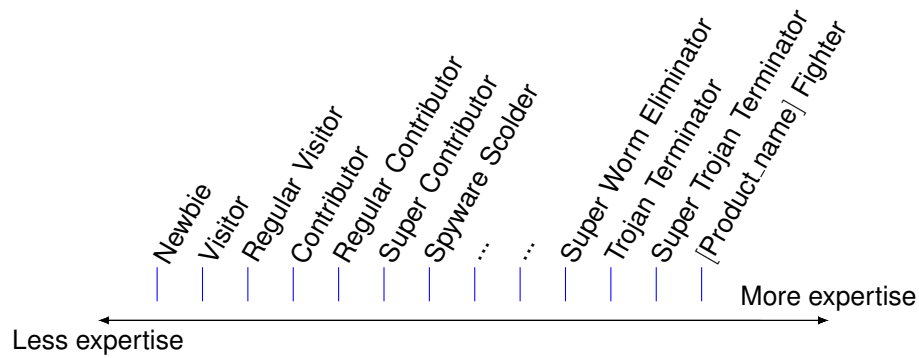


Figure 5.1 – Hierarchy of user ranks based on expertise drawn out of Table B.1 in Appendix B.

dynamic features such as a high degree of contribution to the forum in terms of quality of answers, level of engagement and knowledge they share with users who have a lower degree of expertise i.e. *non-gurus* and interactions with other users in general. All these qualitative features are considered besides other quantitative features by the moderators in charge of identifying this kind of user. The quality of answers, in particular, is not straightforward to assess as it requires subjective consideration by the moderators in charge of assigning ranks and roles to users.

Thus, a user can have two different “labels” attached according to both categorizations, for instance, a *Newbie* user is normally a *Non-guru* user, a *[Product_name] Fighter* is usually a *Guru* and even an *Employee* can be a *Guru* if he or she gathers the necessary qualities to be so.

For the purposes of this research, both categorization were conflated into a more coarse-grained system that keeps in the inherent qualities of a user across both categorizations, rank-based and guru-badge-based. Four categories are comprised in the so created categorization: *employees*, *gurus*, *ranked* and *unranked* users. These groups of forum contributors enclose the different aspects of a consumer related forum scenario and they all perform any of the meta-roles mentioned in Section 1.1.1: Advice Seeker, Advice Giver, Commenter, and Facilitator. *Employees* are current or past workers employed by the software company: in contrast to other users, they may receive financial reward for their contributions to the forum, i.e. managing or contributing to the forum is part of their position duties. *Gurus* do not receive any financial reward for their services. The reward for them has more a subjective character, the main motive for their contribution appears to be the prestige they can obtain, partly via feedback from other users who may reward postings with a positive rating or “kudos” (see Section 5.2) . Common users that are neither *employees* nor *gurus* were split into *ranked* and *unranked* users. *Ranked* users are users with a moderate to high level of expertise. *Unranked* users are composed by individuals holding ranks from *Newbie* to *Super Contributor*, that is users that range from registered ‘lurkers’ to more frequent forum visitors, these users however still do not possess enough experience to be promoted to a higher rank in the hierarchy shown in Figure 5.1.

The annotation dataset of 3,000 posts was randomly chosen in a way that reflected the distribution of individual users belonging to the four different categories in the training dataset. One criterion was that there should be enough heterogeneity of contributions, therefore chosen posts belong to the maximum number of distinct users within each category. The annotation dataset comprises 337 posts that were written by 245 *employees*, 628 by 9 *gurus*, 657 by 657 *unranked* and 1359 by 1353 *ranked* users. This distribution of users reflects 60% of individual users from the full web forum dataset. In the full web forum dataset, 409 users are *employees*, 15 are *gurus*, 19,527 *unranked* and 2,273 *ranked* users. The strategy of choosing the maximum number of distinct users per category causes the average of posts per individual user to be not evenly distributed within each category (ie. 1.38 posts on average ($s = 0.49$) for *employees*, 69.78 ($s = 1.2$) for *gurus*, 1 ($s = 0$) for *unranked* and 1 ($s = 0.066$) for *ranked*). These figures are summarised in Table 5.1.

Table 5.1 also shows figures for a larger dataset compiled for comparison purposes in the analyses described in this chapter. This dataset that comprises 156,132 posts is similar to the one used in previous research ([Mamani Sanchez and Vogel, 2013]) and it will be called from now onwards reduced training dataset (RTD).² In this dataset, employees authored an average of 51.60 posts ($s = 237.96$), gurus 1363.53 ($s = 977.66$), unranked 2.19 ($s = 1.57$) and ranked 34.3 ($s = 180.72$).

Table 5.1 – Frequencies of users and posts across user categories in the Annotation and RTD datasets.

	Annotation dataset				RTD			
	Number of users	Number of posts	Mean	sd	Number of users	Number of posts	Mean	sd
employee	245	337	1.376	0.485	370	19,091	51.597	237.957
guru	9	628	69.778	1.202	15	20,453	1,363.533	977.660
ranked	1,353	1,359	1.004	0.066	2,273	77,960	34.298	180.721
unranked	657	657	1.000	0	17,225	37,771	2.193	1.568

5.2 Ratings

Forum users looking for social recognition are likely to seek to be considered as expert in technical topics according to the nature of the forum. This supposes the existence of an ideal user or “superuser” whose behaviour may be similar to the shown by recognized expert individuals such as employees. Employees are considered as experts *a priori* because of their association with the organization,³ and this may be seen as a certification of their expertise. This status of superuser is embodied by the convergence of linguistic and non-linguistic features in posts of employees (cf. Goffman [1956]) that are deemed as of high quality. Therefore, users seeking recognition are likely to emulate these features in their

²The RTD include blog posts that were dropped in the dataset used in Mamani Sanchez and Vogel [2013].

³Regardless of the fact an employee may not be an expert in all the topics being discussed in the web forum.

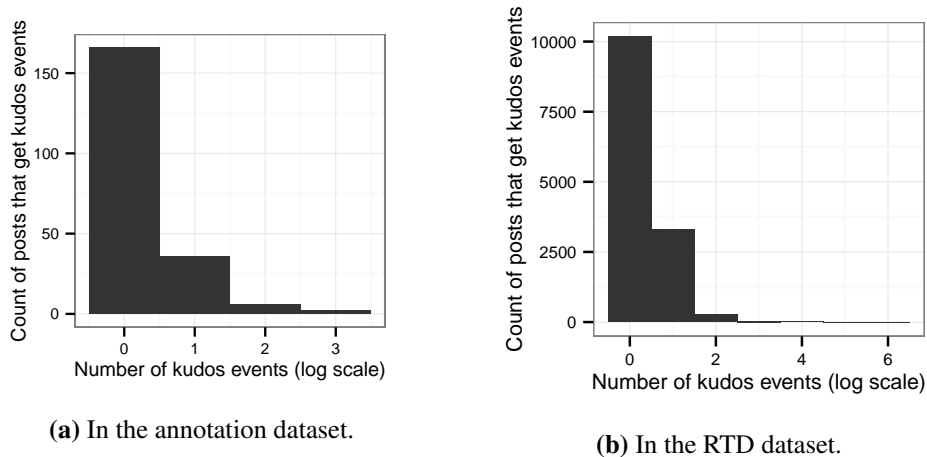


Figure 5.2 – Distribution of the number of kudos-giving events in posts in logarithmic scale.

own contributions to ensure quality.

Besides helping in the promotion of distinguished participants, ratings have been used in research of collaborative filtering to predict users’ interests and behaviour [Wang et al., 2006] and to improve search and recommendation systems Clements et al. [2010].

One strategy used by web communities to leverage the subjective assessment of the quality of posts is to look at feedback given by users that have read a post. Web forum users have the capacity to provide feedback by giving “kudos” to individual postings,⁴ when they find a post useful, valuable or important to be noted. In fact, a user can give kudos to a post for any reason. Theories from social psychology giving accounts of this kind of dynamic have been broadly described in [Cheng and Vassileva, 2005].

For analysis in this study involving ratings, both posts and users can be labelled as “kudoer” depending if they are co-occurrent with kudos. A *kudoer* post is a post that has been given kudos at least once while a *kudoer* user or simply a *kudoer* is someone whose messages have received kudos. Additionally, the number of kudos received by a post overall is recorded as a measure of its rating in comparison to other posts.

Out of 3,000 posts only 210 have at least one *kudo* attribution. This amounts to 7% of the annotation dataset, while the reduced training dataset (RTD) has 9% of *kudoer* posts, nonetheless, this is 13,900 *kudoer* posts. There is a significant difference between these proportions ($p < 0.05$) as posts in the annotation dataset were not selected to reflect a distribution based on kudos but according to user categories. The largest number of *kudoer* posts in the annotated dataset (AnnSet) have been assigned kudos 1 or 2 times as Figure 5.2a shows. Similar trends in the distribution of kudos per post in the RTD are shown in Figure 5.2b. The range of received kudos in the AnnSet is small compared to number of kudos range in the RTD.

⁴“Giving kudos” is achieved by clicking on a designed button using the web interface in the web forum system.

5.3 Signals of emotion

Another dimension for characterising posts is by looking at the sentiment expressed by forum users. As participation in web communities is fuelled not only by expertise but by benevolence exuded by posts, the sentiment expressed in posts was considered to be analysed alongside hedging expressions. Mainstream sentiment analysis research explores words or phrases as signals for conveying polarity of sentiment, be it positive, negative or neutral. In this research, the signals of emotion taken as clues for determining the sentiment are character or pictorial based emoticons mimicking facial expressions in order to convey sentiment polarities.

Some of the research on use of emoticons has explored these signals to produce sentiment scores in question-answering settings [Kucuktunc et al., 2012] and in the training of sentiment classifiers to be applied in independent domains [Read, 2005], as this type of signal of emotion transcends domain specific terms. Emoticons have also been used as a deciding feature for the collection of training data for real-time sentiment analysis in micro-blogging systems [Bifet et al., 2011]. Cultural differences have been analysed by looking at usage of emoticons differences in various topics of discourse such as science and politics [Janssen and Vogel, 2008, Vogel and Janssen, 2009]. More recently in the field of big data analytics, character-based emoticons and “emojis”⁵, have been used as features for classification of large scale data extracted from micro-blogging sites [Bhargava et al., 2013]. In Bhargava et al. research, *emojis* alongside other lexical and shallow stylistic features were used to train machine learning algorithms for authorship attribution.

The set of emoticons chosen for this research are Western-style character-based emoticons and pictorial emoticons. The character-based emoticons were extracted from previous research by Janssen and Vogel [2008], Vogel and Janssen [2009] where emoticons were manually assigned polarity of emotion by consensus between multiple annotators. The set of 45 pictorial emoticons or smilies resemble the same sort of emotions expressed by character-based emoticons (eg. 😊 and 😞). An extra feature of smilies of potential exploit is they characterise distinctions such as gender (eg. 😲 standing for surprised woman and 😲 for surprised man). These smilies were retrieved from the forum management application that provide them to users when writing posts.

Regular expression matching was used to find the occurrence of emoticons. Noisy text adds strain to emoticon matching since it contains elements that seemingly look like emoticons but are not, resulting in false positives such as the examples in Figure 5.3.

Tables 5.2 and 5.3 show the most frequent character-based and pictorial emoticons respectively. For the purposes of this research I will not perform separate analysis for each of them. From now onward I will use the term emoticon to refer to any of these signals of emotion.

⁵Denomination originated in Japan for Eastern-style emoticons intended to be read in horizontal format such as (* _ *) are nowadays of widespread use in mobile phones [Sheu et al., 2011, p. 97]. Emojis are pictorial representations of a wide variety, not solely limited to facial expressions (eg. 🚗, 🍷).

FAT partition of (C:) and the old (F:) is now (E:)... The storage Destination was a Slave Drive (F:) ...
... 7) Selected Recover My Computer. 8) In the drop-list-box above the list of recovery points, I selected ...
Can add URL or use mask (e.g. with ? or *)

Figure 5.3 – Cases of false matches for the emoticons :) , 8) and *) in posts from the web forum under study.

Table 5.2 – Most frequent character-based signals of emotion in the reduced training dataset sorted by decreasing count, characterising E+,E-,E? for positive, negative and neutral polarity emoticons respectively.

	Positive		Negative		Neutral	
	E+	Freq.	E-	Freq.	E?	Freq.
1	:)	2032	!!!	2462	\$\$\$	19
2	:-)	856	???	1740	()	18
3	;))	588	:(437	(=	13
4	:D	433	!?!?	189	\$\$	12
5	:P	168	:-((95	(D)	5
6	;-)	166	:/	33	<=	3
7	=)	163	;-((6	:>)	2
8	:-D	81	:-((2	\$\$\$\$	1
9	8)	77	>;	2	{}	1
10	=>	70	:X	2	I	1





























Only around 8% of posts (237 out of 3,000) in the annotation dataset contain emoticons, therefore some analyses do not achieve significance, as the interaction of emoticons and other features draws out an even smaller number of observations. The RTD dataset shows a similar trend with 9% of posts containing emoticons, that is 13,936 out of 141,338 posts.

Per user category, the portion of posts that include emoticons varies from 2% to 10%, where posts from *employees* make up of them the group that use emoticons the least with only 2.37% of their posts containing emoticons. This ratio is maintained in the RTD dataset. *Gurus*'s posts have emoticons 6.69% of the time, *unranked* users 8.68% and ranked users 9.57%. There is no statistically significant difference in the use of emoticons by *ranked* and *unranked* users, while the difference is significant ($p < 0.05$) when comparing ratios of posts by *gurus* to *ranked* users, but not when comparing to posts by *unranked* users. Therefore, *ranked* users is the group that shows the highest use of emoticons.

5.4 Post labelling strategies based on hedges

In order to have an overall view of how hedges per category are distributed in posts and then find out about correlations with other forum metrics, some hedge discretization functions were devised:

Table 5.3 – Most frequent pictorial signals of emotion in the reduced training dataset sorted by decreasing count, characterising S+,S-,S? for positive, negative and neutral polarity smilies respectively.

	Positive		Negative		Neutral	
	S+	Freq.	S-	Freq.	S?	Freq.
1		2,713		782		625
2		1,750		349		11
3		1,089		305		9
4		26		303		8
5		19		29		0
6		17		19		0
7		14		16		0
8		13		15		0
9		13		13	-	-
10		10		11	-	-

- (A). **Binary discretization.** A post is called HEDGED if it contains at least one hedge of any category and UNHEDGED otherwise. A post is assigned a label of C -HEDGED where $C \in \{\text{Single, NCK, Syntatic, Other}\}$ categories of hedges, if there is at least one hedge belonging to category C within the post and UNHEDGED otherwise.
- (B). **Co-occurrence discretization.** This is based on the simultaneous occurrence of at least one hedge of the four categories. If $H = \{\text{Single, NCK, Syntatic, Other}\}$ is the set of all four categories, the labels are given according to the power set of H , $\mathcal{P}(H)$.⁶ The label UNHEDGED corresponds to the empty set in $\mathcal{P}(H)$. For instance, a post containing instances of at least one Single and NCK and Syntactic hedges is classified as SINGLENCKSYNTACTIC-HEDGED. There are 16 possible posts categories according to this strategy.
- (C). **Subset co-occurrence discretization.** This assigned labels based on elements of $\mathcal{P}(H')$ were H' is a subset of $H = \{\text{Single, NCK, Syntatic, Other}\}$.

These methods do not weigh up the importance of hedge frequency in a post.

5.5 Correlation between post characterizations

In this section, analyses to explore association between hedges and other features in posts are described. The hypotheses taken into account for this exploration are:

Hypothesis 1: Hedges are used qualitatively and quantitatively differently in posts written by users in each of the four user categories.

Hypothesis 2: The inclusion of hedges in a post increases its likelihood of getting kudos.

⁶A distribution of posts in these subsets is depicted by the diagram in Figure 4.4.

Hypothesis 3: Not-Claiming-Knowledge phrases are a distinctive category of hedges that increases the likelihood of a posts getting kudos in comparison to other categories of hedges.

Hypothesis 4: The inclusion of hedges in posts concurrently with other features in posts such as the number of views and size of a post increases its likelihood of receiving kudos multiple times.

5.5.1 Methods of analysis

Occurrences of hedges in both annotation and RTD datasets are extracted at the document level and represented either as a categorical or numerical feature. Categorical representations of hedges are obtained by applying any of the strategies explained in Section 5.4. Numerical representations are drawn from hedge frequencies in each post. Representation of kudos follows similar fashion as categorical and numerical variables.

To explore correlations of kudos given to posts and hedges and emoticons, statistical models are proposed where the independent variable is a variable representing kudos and dependent variables comprehend representation of hedges, emoticons and other features such as a posts' size, number of days it has been published and its visibility.

For examining the correlation between hedges and user categories, graphical methods such as Vocabulary growth curves are applied to sampling points extracted from individual occurrence of hedges.

Statistical tests comprise comparison of proportions, logistic regression for generalised linear and negative binomial models, and general linear hypothesis testing are applied over categorical post labels and continuous post's features. Detailed explanation about method set up is given in follow-up sections.

5.5.2 Correlation of hedging and user categories

The relation between kudos and user categories is evident because kudos usage is one of the criteria taken into account to promote individuals from one category to another more privileged one. A ranked user increases his or her chances of becoming a guru if his or her posts get more kudos than the rest. Similarly, an *unranked* user may become ranked by the same means. The largest groups of *kudoer* posts were written by gurus and ranked users if only kudoer posts are observed (see Figure 5.4), while the largest proportion of *kudoer* posts falls in the guru category (15.61% and 16.59% in the Annset and RTD respectively). *Kudoer* posts by employees constitute the second largest proportion (13.95% and 13.22%). The proportion of ranked posts that are *kudoer* in the Annset differs significantly from the one in the RTD (3.90% and 8.86%), however, as the RTD is a more representative dataset, the proportion of 8.86% is taken to be more realistic. The proportions of *kudoer* posts written by *unranked* users are the smallest ones (1.83% and 2.58%). This clearly shows the association of kudos and non-newbie user categories. Kudos in posts are used as a proxy feature to be explored in statistical models in Sections 5.5.3 and 5.5.4. No statistical models

to account for the correlation of hedges in post and the user category of a post's author are proposed. Nonetheless, in this section I will describe significant differences of how hedges are used in posts across the different user categories.

Figure 5.5 shows the overall cross-distribution of posts according to hedges occurrence and user categories. The percentages in this plot refer to proportion of HEDGED and UNHEDGED posts in each user category. Although the largest number of posts with hedges falls in the *ranked* user category, there are neither significant differences between proportions of HEDGED posts for the four different categories compared one to one nor between HEDGED and UNHEDGED posts proportions in each category in the Annset.⁷ However, in the RTD, there are significant differences in the proportion of HEDGED posts written by employees and all the other categories of users. HEDGED posts by *ranked* users proportion also differs significantly from those by *gurus* and *unranked* users. However, there is no significant difference of HEDGED posts by *unranked* users when compared to *gurus*. On the other hand, *ranked* users is the only user category of posts whose proportions across HEDGED and UNHEDGED posts are not significantly different (50.09% and 50.85% respectively). So, at least from these differences of proportions it looks like the presence of hedges is not a suitable criterion to distinguish between *gurus* and *unranked* users or to depict behaviour by *ranked* users. These caveats prevent supporting Hypothesis 1: hedge use in posts does not enable drawing significant quantitative distinctions between posts written by users of different categories.

Looking at individual hedge categories, percentages of posts that have each one of the hedge categories are shown in Table 5.4. There is a significant difference in the proportion of NCK-hedged posts by employees in comparison to the proportions in the remaining three user categories. NCK-hedged posts by employees constitute the smallest percentage of the total of NCK-hedged posts. The style used by employees in their posts may explain this difference since they engage in forum conversation as part of their position duties at the company, so they are less likely to make subjective comments; therefore their posts would not use *I* or *my* as frequently as compared to other users. There are not other significant differences of NCK-hedged posts between the other user categories, so it could be said that all users but employees use NCK phrases in a similarly quantitative way.

Unranked users have the largest proportion of Single-hedged posts (65.60% of the posts by *unranked* users). Difference with other user categories is significant. On the other hand there are not any more significant differences when comparing proportions in other user categories.

There is a significant difference in the number of *Other-hedged* posts by *gurus* and *ranked* and *unranked* users. Posts by *gurus* being the smallest group in *Other-hedged* posts, it seems that *gurus* are less likely to use *Other-hedges*. A *guru* would not use expressions such as *I'm still learning* or *I'm not too techy* (see Table D.3 in Appendix D).

Regarding Syntactic hedges, *employees* have the highest percentage of posts containing

⁷For measuring the difference in proportions significance here and in subsequent comparisons, a two-sample test of proportions is used, where significant differences are considered for $p < 0.05$.

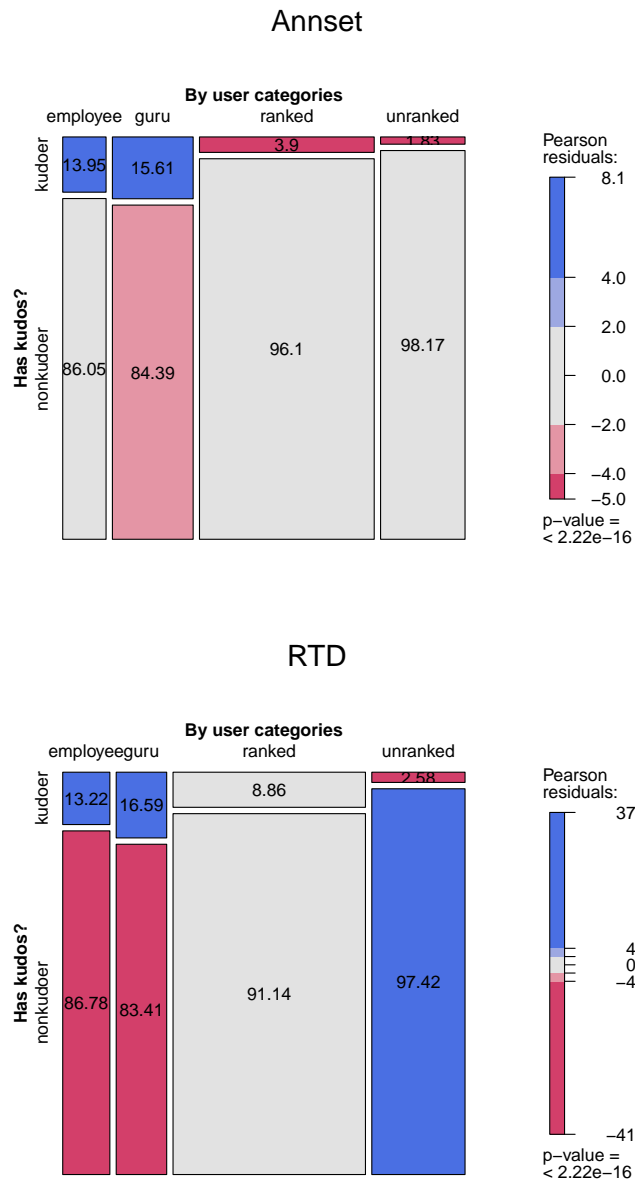


Figure 5.4 – Cross-distribution of posts by user category and kudos in the annotation (Annset) and reduced training dataset (RTD) datasets.

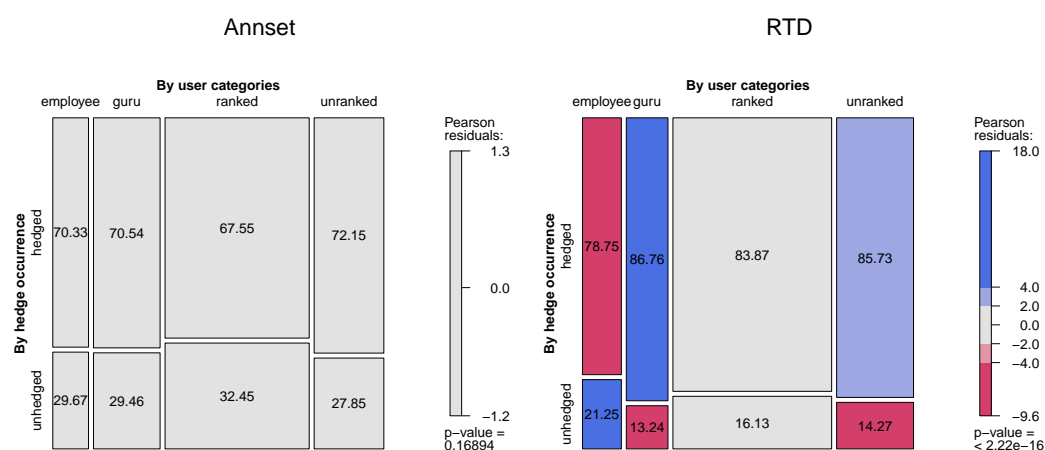


Figure 5.5 – Cross-distribution of posts by user category and hedges in the annotation and RTD datasets.

this kind of hedges, while this percentage is the smallest one in unranked posts. Differences in proportions shown in Table 5.4 are significant with exception of the comparisons between proportions in posts by *ranked* and *unranked* users, and between proportions of posts by employees and gurus.

From this it could be said that in the use of Syntactic hedges there are two differentiated groups of posts: the one by employees/gurus and the one by ranked/unranked users.

Table 5.4 – Distribution of posts containing particular hedge categories calculated within user categories in the Annsset.

	Single (%)	NCK (%)	Syntactic (%)	Other (%)	Hedged (%)
employee	59.05	14.84	38.28	9.20	70.33
guru	58.76	27.71	32.64	6.05	70.54
ranked	59.16	26.05	20.24	10.08	67.55
unranked	65.60	25.57	19.63	10.65	72.15

Lexical richness through user categories There are 784 different hedging types in the Annsset. In gurus' posts they sum up to 265 hedge types overall, employees 208, ranked users 507 and unranked users 329. Ranked posts have a wider range of types, however their contribution to the whole of posts is also outstanding (cf. Section 5.1). Raw frequencies show that posts by *gurus* have 1,770 hedge occurrences, *employee* 1,164, *ranked* users 3446 and *unranked* 1,814 hedge occurrences overall. These observations are summarized in Table 5.5 alongside type-token ratio (TTR) for hedges. The type-token ratios show that posts by *unranked* users have fewer hedge types than other posts, while posts by *ranked* users is the group that have the largest number of hedge types.

Table 5.5 – Distribution of frequencies of hedge tokens (occurrences) and types per user category in the Annset. Type token ratios (TTR) are also shown per user category.

	Hedge occurrences		Hedge types		TTR
	%	Frequency	%	Frequency	
employee	14.21	1164	15.89	208	17.87
guru	21.6	1770	20.24	265	14.97
ranked	42.06	3446	38.73	507	14.71
unranked	22.14	1814	25.13	329	18.14

To corroborate insights about lexical richness, vocabulary growth rates are calculated by dividing the number of hedges *hapax legomena* by the number of hedge occurrences in each user category. To compare mean vocabulary growth rates from different user categories the independent student test was conducted. For this comparison, frequency of hedges occurrences was restricted to 1,150 since is approximate to the employee category's overall hedge frequency, posts from this category have the smallest overall number of hedge tokens in comparison to other user categories (cf. Table 5.5). At the sampling point of 1,150 occurrences, the mean rate in ranked posts is 0.607, while for unranked user it is 0.593. For gurus is 0.59 and for employees 0.56. The only significant difference of means is between employees and ranked users ($p < 0.005$). At the sampling point of 1,750 hedge occurrences, mean vocabulary growth rates are calculated only for gurus (0.58), ranked (0.602) and unranked (0.571) users. The mean vocabulary rate growth for ranked is significantly different from the ones for unranked users ($p < 0.05$) and gurus ($p < 0.05$), but the latter do not differ from each other.

The corresponding vocabulary growth curves are plotted for posts from each user category. The vocabulary growth plot in Figure 5.6 presents a comparison of vocabulary size for *hapax legomena* related to hedge types for the sampling point of 1,150 hedges. Each curve represents how the occurrence of hedge *hapax legomena* grows for each user category. These vocabulary growth curves for *hapax legomena* are plotted based on the number of new hedge types per every 50 hedge occurrences and counted in the order of publication time of the posts they are contained in. This means the 50 first hedge occurrences were taken from a x_0 number of posts from the annotated dataset that were first published, the next sampling point is taken at additional 50 hedge occurrences that are taken from the x_0 number of posts plus a x_1 number of posts that were published subsequently.

Until the sampling point of 900 hedge occurrences two main trends can be observed: the curves for guru and employees separate from *ranked* and *unranked*. From there on, the curve for ranked users soars upwards in comparison to the unranked one (Figure 5.7). The high vocabulary richness in ranked users' posts may be attributed to the large number of posts and users in this category. This trend prevails at sampling points greater than 900 hedge occurrences as Figure 5.7 indicates, where *hapax legomena* vocabulary curves are plotted for hedges in posts by gurus, ranked and unranked users. Intuitions about the

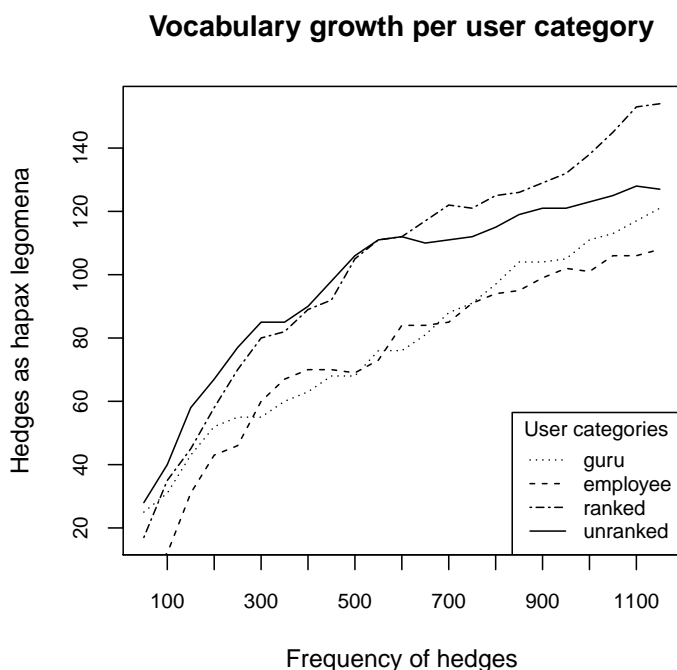


Figure 5.6 – Vocabulary growth curve for hapax legomena hedges across user categories in the annotation dataset.

lexical richness in hedges increases across time could emerge, however I will not address this subject any further.

It is known that type-token ratio and vocabulary growth rate are measures that could poorly represent lexical richness when there is disparity in document size [Baayen, 2008]. This is what occurs in both datasets, as the standard deviation and coefficient of variance in Table 5.6 show. This table also shows that the size of posts by gurus is less disperse than in posts from other categories ($cv = 1.03$ and $cv = 1.03$ in each dataset respectively).

Although these shallow measures indicate that posts from ranked users show high hedging lexical richness, there is not enough evidence to support Hypothesis 1 regarding the qualitative use of hedges across user categories since there is a large disparity in posts' sizes as Table 5.6 shows.

Table 5.6 – Average and standard deviation of post's size in number of words across user categories in the Annset and RTD.

User category	Annset			RTD		
	Mean	sd	cv	Mean	sd	cv
employee	129.79	453.79	3.5	64.15	95.20	1.48
guru	73.74	75.76	1.03	77.76	81.03	1.04
ranked	93.27	119.12	1.28	81.04	97.94	1.21
unranked	108.88	118.19	1.09	97.48	113.61	1.17

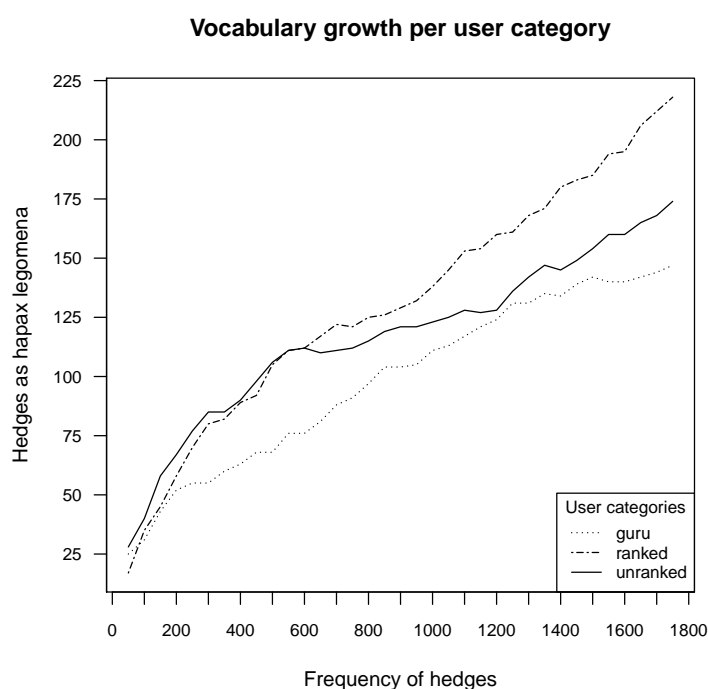


Figure 5.7 – Vocabulary growth curves (hedges) for three user categories: gurus, ranked and unranked in the annotation dataset.

Table 5.7 – Overall counts of users and posts in the annotation dataset accompanied by user and posts count and percentages at the sampling point of 1100 hedges.

	For whole hedge occurrences		For first 1100 hedge occurrences			
	User count	Post count	User count	%	Post count	%
employee	185	235	182	98.38	232	98.72
guru	9	440	9	100.00	382	86.82
unranked	461	461	383	83.08	383	83.08
ranked	902	906	550	60.98	552	60.93

5.5.3 Correlation of kudos and hedges in posts

The correlation between kudos and hedges in posts will be explored from two angles: having kudos as a categorical feature of posts and as a numerical attribute.

Models with discrete representation of kudos

Kudos as a categorical descriptor divides posts into *kudoer* and *nonkudoer* (cf. Section 5.2). The proportion of *kudoer* posts is pretty small in both annotation and RTD sets (7% and 9% respectively), which makes a study of correlation between the *kudoer* and hedge features in a post fairly prone to error when working with the Annset because of its size. Because of this, most of the analysis of correlation between both features will be done over the RTD. Nonetheless, I also show results from exploratory analysis in the annotation

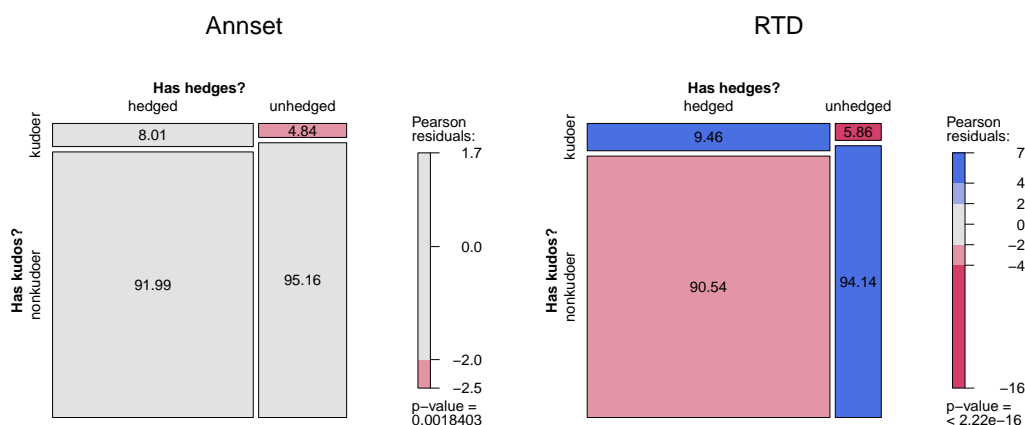


Figure 5.8 – Cross-distribution of posts by kudos and hedges in the annotation (Annset) and reduced training dataset (RTD) datasets.

dataset.

In the annotation dataset, 69% of posts (2,072) are HEDGED while the remaining are UNHEDGED. Most of *kudoer* posts are also HEDGED (79%), while 67.76% of *nonkudoer* are *hedged*. In contrast, the proportion of *kudoer* posts is very small in HEDGED and UNHEDGED posts; 8% in HEDGED posts and 4.7% in *unhedged* posts.

Similarly, 89.5% of *kudoer* posts and 83.55% of *nonkudoer* posts are HEDGED in the RTD. *kudos* appear only in 9.46% of HEDGED posts and 5.86% of *unhedged* posts. These percentages are better illustrated in Figure 5.8, test for difference of proportions indicated all these differences are significant ($p < 0.05$). These differences suggest there is more likelihood that *kudoer* posts have hedges within.

Table 5.8 – Overall distribution of posts according to co-occurrence of hedges and kudos.

	Annotation dataset				RTD			
	HEDGED	%	UNHEDGED	%	HEDGED	%	UNHEDGED	%
<i>kudoer</i>	166	5.57	44	1.48	12,353	7.96	1,449	0.93
<i>nonkudoer</i>	1,906	63.94	865	29.02	118,196	76.12	23,277	14.99

Table 5.8 shows more detailed information about the concomitant distribution of posts according to kudos and hedges in Figure 5.8. Posts with co-occurrent hedges and kudos is the second smallest group, while posts with hedges and without kudos is the largest one. This is due to the small number of posts with kudos. Therefore, the RTD is chosen for building some statistical models since it contains a larger proportion and number of posts concurrently being *kudoer* and HEDGED.

Next, logistic regression models will be fitted for explanatory purposes of the contribution of hedges to the likelihood of a post being *kudoer*. Binomial generalized linear models with logit link function will be fitted since a categorical representation of kudos is being addressed, therefore the dependent variable for these models indicates whether a posts is

kudoer or *nonkudoer*.

To explore the contribution of distinct hedge categories to the likelihood of a post getting kudos, an independent categorical variable is used to represent different posts categories by the types of hedges they contain. This discretized variable is a coding obtained by applying posts labelling strategies described in Section 5.4.

The first model `AllCats1` has as independent variable a hedge-based post type whose values were assigned according to **Co-occurrence discretization**. The possible values for this variable are shown in Table 5.9. Tukey test for multiple comparisons of means is applied to a logistic regression model. Simultaneous confidence intervals at 95% are built for pairwise comparisons.⁸ Since there is large number of comparisons made, a Hasse diagram is used to show significant comparisons in Figure 5.9.

A Hasse diagram is a lattice-like representation for partially ordered sets. As the set of factors in a statistical model keeping a one-to-one odds ratio relationship also defines a partially ordered set, I chose this representation to illustrate significant comparisons of factors, in this case the factors correspond to the different values the independent variable representing post categories in `AllCats1` takes (Table 5.9).

Table 5.9 – Posts categories considered in model `AllCats1` assigned according to co-occurrence discretization. Frequency of posts in groups corresponding to each label are shown in the last column.

	Label	Hedge category in a post				Number of posts
		Single	NCK	Syntactic	Other	
1	NCK		✓			2,233
2	NCKOth		✓		✓	38
3	Oth					281
4	Sing	✓				34,856
5	SingNCK	✓	✓			8121
6	SingNCKOth	✓	✓		✓	403
7	SingOth	✓			✓	1,059
8	SingSynt	✓		✓		46,035
9	SingSyntNCK	✓	✓	✓		22,944
10	SingSyntNCKOth	✓	✓	✓	✓	2,634
11	SingSyntOth	✓		✓	✓	2,566
12	Synt			✓		7,877
13	SyntNCK		✓	✓		1,310
14	SyntNCKOth		✓	✓	✓	40
15	SyntOth			✓	✓	152
16	unhedged					24,726

Each of these labels for a post category is represented as a vertex that has a rectangular shape. Vertices corresponding to posts categories that do not keep significant difference from other categories are completely disconnected from other vertices.⁹ Unidirectional

⁸This means that there is a probability of at least 95% that all estimates for each group are contained within their corresponding interval. This value of confidence is used by default in the package `multcomp` [Hothorn et al., 2008].

⁹In Figure 5.9, disconnected vertices were removed from the diagram for the sake of clarity.

edges represent significant odds ratio relationships between two categories. Be a , b and c vertices that correspond to post groups, an edge goes downwards from vertex a to vertex b if and only if the likelihood of posts in b getting kudos assigned is significantly lower than the likelihood of posts in a . At the same time, there is no c whose posts have more likelihood than in b and lower likelihood than in a . If there is no path of edges connecting any two vertices, it means that the difference between those two groups is not significant in relation to the measure being used (ratio of kudos). Therefore, categories on top in the diagram correspond to higher odds of leading to kudos in comparison to categories in the bottom. The “dot” vertex is used as an intermediate vertex for a compressed representation of all possible significant comparisons. There is a significant odds ratio relationship between any two vertices connected through a dot vertex. The shade of gray in each vertex is an indicator of the means of *kudoer* posts size in each post category. The shade of grey that colours a vertex represents the relative frequency of *kudoer* posts to the total number of posts in the group. Lighter shades are related to a higher frequency than darker shades. The vertex size is proportional to the number of posts in each category. For instance, Figure 5.9 shows a downward edge from the vertex SINGSYNT to SING vertex throughout a dot vertex. This indicates that SINGSYNT-hedged posts are more likely to get kudos assigned on average than SING-hedged ones. The diagram shows that vertices SINGSYNT and SING are the largest ones, this is due to posts with only Single hedges (SING) and with exclusively both Singular and Syntactic hedges (SINGSYNT) are the most frequent ones in the RTD (frequencies in Table 5.9 verify this because there are 34,856 posts labelled as SING and 46,035 labelled as SINGSYNT). Furthermore, the group of SINGSYNT-hedged posts has a higher ratio of kudoer posts compared to the group of SING-hedged ones, this is revealed in the colour if the corresponding vertices: the vertex SINGSYNT is lighter than the vertex SING. Hasse diagrams will be used throughout this chapter to aid in showing observations from models that involve a large number of comparisons.

The main findings after fitting the `AllCats1` model are:

- From the 16 categories of posts enumerated in Table 5.9, 11 keep a significant relation with at least one other post category. Therefore, from now onwards only the relevant categories shown as vertex in Figure 5.9 will be referred to when describing the results from this analysis.
- Posts with one exclusive category of hedges. There is no significant difference between exclusively *Single* and exclusively *NCK-hedged* posts. Similarly these posts do not differ from posts with only *Syntactic* hedges. There is no statistically significant difference between posts containing exclusively *Other* hedges and any of the remaining categories of post.
- There is no significant evidence to confirm the difference between posts with exclusively *NCK* hedges and *unhedged* posts. Similarly, the difference between *unhedged* posts and those containing *SINGULAR* and *OTHER* hedges concurrently (*SINGOTH*)

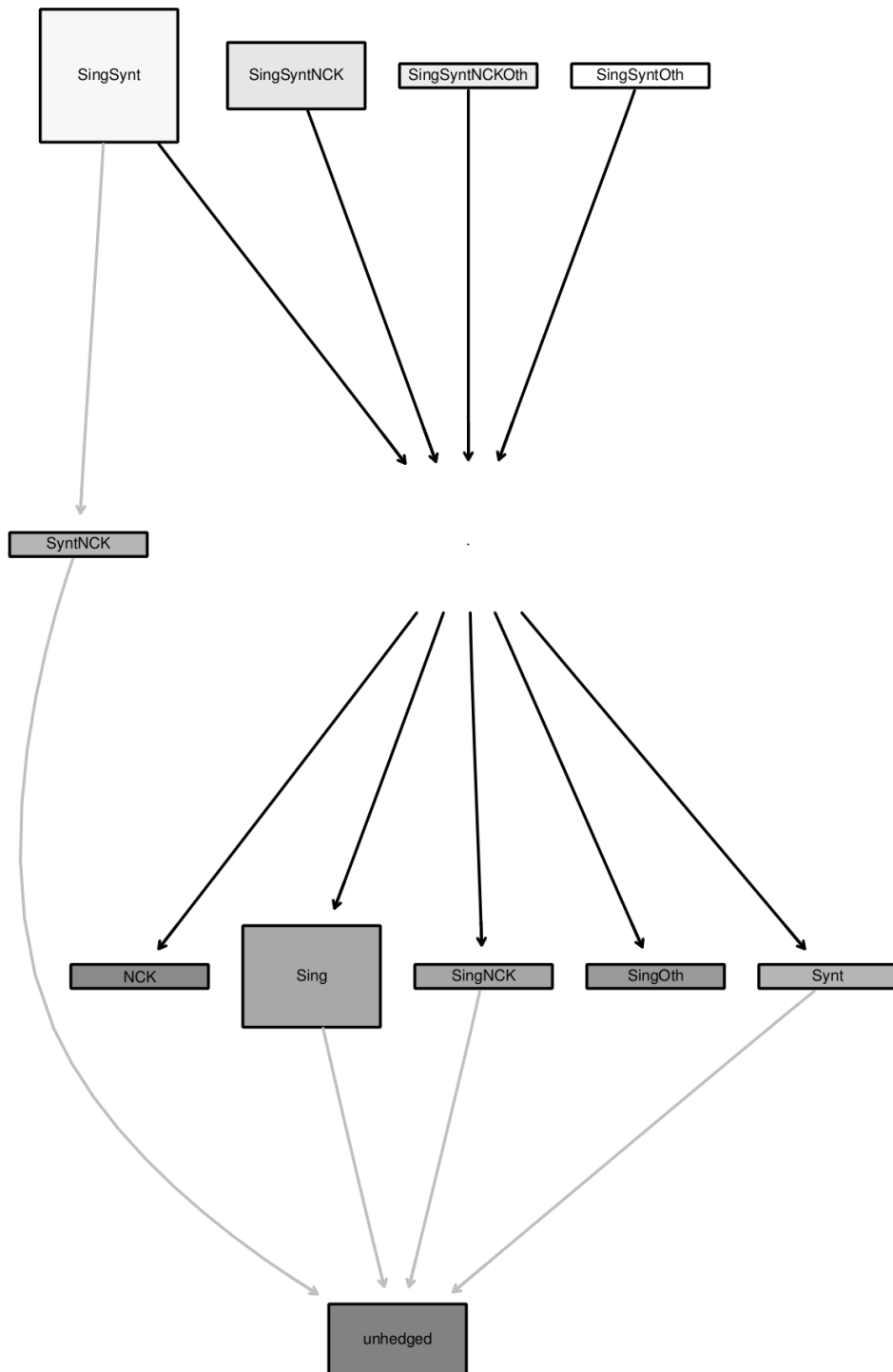


Figure 5.9 – Significant Pairwise comparisons in model AllCats1.

is not significant. All the other 9 categories of posts are significantly different and all they are more likely to lead to a post getting kudos than *unhedged* posts.

- Posts containing all four categories of hedges concurrently (SINGSYNTNCKOTH) are more likely to be *kudoer* than posts with only one of these three kinds of hedges {Single, NCK, Syntactic}. Similarly, the later kinds of posts are less likely to get accorded kudos than posts containing all of these three categories (SINGSYNTNCK). It seems from this, that posts with heterogeneous types of hedges (3 or 4 categories) within are more likely to get kudos than their one-category counterparts. Figure 5.9 shows vertices corresponding to SINGSYNTNCKOTH and SINGSYNTNCK on top of the diagram. Significant differences between these groups and SING, NCK, SYNT are illustrated by edges running downwards.
- In particular, *SingSynt*, *SingSyntNCK*, *SingSyntNCKOth* and *SingSyntOth* categories of posts contain all SINGLE hedges and all these four post categories are more likely to get kudos than exclusively *Single*-hedged posts. Similarly to the previous point, there are edges running down from these groups to the SING vertex.

Confidence intervals and probabilities for significant differences for model `AllCats1` are shown in Figure 5.10. Differences of estimates in log odds ratios are shown in the left-hand plot, while probabilities are shown in the right-hand plot. Those confidence intervals above zero signal that the term in the left of the comparison with more likelihood than the term to the left. Conversely, confidence intervals below zero point out that the right term in the comparison is more likely to be *kudoer* than the left term.¹⁰ A confidence interval central point indicates the calculated estimate for the difference in each pairwise comparison. Likewise, the plot in the left shows the probabilities for each comparison between the first term and the second term. For instance, the probability of a UNHEDGED post being *kudoer* in comparison to a SYNTNCK-hedged post is 0.41; this is revealed by the value for the tick-mark labelled as `unhedged - SingNCK` in the right-hand plot of Figure 5.10. The highest probabilities occur when posts with combination of Singular and Syntactic hedges plus optionally NCK and Other hedges are compared to NCK-hedged posts and UNHEDGED posts (cf. values `unhedged - SingSyntOth`, `unhedged - SingSynt`, `unhedged - SingSyntNCKOth`, `unhedged - SingSyntNCK`, `SingSyntNCK - NCK`, `SingSyntNCKOth - NCK`, `SingSynt - NCK`, `SingSyntOth - NCK` for tick-marks in Figure 5.10). Table 5.10 shows these probabilities in detail.

Since SYNTACTIC hedges constitute a very ambiguous category of hedges, a model using an independent variable resulting from **Subset co-occurrence discretization** labelling strategy, `AllbutSynt1`, is proposed. As in model `AllCats1`, a Tukey test for multiple pairwise comparisons at 95% of confidence level was used to find out which categories of posts out of 8 are more likely to be *kudoer*. These categories are the ones that constitute

¹⁰Some non-significant comparisons are shown for the sake of illustration: those whose confidence intervals cross the y-axis at zero.

Table 5.10 – Probabilities of a post being *kudoer* in pairwise comparisons.¹

Compared to	NCK-HEDGED	UNHEDGED
SINGSYNTOTH-HEDGED	0.663	0.675
SINGSYNT-HEDGED	0.656	0.668
SINGSYNTNCKOTH-HEDGED	0.645	0.658
SINGSYNTNCK-HEDGED	0.637	0.650

¹ Probabilities in comparisons with UNHEDGED posts are the complementary probability of the values shown in Figure 5.10.

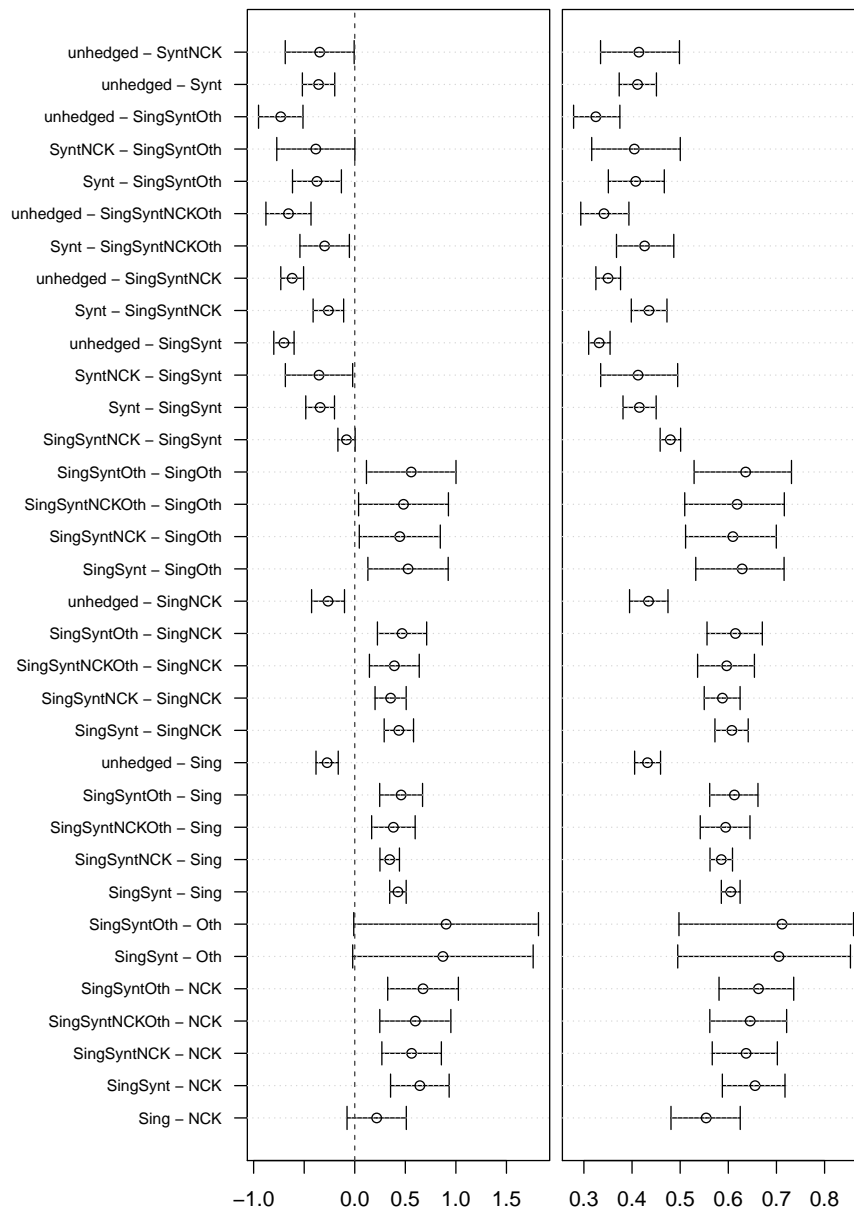


Figure 5.10 – Confidence intervals for difference of estimates of effects (left) and actual probabilities (right) for model AllCats1 at 95% confidence level. Only intervals placed below and above zero are significant. Probabilities are transformed from logit scale estimates, those below the 0.5 mark favour the left term in the comparison.

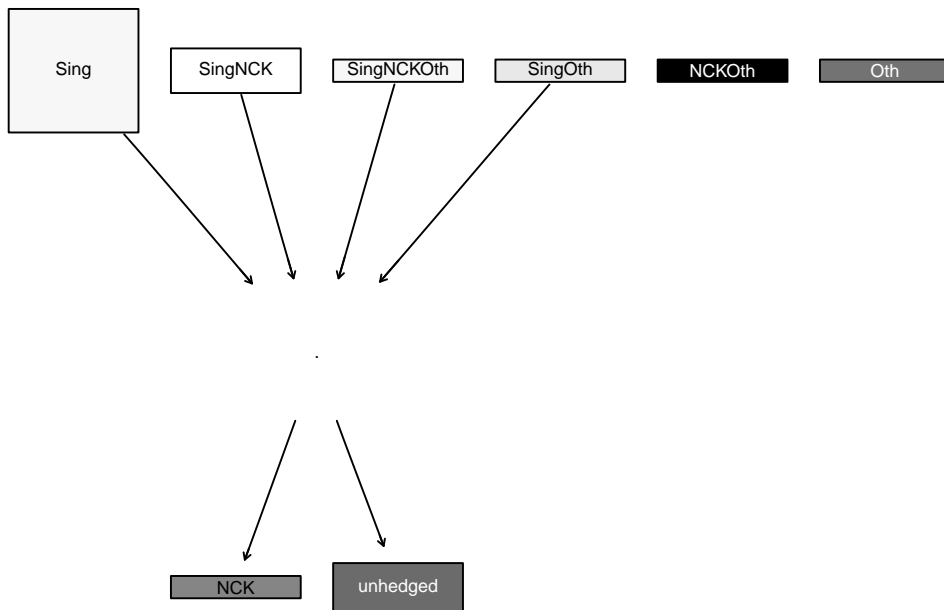


Figure 5.11 – Significant Pairwise comparisons in model `AllbutSynt1`.

labels for vertices in the Hasse diagram in Figure 5.11. The relation of significant comparisons in model `AllbutSynt1` is depicted in this diagram.

Posts containing only Single hedges are significantly more likely to get kudos attributed than posts with NCK hedges on their own and UNHEDGED posts. Post with Single hedges accompanied by NCK, Other and a combination of both are more likely to get kudos than NCK-HEDGED and UNHEDGED posts.

Posts with Other hedges on their own and accompanied by NCK hedges do not keep a significant difference from other categories of posts.

In model `AllCats1`, the difference between estimates for exclusively SINGLE-hedged and NCK-hedged posts was irrelevant. On the contrary, this difference is significant in model `AllbutSynt1`: Single-hedged posts are more likely to get kudos than NCK-hedged posts. However, the difference between SINGLE-hedged and SINGLE-NCK-hedged posts is still not significant.

At least in three categories of HEDGED posts, the presence of Syntactic hedges makes them more likely to get kudos: SINGLE-HEDGED, SINGNCK-HEDGED and SINGOTH-HEDGED posts increase their likelihood to get kudos when they include Syntactic hedges too. On the other hand, the lack of account for Syntactic hedges does not affect in essence the likelihood of getting kudos for NCK-HEDGED posts when compared to Single-hedged posts, the probability of SINGLE-HEDGED posts getting kudos is 0.585 when compared to NCK-HEDGED posts in model `AllbutSynt1` while the probability of SINGLE-SYNTACTIC-HEDGED posts is 0.587 in model `AllCats1`.

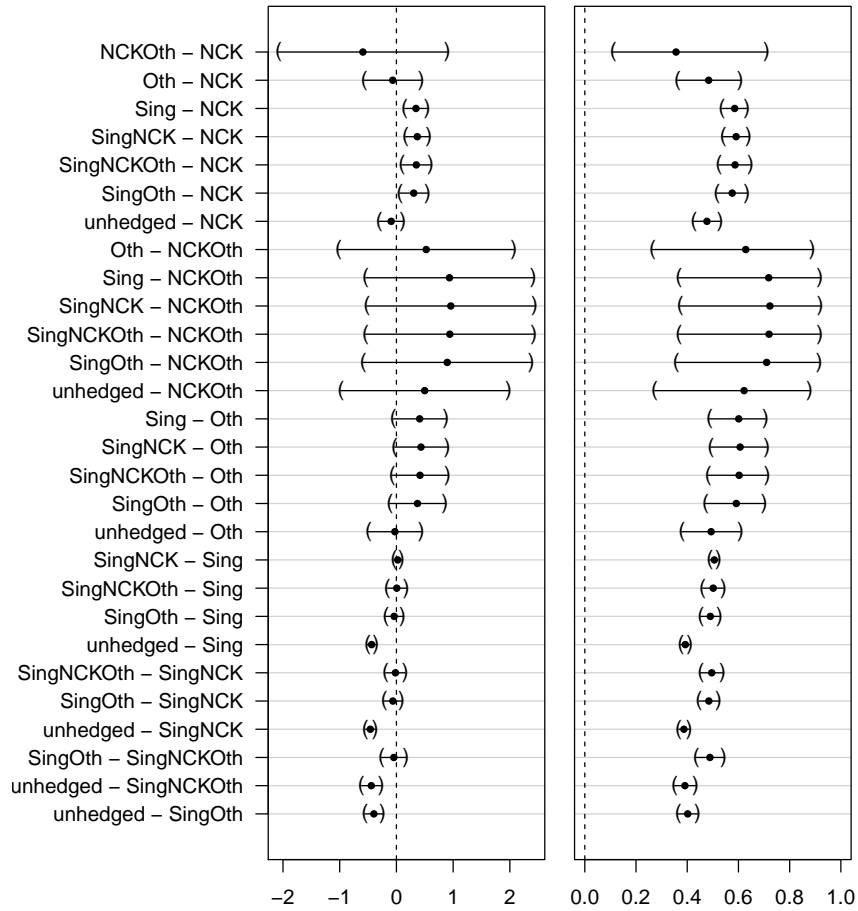


Figure 5.12 – Confidence intervals (left) and probabilities (right) for difference of means of effects for model AllbutSynt1.

Table 5.11 summarizes the probabilities of the aforementioned kinds of posts of getting kudos when compared to NCK-HEDGED and UNHEDGED posts. These are probabilities that fall in the middle point of confidence intervals for each significance difference shown in Figure 5.12. More precise figures including lower and upper bounds for these confidence intervals are shown in Table F.2 in Appendix F. The plots in Figure 5.12 show confidence intervals for differences of all the effects means (left plot) and probabilities (right plot) at 95% of significance for the model AllbutSynt1 that do not take Syntactic hedges into account.¹¹

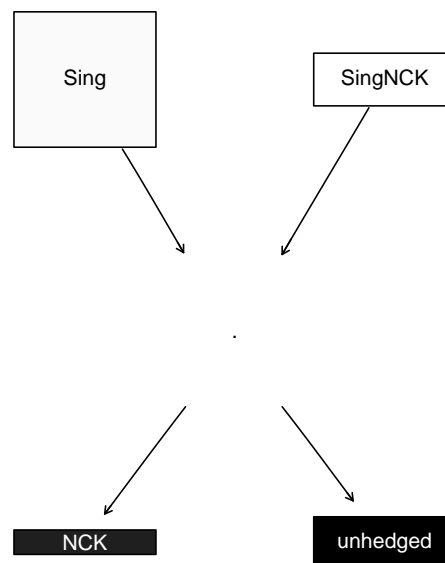
As the main categories in this study are Single and NCK hedges, a third model was built on an independent variable whose values are assigned using the Subset co-occurrence discretization labelling strategy over only these two categories of hedges. Thus, four categories of posts are considered for this SinglevsNCK model: SINGLE-HEDGED, NCK-HEDGED, SINGNCK-HEDGED and UNHEDGED. Significant differences can be observed in Figure 5.13. Posts only containing Single hedges or with mixed Single and NCK hedges are most

¹¹Significant confidence intervals do not cross the zero axis in the left plot of difference of means.

Table 5.11 – Probabilities of a post being *kudoer* in pairwise comparisons in model `AllbutSynt1`.¹

Compared to	NCK-HEDGED	UNHEDGED
SINGNCK-HEDGED	0.591	0.613
SINGNCKOTH-HEDGED	0.587	0.609
SING-HEDGED	0.585	0.608
SINGOTH-HEDGED	0.576	0.598

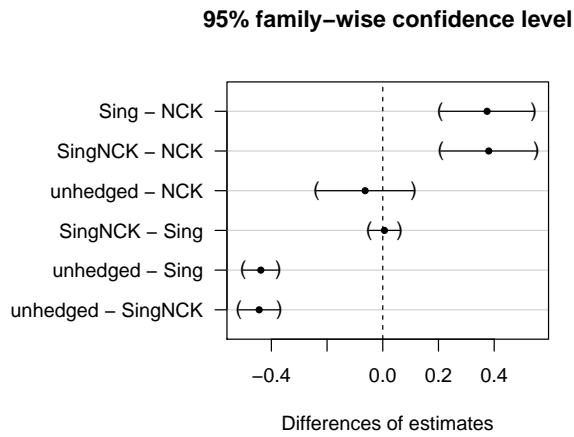
¹ Probabilities in comparisons with UNHEDGED posts shown in this table are the complementary probability of the values shown in Figure 5.12.

**Figure 5.13** – Significant pairwise comparisons in model `SinglevsNCK`.

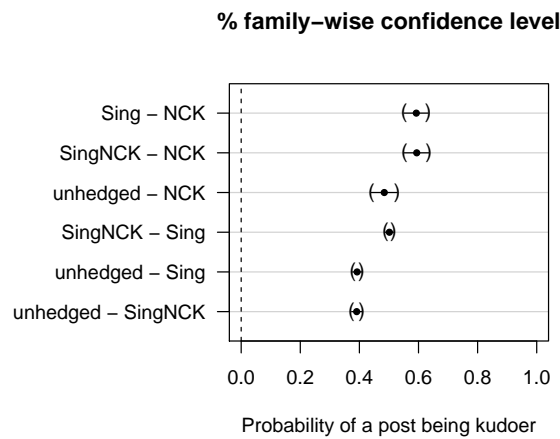
likely to render kudos than posts that exclusively have NCK hedges. These posts (SINGLE and SINGLE-NCK HEDGED) are also more likely to render kudos than UNHEDGED posts. There is no significant difference between SINGLE-HEDGED and SINGNCK-HEDGED posts and between NCK-HEDGED and UNHEDGED posts as Figure 5.14a reveals. This figure shows the corresponding confidence intervals for comparisons of means of each category of post. Again, significant differences in comparisons have confidence intervals that fall above or below the 0 axis.¹² Confidence intervals for probabilities of a post getting kudos is shown in Figure 5.14b, the averages are shown in Table 5.12.

The three models `AllCats1`, `AllbutSynt1` and `SinglevsNCK` show that any kind of HEDGED post has more probability of getting kudos than UNHEDGED posts. The Akaike Information Criterion (AIC) was used to assess the goodness of fit in each model in relation to the other models. Table 5.13 presents the AIC values for all the models proposed, where

¹²A difference of coefficients being zero would mean there is no difference.



(a) Confidence intervals for difference of estimates of effects for model for Single hedges and NCK.



(b) Probabilities for difference of estimates of effects for model for Single hedges and NCK.

Figure 5.14 – Plots of confidence intervals and probabilities for model `SinglevsNCK`.

Table 5.12 – Probabilities of a post being *kudoer* in pairwise comparisons in model `SinglevsNCK`.¹

Compared to	NCK-HEDGED	UNHEDGED
SINGNCK-HEDGED	0.594	0.609
SING-HEDGED	0.593	0.608

¹ Exact figures for all the intervals for this model in Table F.3 from Appendix F.

the one corresponding to `AllCats1` is the lowest one showing that a model with a more fine-grained distinction of hedges fits better the data.

Table 5.13 – Results from applying the AIC test over the models for *kudoer* posts.

Model	df	AIC
<code>SinglevsNCK</code>	4.00	92788.00
<code>AllCats1</code>	16.00	92362.58
<code>AllbutSynt1</code>	8.00	92792.26

The significant differences from pairwise comparisons in models `AllCats1`, `AllbutSynt1` and `SinglevsNCK` reveal that Hypothesis 2, that the inclusion of hedges in posts increase their likelihood of getting kudos attributed, holds true when comparing HEDGED posts to those without any kind of hedge (UNHEDGED). HEDGED posts in these models have different realisations: in model `AllCats1` there are 15 categories of hedged posts, in `AllbutSynt1` there are 7 and in `SinglevsNCK` there are 3. In these models, posts in categories where Single hedges are included have more likelihood of being kudoer than unhedged posts. Syntactic hedges are only considered in model `AllCats1` and categories of posts including this kind of hedges are all more likely to get kudos accorded than unhedged posts. In these three models, groups of posts with only Other hedges, NCK hedges or combination of both do not keep significant difference with unhedged posts. Moreover, because of this not significant difference, posts with only NCK hedges or accompanied by Other hedges are less likely to be kudoer than group of posts that keep significant difference with unhedged posts. This leads to conclude that there is no enough evidence to support Hypothesis 3, that Not-Claiming-knowledge (NCK) is a category of hedges whose inclusion in posts increased their likelihood of getting kudos accorded.

Models with continuous representation of kudos

In the previous section, the models described use a discretized variable drawn out of the number of kudos (*kudoer* and *nonkudoer*). On the other hand, a numerical representation of kudos corresponds to the number of kudos-giving events related to the posts. This means that if a post has n kudos, there were n situations where n different users gave kudos to the post.

For modelling the numeric representation of kudo-giving events, three other variables are taken into account: the number of days a post has been online, the number of words the post contains and the number of views a post has received.

The number of days a posts has been online corresponds to the number of days elapsed from the time the post was first published until the cut-off day for the dataset used in this study.¹³ This variable was included in the models because it could be claimed that the

¹³The collected dataset comprises all the posts published during 2 years, 6 months and 5 days, calculated from 2008-04-07T19:46:59+00:00 to 2010-10-12T11:24:16+00:00 (cut-off time).

longer a post is online the more times it gets accorded kudos. The number of words is a rough representation of a post size because of the noisy nature of text in posts.

A high number of zeros is observed in the count of kudo-giving events as Figures 5.15a and 5.15b show.¹⁴ As mentioned earlier in this chapter, 7% and 9% of the posts are *kudoer* in the Annset and RTD respectively. A ratio D of the variance by the mean in the number of kudos reveals overdispersion ($D = 3.7$ and $D = 12.97$ in each dataset).

Therefore, the number of times kudos are given to a posts was modelled as a dependent variable following a zero-inflated distribution. A zero inflated distribution accounts for excess zeros by combining two distributions: a regular count distribution such as Poisson or Negative Binomial and a degenerate distribution with point mass at zero. When a Poisson distribution is chosen, this is commonly called Zero-Inflated Poisson or ZIP distribution, while when a Negative Binomial distribution is the chosen on, it is called Zero-Inflated Negative Binomial or ZINB. Primarily, a negative binomial model with zero inflation was fitted to assess the contribution of hedges to the number of times kudos were given to a post. In this way, a Zero Inflated Negative Binomial (ZINB) regression gives account of excess zeros as having two different types: sampling zeros and structural zeros. In the process of kudo-giving events, sampling zeros come from when a user has viewed a post and he or she has not deemed it worth of kudos. Structural zeros come a process where a user has not seen a post, therefore the post is not given kudos, the process is ineligible to have a kudos-giving event as an outcome.

Therefore, the ZINB model `ZINBcathedges` is built comprising these two processes. The first part or count model is modeled as a regular negative binomial regression accounting for sampling zeros and the second part or inflation model is modeled as a binomial distribution.

In the RTD, outliers in the number of kudos and hedges variables were dropped from the dataset. This constitutes a very small decrease (19) in the number of posts but outlying number of kudos and hedges may affect the model outputs since the means affects the distribution. Dropping the outliers reduces the overdispersion to $D = 3.16$, therefore the dataset was reduced in number.

The count model in `ZINBcathedges` has as predictors variables representing: whether a post has hedges or not, number of words and number of days online. The zero inflated model has number of views as unique predictor. The dependent variable is the number of kudos a post receives.

The odds ratio of getting false zeros in the number of kudos decreases marginally (odds ratio of -0.4018) with each new post view. The incidence of kudos-giving events for UN-HEDGED posts is 0.669 times the incidence rate for HEDGED posts, holding the other variables constant. For every one day increase in the number of days online (`numdays`), the probability of increasing the frequency of kudo-giving events slowly decreases (multiplied by a factor of 0.997). Also, a unit-increase in the number of words (`numwords`) suggests

¹⁴ This distribution is also shown in Figure 5.2.

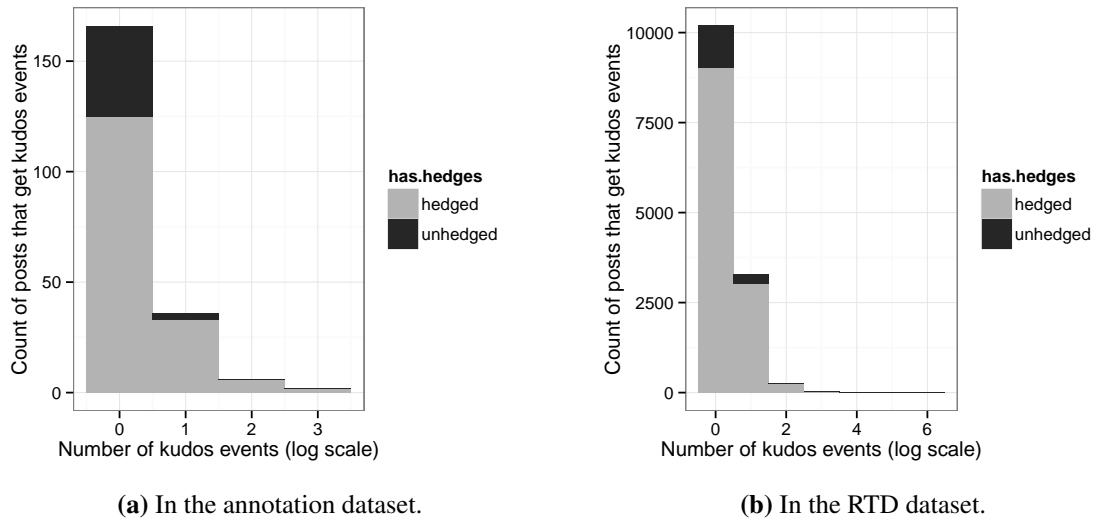


Figure 5.15 – Distribution of the number kudos-giving events in posts in logarithmic scale in HEDGED and UNHEDGED posts.

that the average of kudo-receiving events increases by a factor of 1.003. All these variables affect significantly the number of times kudos are received although marginally. These results are shown as coefficients in Table 5.14 alongside standard errors for each predictor variable. Coefficients for alternative Zero-Inflated Poisson ($ZIP_{cathedges}$) and Negative Binomial ($NB_{cathedges}$) models are shown alongside.

The model also shows that the probability of getting false zeros is 0.494. The log odds of getting zero inflated kudo-giving events slightly decreases with every new view (multiplied by a factor of $0.99947 = e^{(-0.00052)}$).

Table 5.14 – Coefficients and standard errors (S.E.) for three explanatory models of the number of kudo-giving events: a Zero-Inflated Negative Binomial model ($ZINB_{cathedges}$), a Zero-Inflated Poisson model ($ZIP_{cathedges}$) and a Negative binomial model ($NB_{cathedges}$). Coefficients for all variables are statistically significant ($p < 0.0001$).

	$ZINB_{cathedges}$		$ZIP_{cathedges}$		$NB_{cathedges}$	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Intercept	-0.97587	0.03815	0.25448	0.01593	-1.47138	0.01957
has.hedges (unhedged)	-0.40184	0.03145	-0.46392	0.02827	-0.41047	0.03099
numdays	-0.00251	0.00005	-0.00227	0.00002	-0.00215	0.00004
numwords	0.00289	0.00010	0.00183	0.00004	0.00299	0.00008
Zero inflated part						
Intercept	-0.02222	0.0897	1.5628	0.01187	-	-
numviews	-0.00052	0.00001	-0.00011	~0	-	-

To determine the improvement of using the presence of hedges as a predictor variable, the model $ZINB_{cathedges}$ was also compared to a model $ZINB_{nohedges}$ where only the variables number of days and number of words are considered. The Vuong test [Vuong,

Table 5.15 – Goodness of fit comparison of models to account for number of kudos events. Negative values for AIC favour the left side of comparison as smaller figures suggest a better fit. Positive values for the Vuong measure indicate the left model better fit compared to the model in the right.

	Measure	NBcathedges	ZIPcathedges	ZINB-nohedges
ZINBcathedges	Vuong	9.077069	9.19838	-
	AIC diff.	-343.9	-5272.8	-165
NBcathedges	Vuong	-	8.515973	-3.85802
	AIC diff.	-	-4928.9	178.9
ZIPcathedges	Vuong	-	-	-8.992
	AIC diff.	-	-	5107.8

1989] was used to compare to an alternative regular negative binomial model alongside the AIC measure. Results from comparing models using these measures are shown in Table 5.15. It can be seen that the Zero-inflated negative binomial model `ZINBcathedges` is the one which best fits the data. A likelihood ratio test was also applied comparing the Zero-Inflated Negative binomial model and the regular negative binomial one, with an approximated χ^2 value of 347.86 and significant ($p < 0.0001$), which shows that the model `ZINBcathedges` with more degrees of freedom fits the data better than the NB alternative. It also follows that `ZINBcathedges` fits the data better than the model `ZINB-nohedges` where the variable `has.hedges` was dropped out (the AIC value for `ZINBcathedges` is smaller than for `ZINB-nohedges`¹⁵).

However, this model without the variable `has.hedges` fits the data better than the alternative `NBcathedges` and `ZIPcathedges` models which shows that a Zero-Inflated Negative Binomial is better than both a ZIP model or NB model to produce an explanatory model of the likelihood of a post receiving kudos repeatedly. This is indicated by the negative values for the Vuong measure and positive difference of AIC values when comparing these models, one to one (see Table 5.15).

Again, these results show the inclusion of hedges benefits posts in making them more likely to get kudos repeated times, which gives support to Hypothesis 4, that hedge occurrence in posts increases a post's likelihood of receiving kudos repeated times in comparison to UNHEDGED posts.

5.5.4 Interaction with signals of emotion

In the annotation dataset, there is no significant difference in the proportions of posts having emoticons and hedges or no hedges, 8% of each, posts with and without hedges have emoticons at the same time.

Emoticons were found in small proportion in HEDGED and UNHEDGED posts (8.98%

¹⁵The Vuong test was not applied here because `ZINB-nohedges` is a nested model of `ZINBcathedges` since both models only differ in one variable but assume the same Zero-Inflated Negative Binomial distribution.

and 8.57% respectively). The difference in proportions is significant ($p = 0.043$) which would show that there are more HEDGED posts that have emoticons compared to UNHEDGED posts. On the other hand, in 13,839 posts with emoticons, 84.68% of them has hedges 84,02% of posts without emoticons have hedges as well, however this is 118,830 posts which makes this small difference significant.

Table 5.16 – Overall distribution of posts according to co-occurrence of emoticons and hedges.

	Annotation dataset				RTD			
	HEDGED	%	UNHEDGED	%	HEDGED	%	UNHEDGED	%
has emoticons	115	5.46	43	2.04	11,719	7.55	2,120	1.37
no emoticons	1,355	64.28	595	28.23	118,830	76.53	22,606	14.56

Posts were categorised according to emoticon use by applying a criterion that takes into account the polarity of emoticons. The prevalent post polarity was chosen according to a majority category strategy upon emoticons polarity, eg. positive posts have mostly positive emoticons. According to polarity, posts in the RTD are distributed in this way: 5.27% of posts are positive, 3.05% negative, 0.59% neutral and 91.09% have no emoticons.

The models proposed to explore the contribution of emoticons and hedges to post rating is a generalized linear model that has as dependent variable the discrete labelling of *kudoer* posts. The independent variable represents the interaction of posts by polarity and posts labelled according to Binary discretization over hedges. This interaction is compared by carrying out multiple comparisons of means using Tukey contrasts.

In the first model `BinHedgesEmots`, hedges are presented as `has.hedges` categorical variable that has `hedged` and `unhedged` as possible values. Emoticons are represented by a categorical variable that signals a post's sentiment polarity and whose possible values are `{pos, neg, neut, unhedged}`.

Significant differences between interactions compared one to one are illustrated by the Hasse diagram in Figure 5.16. The result of pairwise comparison of interactions is represented as partial order in the odds ratio relation of a post being *kudoer* between any two interactions. As this figure illustrates, HEDGED-negative posts and UNHEDGED-unemoted posts are the less likely to get kudos accorded in comparison to all other interactions. The only interaction where there is not enough evidence to claim anything regarding its difference with other post categories is the one that is UNHEDGED and has neutral emoticons alongside, mostly due to a small number of neutral posts overall.

If a post is HEDGED, it is more likely it will be *kudoer* if it contains positive emoticons alongside than negative ones or no emoticons at all. However, there is no significant difference with those that contain neutral emoticons, If a post is UNHEDGED, it is more likely to get kudos if this has positive polarity in comparison to having negative polarity. There is statistically significant difference between positive and negative posts and UNHEDGED posts.

UNHEDGED posts with positive polarity are more likely to be *kudoer* than negative ones.

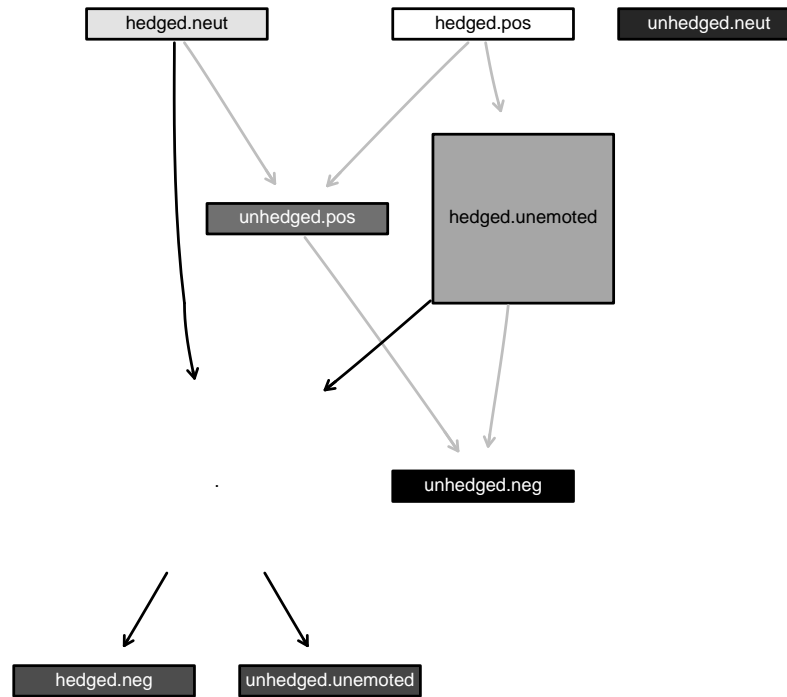


Figure 5.16 – Hasse diagram for the model `BinHedgesEmots`.

The next model `SingNCKEmots` comprises a categorical representation of hedges that considers the main two categories of hedges: Singular and NCK phrases. Therefore, representation of the interaction between post polarity and hedge-based post category is used as independent variable. Significant differences calculated by the Tukey pairwise comparison method are shown in Figure 5.17.

Positive posts with Singular and a combination of Singular and NCK hedges are the most likely to get kudos than other resulting from interactions between hedges and emoticons. Three types of posts are the less likely to be *kudoer*: UNHEDGED-negative, UNHEDGED-unemoted, and Singular-hedged with negative polarity.

When a post is positive, it is more likely to get kudos when it is HEDGED compared to being UNHEDGED. Likewise, if a post is unemoted, it is more likely to be *kudoer* when it is HEDGED. For negative and neutral posts, there is not significant evidence to signal difference between HEDGED and UNHEDGED posts. Other four categories of posts do not show significant difference with any other post category: NCK with either negative, positive or neutral polarity, and UNHEDGED-neutral posts.

Regarding posts with significant mutual differences (those whose corresponding vertices have at least one edge connecting it to any other vertex), positive posts are more likely to get kudos if these co-occur with Singular and combination of Singular and NCK hedges than UNHEDGED posts. If posts are unemoted, likewise than with positive posts, if they have Single and combination of Single and NCK hedges, they are more likely to be *ku-*

doer than UNHEDGED posts. When posts are negative or neutral, there are not significant comparisons to consider amongst the different types of HEDGED or UNHEDGED posts they co-occur with.

In model `BinHedgesEmots`, whenever posts are HEDGED, positive are more likely to be *kudoer* than UNHEDGED ones and the latter ones more likely to be *kudoer* than negative ones. Similar observations are drawn in model `SingNCKEmots` with relation to Single-hedged and SingleNCK-hedged posts. These post categories are not affected by the polarity of emoticons they contain in comparison to the observations in model `BinHedgesEmots`, the odds ratio relation still holds in relation to positive, unemoted and negative posts: positive are more likely to be *kudoer* than unemoted and these one more likely than negative ones.

5.6 Conclusions

This chapter has empirically explored the distribution of hedges in web forums with the purpose of finding out how they are used with respect to other pseudo and non-linguistic features.

Research in social motivations in online communities has highlighted how proactive participation is encouraged by rewarding users who actively engage in it, and that finding representative characteristics of these users is relevant because such kind of users are considered reputed members of the web forum communities, who besides pro-activity, have high expertise and other desirable skills. Nonetheless, identifying this kind of individual is not trivial.

With the motivation of exploring these user's characteristics in posts from the web forum community under study, I proposed posts categorizations based on other relevant post features (its author's user category, ratings given to it and polarity of sentiment in the post) to analyse correlations between hedges use and these other features in posts.

I have described posts according to each of these categorizations and based on them, I proposed some statistical models with an explanatory intent, since the main purpose is related to the question of how hedges are used in this domain and how they correlate with other web posts features, and whether the use of hedges may aid in distinctively characterising users playing particular roles in the web forum community.

I have combined certain criteria used in web communities for promoting users to higher ranks in the community hierarchy into two types: de facto and dynamic qualities, de facto qualities are descriptive according to pre-defined roles of users (eg. appointed moderators), while dynamic qualities emerge from the user's participation in the forum such as frequency of visits, and the quality of their posts.

In the web forum under study two categorizations of users were found: one based on ranks, where promotion to higher levels can be achieved by improving some dynamic qualities and the other one based on roles, where prominent individuals are assigned the role of 'guru', based on dynamic qualities, one of them being the quality of their posts assessed

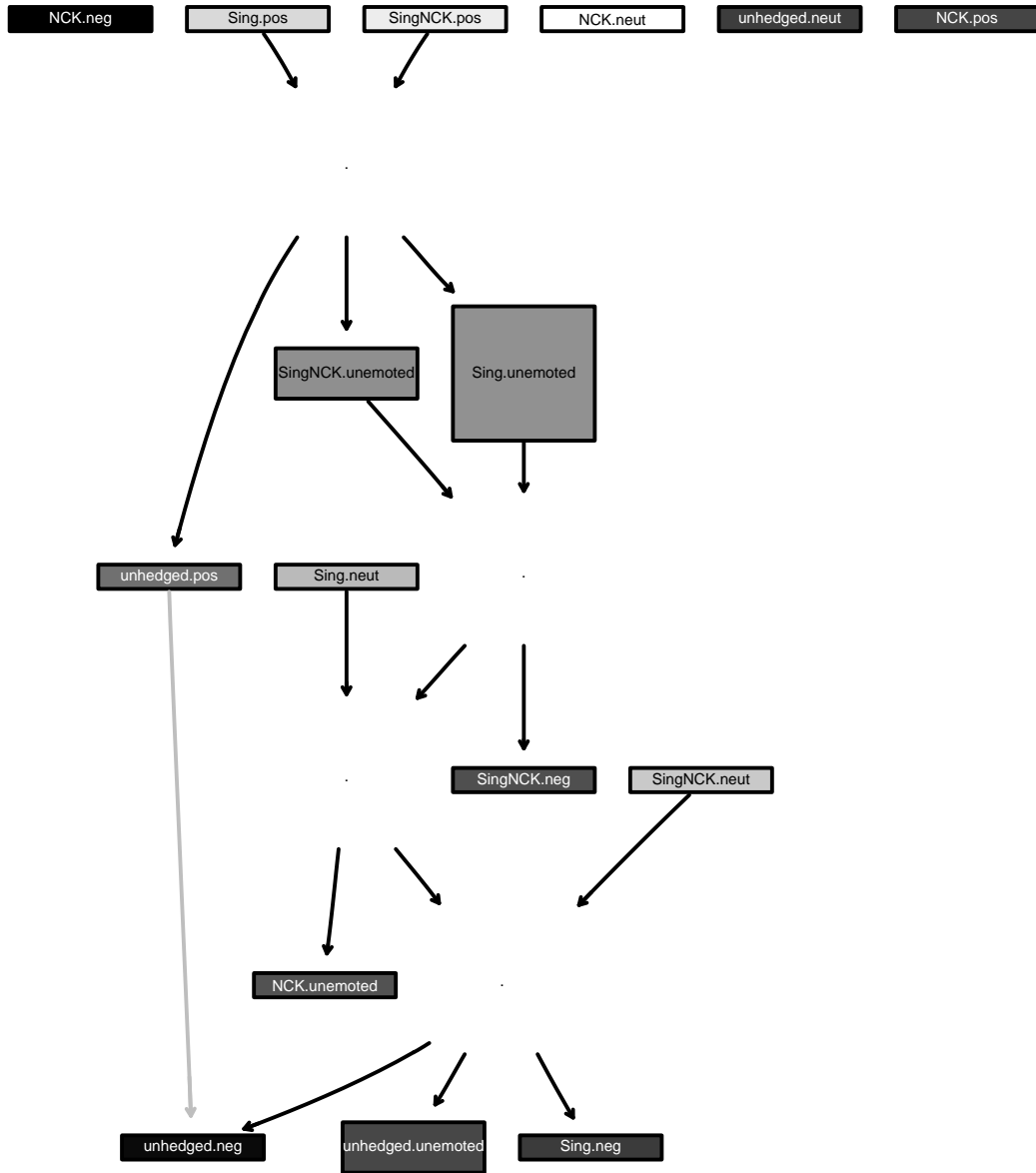


Figure 5.17 – Hasse diagram for the model SingNCKEmots.

in a strict manner by forum moderators; a post's quality is a more subjective criterion that depends on the judgement of forum moderators.

I conflated both categories into one, and each individual is given a label according to the following denominations: *employees*, *gurus*, *ranked*, and *unranked* users. Thereafter, I assigned each post a category according to the one its author holds. Particularly, *ranked* users are individuals who have actively contributed to the community but still have not earned enough prestige or expertise to be qualified as *gurus*, while *unranked* users range from 'lurkers' to users who have shown only a little amount of contribution.

This categorization influenced the way the annotated dataset (Annset) and reduced training dataset (RTD) were built, as I designed these to be stratified resembling the ratio of contributions per user category in the original dataset.

I chose ratings given to posts as a feature that measures the post quality by the community peers and categorize them according to it. A user assigns high ratings ('kudos') to posts that he or she considers as being insightful, useful, or because, all things being equal, he or she just likes the post. Nonetheless, the reasons for a user to give kudos to posts are essentially subjective.

I found that only 7% of posts in the Annset and 9% in the RTD have been given kudos at least once (*kudoer*), since every user may give kudos to a post once. This is congruent with the idea that kudoer posts are outstanding contributions.

Another feature considered as a criterion for categorizing a post is the polarity of sentiment it evokes. I deemed it as important and also there is evidence from literature that the trust that makes users participate in online communities comes from their perception of expertise and benevolence in members of such a community. While most of the studies in sentiment analysis use pure linguistic analyses, I chose a set of emoticons from earlier research that signal a particular polarity of sentiment. According to this feature, I categorized each post as being either positive, negative or neutral. This feature also proved to be sparse in the datasets; 8% and 9% of posts have at least one emoticon in the Annset and RTD respectively. Comparing proportions of emoticon's use across user categories showed that ranked users are the ones that use them most frequently in their posts.

Since a hedge may belong to any of the following four categories: Single, NCK, Syntactic and Other, I designed some post labelling strategies according hedges per category occurring in them, such as: a post is either HEDGED or UNHEDGED, a post is SINGLENCK-hedged when only Single and NCK hedges appear on them and so on. In the same way a post can be labelled as UNHEDGED if no hedge occurs in it. This labelling is useful to characterise posts according to posts in a categorical way, and defined groups or subsets on posts that were then compared by using statistical models.

In this chapter, I have formulated four hypothesis that led the analysis of the proposed statistical models: a) Individuals from different categories do not use hedges according to the same qualitative and quantitative patterns, b) Including hedges in a post increases its likelihood of receiving kudos, c) NCK epistemic phrases in a post lead to an increased likelihood of being assigned kudos, and d) the use of hedges in posts alongside other features

increases the likelihood of more users giving kudos to the post.

I have verified that *gurus* and *ranked* users are users whose posts receive kudos frequently, while *unranked* users have the least proportion of kudo-receiving posts. Therefore, kudos are deemed as a proxy category in the statistical models proposed in this chapter as the higher the proportion of kudos in posts, the higher expertise they appear to be signalling.

I showed that although there are some significant differences between posts from users from the various categories, the sole occurrence of hedges is not particularly a good criterion to distinguish between posts from *gurus* and *unranked* users. A proper model that takes into account subjective features in posts should enable this kind of distinction.

Notwithstanding, there are some useful qualitative observations such as: on average, *employees* use fewer NCK-hedges in their posts, probably due to their communicative style is less likely to include using expressions in the first person, since they perform duties as representing the organization behind the web forum. Similarly, posts by *gurus* include Other-hedges less frequently, this category of hedges comprises domain-tailored expressions such as *I'm still learning* or *I'm not techy*, that expert users are not likely to utter.

In terms of lexical richness, posts from *ranked* and *unranked* users are the ones where hedge types are more productive in comparison to posts from *gurus* and *employees*. However, this may be due to the number of users in the latter categories being lower than in the former ones.

Analysis of interaction between kudos and hedges was set up as logistic regression models both for categorical and continuous representation of kudos given to posts.

The statistical models I proposed were modelled over the RTD as it has representative features in a higher frequency than in the Annset due to its size (number of posts in it).

Since kudos is represented as a categorical feature of posts, I proposed three main generalized linear models to analyse the interaction between kudos and various categories of hedges in posts, namely: `AllCats1`, `AllbutSynt1` and `SinglevsNCK`. All of these models have kudos as binomial categorical variable, whose values are *kudoer* and *nonkudoer* according to whether it was given kudos at least once or not. `AllCats1` is the model that has as predictor variable a representation of all possible hedge categories in a post. `AllbutSynt1` is similar to `AllCats1`, but in this one the predictor variable was given values according to all hedge categories excluding Syntactic hedges.

`SinglevsNCK` has as possible values for the predictor variable a combination of values according to solely Single and NCK hedges occurring in posts.

Based on these models, a Tukey test for pairwise comparison of means was applied to identify relevant statistical differences between groups of posts categorized according to hedges, and their likelihood to be associated to *kudoer* posts.

These models showed that posts containing exclusively one category of hedges share the same likelihood of having kudos. Most of the relevant groups of posts that have a combination of the main types of hedges have more likelihood of having kudos to them than UNHEDGED posts. However, there is not significant evidence to support the hypothesis that posts exclusively containing NCK hedges are more likely to have kudos awarded than

UNHEDGED posts. More heterogeneous posts according to the hedging categories they contain are more likely to get kudos than posts with less varied hedge types.

Without taking into account occurrence of Syntactic hedges, the group of Single-hedged posts are more likely to have kudos assigned than the group of posts with NCK hedges only. I have provided values for metrics for model goodness of fit that show that a model with more fine-grained categories of hedges fits the data better than the other models. These models highly suggest that Hypothesis 2, that hedges included in posts increase their probabilities of getting kudos assigned, holds. However, Hypothesis 3 does not hold in the same way as the contribution of NCK hedges to this likelihood could not be proven.

Further, I have proposed a logistic regression model whose dependent variable is a continuous representation of kudos, to account for the scenario where a post can be assigned kudos repeatedly, as each user from the community can potentially give kudos to this post.

Since only a minimal percentage of posts in the dataset were assigned kudos at all, this scenario was modelled as a statistical process for the case where no kudos are assigned to a post, ie. zeros as values in the dependent variable; this lack of kudos comes from two possible sources: either an individual has not viewed a post and therefore not given kudos to it, or the individual has viewed the posts but deemed it not worthy of kudos. Because of this, I built a model contemplating a Zero-inflated Negative Binomial distribution that has additionally as predictors the number of days a post has been online, the number of words and the number of views a post has gotten and whether it is hedged post or not.

I compared this model to alternative models such as Zero-inflated Poisson and regular Negative binomial models with the same parameter and showed that the Zero-inflated negative binomial model fits the data better than the former ones and accounts for zeros in the kudos-given variable that were not caused because users did not deem it of being worthy of kudos (0.49 of probability that zeros in this variable are caused by zero views). Also, this model was shown to fit the data better than a model where the variable representing hedges is not included.

Finally, the inclusion of emoticon as signals of emotion was explored by proposing two logistic regression models that included polarity of sentiment as part of a predictor variable. One has a binomial representation of hedges in a post (hedged, unhedged) and the second one has values assigned according to the main categories of hedges Single-hedges and NCK phrases. This variable was combined with the values for polarity in a post into a composite variable representing hedge types and sentiment polarity. The groups formed from these interactions were also tested for significant differences using the Tukey test for pairwise comparison of means.

These models showed that whether a post is HEDGED or UNHEDGED, it is more likely to be awarded kudos if it expresses positive sentiment in comparison to showing negative polarity of sentiment. Overall, posts that are UNHEDGED and have no emoticons within are the least likely to have kudos assigned, alongside posts that are hedged and express negative sentiment. I have also shown that when a post has positive polarity of sentiment, its chances of having kudos assigned increases if it features hedges. There are no conclusive

observations about neutral polarity of sentiment in posts, mostly due to low frequency of neutral emoticons in posts.

Overall, I have observed that the use of NCK can lend itself to richer interpretations than other kinds of hedges taking into account other posts features than from the use of Single hedges, since NCK is a more complex category that takes into account the user's involvement, intentions and mental state. For instance, the fact that NCK epistemic phrases are written in the first person makes them straightforward to interpret them according to features in distinct user categories e.g. employees are less likely to use this kind of expression because of the characteristics of their position in the community.

Chapter 6

Future of Hedging Detection in Web text

This chapter describes potential paths of work that may be taken following up this research. Some suggestions across this chapter emerge from limitations of the current research that did not enable performing such experiments. Other suggestions divert more substantially from the present research but are, nonetheless, inspired by it. I have divided related topics that could be developed into future research in two aspects: improvements in the overall detection of hedges in Section 6.1 and application of hedge detection to other natural language processing tasks in Section 6.2.

6.1 Improvement of Automatic Detection of Hedges

The hedge categorization provided by this study constitutes a foundation for the building of a machine learning based method for automatic identification of hedges. So far, the results provided are an outcome from a simple string pattern matching process. Although some categories of hedges such as NCK phrases have been seen as less ambiguous regarding their speculative meaning than Single hedges, a pattern matching based method for linguistic hedge identification maintains all the limitations that machine learning methods try to address in natural language processing. The following sections cover various aspects that can be addressed to improve on my contributions looking forward to the construction of such machine learning method.

6.1.1 Dealing with noisy text

Text in web forums is naturally noisy. Besides non-linguistic items, dealing with misspellings and typos turns out problematic for most kinds of automatic linguistic analysis. Noisy text affects any semi-automatic or automatic processing task in two levels: a) Construction and b) Functional.

At the Construction level, noisy text affects tasks when it involves manual annotation

or corpus linguistic studies which in turn involve keyword pattern matching to be used in other tasks such as pre-annotation, concordancing or parsing. In applying pattern matching to noisy text, positive cases may be missed because of misspellings, resulting in a larger number of false negatives and probably into an increase of false positives.

At the Functional level, when a natural language system has to deal with noisy text when used as a front-end product, similar caveats as at the Construction level may arise. In addition to what is required from that front-end product that deals with an ever transforming language, e.g. neologisms, the system has to be adaptive to new forms of text noise.

The deployment of solutions for transforming noisy text to non-noisy text is not straightforward, e.g. a solution based on the use of a regular spell-checker tool will not always provide an accurate corrected alternative to noisy text strings. For instance Zainkó et al. [2010] points out that a spell-checker excluded many valuable forms from processing. Addressing this kind of issue when dealing with noisy text is outside the scope of this research. There are efforts that work toward solving issues arising from natural language processing in noisy text such as in [Foster et al., 2011]. Also, [Eisenstein, 2013] explores various issues related to noisy text in web and potential solutions.

Noisy text is pervasive in social media and user generated content in general. One possible solution in the case of misspelled words in forums is by looking for the right term in a thread of posts since the correct term is likely to be found there. An adaptive model would be useful in addressing noisy text not only for the detection of hedges but for any other kind of automatic language processing task in general.

6.1.2 Scope

In this study, I have looked at scope of a hedge limiting it to the sentential level. In particular, the annotated scope was chosen to represent constituents in a sentence that are affected by a hedging expression. Nevertheless, the scope of a hedge could be devised to have a broader meaning that includes all the participant elements in hedging. All these elements constitute a context that varies according to each hedge's linguistic realisation, and often these elements are not uniquely placed in the same sentence where the hedging expression appears. Decisions an automatic method for identifying hedges makes do not uniquely depend on information extracted from the sentential level (see Chapters 3 and 4).

Further research on how this context is realised at different levels of discourse could be done. The scope of a hedge could involve more than a couple of sentences if linguistic constituents that affect hedging and linguistic constituents affected by hedging are taking into account. For instance, conditionals often depend on what has been said in previous sentences, and at the same time the propositions affected by hedging could span beyond the sentence where the hedging expression has been used. This study would be worth carrying on not only on informal language styles but on others such as academic writing. The conceptualization of what constitutes the scope of hedging could improve the performance of automatic identification of hedging.

6.1.3 Exploring association to other post features

The presence and frequency of hedges in posts were used as features to explore how they are associated with other features namely: the user category of post's writers and kudos given to posts. Alongside hedges, other post features such as post size, sentiment polarity and number of views were explored to be used as predictors of kudos and user categories. Other features such as number of misspellings, number of sentences in posts or whether the post has been edited or not were not considered for the explanatory models in Section 5. These and other features may be used to create a better model tailored to the dataset used in this study.

These features are centered around individual posts, but features centered around users could be also considered. For instance, the contributions of a user as a whole during their lifetime participation in a web forum could be leveraged in terms of hedge use frequency, amount of time spent online, how long she or he takes to be promoted to the next level in a user hierarchy based on merit, overall sentiment expressed, and so on. Consequently, correlation models centered around individuals could be created by considering these features.

Logistic regression models explaining correlations between hedges and user categories could be proposed instead of simple correlations as the ones drawn out in Section 5.5.2. For instance, a multinomial logit model could take the user category of a post's author as a dependent variable and use other variables mentioned earlier as explanatory variables. This kind of model makes sense in this case because it is a categorical variable that takes more than two values: {employee, guru, ranked, unranked}. In this sense, the post's writer could be seen as a process of sequential choices between every two values and therefore each choice could be modelled as an ordinary logistic regression.

6.1.4 Enriched information about hedges

Since this research resulted in producing a set of new-found lexical devices to convey hedging, the next step is to ensure validation of these lexical resources is by the manual annotation by multiple individuals. The main categories to be addressed in this task would be Single hedges and Not-Claiming-Knowledge epistemic phrases. The reason is they are more frequent and have a more regular morphology than Syntactic and Other hedges. The degree of agreement by annotators could be then measured by applying statistics similar to Cohen's Kappa statistic.

Likewise, more detailed information about hedges can be extracted from the current annotated dataset. There are at least two possible paths to follow: First, a characterisation that addresses the lexical and syntactic functions of hedges, and other one that regards the hedge's position in a sentence. A deeper lexical-syntactic characterisation of hedges in informal language style would be the next step toward building an automatic system for hedge detection.

Regarding Single hedges, the current study used as a starting point the work of Rubin et al. [2005a]. Further studies on grammatical information for hedges could be made in a

similar way to Rubin et al.. Grammatical patterns could be derived from some hedge types, particularly from hedges belonging to the Not-Claiming-Knowledge category. Throughout this document, I emphasized that Not-Claiming-Knowledge epistemic phrases that are verb-centered follow a pattern composed of a first person pronoun plus a hedging particle (cf. Section 3.3.4 and Table 3.3 in particular). Extended NCK epistemic phrases include modifiers such as adjectives and adverbs, therefore it is a productive category of hedges. Syntactic and semantic patterns created from these phrases could be used to devise methods for automatic hedge identification, methods based on adaptive machine learning in particular and high recall-savvy methods in general. In Section 6.2.3, I provide some examples of how these patterns could apply to online conversations in other domains.

Regarding a hedge's position in-sentence, information about the hedge's scope or whether a phrasal hedge is embedded within another hedge scope in the same sentence. Clause dependencies could also be explored if syntactic sentential structures are explored.

Co-occurrence of one hedge with other hedges in-sentence is also worth exploring as it looks very likely than some hedge types are used often in conjunction. For instance in (208), the NCK phrase *I am not sure* and the Syntactic hedge *if* co-occur and this is a common construction. It would be, nonetheless interesting to find out which other co-occurring patterns are frequent.

(208) I am not sure if he went to the conference.

6.1.5 Hedges and Questions

Inherently, interrogative questions are used to express doubt. In this study I have not analysed hedging in questions. This would pose a future path of research to find out if cues of hedging play a different role than in non-interrogative sentences. Consequently, the fact that interrogative sentences appear in a post would possibly affect the labelling strategies based on hedge occurrence (cf. Section 5.4). Questions in posts could also be used as variable to improve the fitting of statistical regressions models alongside the variables that represent the use of hedges in a post.

6.1.6 Degrees of uncertainty

In this study, I have not addressed the distinction between levels of uncertainty conveyed by hedges. In contrast, other studies such as the one conducted by Rubin et al. [2005a] explored the gradation of certainty/uncertainty and associated lexical items. A similar taxonomy could be used to characterise the hedging intensity in different hedge types. For instance, there is a clear difference between *I don't know* and *I hope* in sentences (209) and (210), since *I don't know* signals more uncertainty than *I hope*, which can be used also to provide a suggestion in a polite way.

(209) **I don't know** how to configure this software application.

(210) **I hope** it helps.

6.2 Applications of automatic hedge detection

6.2.1 Sentiment analysis

Hedges are mostly deemed as expressions conveying neutral opinions. From the vast state of the art research in sentiment analysis, Wiebe et al. [2005] is the only work that has taken into account speculative expressions alongside linguistic expressions of positive and negative sentiment. The annotation of both kinds of expression are considered as features for the automatic detection of opinions. Following such approach addressing social media texts, suggests an interesting research path as linguistic expressions of sentiment showed to be a productive category in informal language.

6.2.2 Hedging in dialogue

The dataset used in this research was extracted from a web forum, a natural dialogue environment. However, analysis of hedges was circumscribed to individual posts. Some superficial profiling of how hedges are used in dialogue dynamics are carried out by looking at whether a post is either starting a conversation thread or it is answering to another post in the thread.

Further work could focus on the role that hedges play in dialogue and how interlocutors react (linguistically) to propositions where hedges are used. As seen earlier, Kärkkäinen [2003] showed how epistemic phrases are used to express personal stance in a spoken language context. A similar approach could be followed in written dialogue having Not-Claiming-Knowledge epistemic phrases as linguistic cues.

For instance, the dialogue in (211) shows interaction between three users in a scenario of Advice-Seeking/Advice-Giving/Commenting.¹ Each of the users' contributions contains some sort of hedge. Studying hedges in a dialogue situation was not viable in my research since one drawback of the dataset used in this research is the lack of information about which post is being addressed when a user posts a contribution. In some cases, this information is available by interpreting the post's contents. For instance in the post by USER3 below, he or she addresses USER1 in (211) and later USER3 addresses USER2. However, this information is not encoded in any other manner in the platform supporting the web forum. This limitation makes necessary the devise of methods to identify peer-to-peer dialog participants prior to analyzing use of hedges or any other kind of phenomena in a dialogue environment.

(211) USER1: I have always used [company_brand_name] products. This year I downloaded the [product1] trial version. I find it similar to [product2], but it has clean up and back up options which I do myself anyway. Don't want to pay extra for this. I **tried** to download the [product3] trial version to compare the time hog, as [product3]

¹cf. introductory notes in Section 1.1.2

claims to perform faster than previous versions. [product1] **seems** to perform faster than previous versions of [company_brand_name] that I have used.

I am unable to download the [product3] trial version, because I get a message saying I have [product1] installed which has similar properties.

Does anyone know a way around this? Do I have to delete [product1] trial?

or Does anyone know the answer to my question? Is [product3] as slow as [product2]?

thank you in advance

Post: 731

USER2: Hi USER1, I'm afraid that [company_brand_name] and N360 can not coexist on the same machine (unless you have virtual machines set up in VMWare or the likes).

I'll give you a short and a long answer to the [product3] vs. NIS 2007 question

Short Answer: NIS 2008 is faster than NIS 2007 😊

Longer answer: It's reasonably fair to say that up to the 2006 versions of our products, we were adding features at the expense of performance. Everything we have done since then has been done with performance in mind. [product2] was faster than [product4], and in turn [product3] is faster than [product2]. Likewise [product1] 2.0 is less intrusive than 1.0.

We are not compromising security to enhance performance. We are listening to feedback and making intelligent decisions about what needs to be there for everyone, vs. what **some** people need for specific reasons. All this, plus **some** hefty code optimization.

The add-on pack concept is a case in point (your choice to add free features, at the expense of **a little** performance):

[quotation]

I've worked here **over** 12 years and I've never seen such focus and dedication put into driving better and better, lighter and lighter code.

USER2

USER1: Which product **would** you recommend for an advanced user, ie .. does regular backups, uses utilities to clean PC, (running WXP2) ?

Would you recommend NIS or 360?

The add on pack I download is antispyware component only. I like this feature. 😊

Don't need parental controls now.

USER3: [quotation]

USER2, This is a bit OT but I really appreciate your honesty in the above appraisal of NIS evolution over the years. And I agree completely - [product3] is very unobtrusive and does not appreciably slow down my machine. I appreciate the efforts that [brand_name] has put into improving this product. And I am glad to finally have these forums in which to state my thoughts, questions and comments. Thank you.

USER3

Post: 775

USER2: Hi USER1, **If** you're already doing the tune-up and backup work using other apps (**I hope** I am understanding your request correctly), **I would** definitely go for NIS rather than N360.

USER2 Post: 776

Where the addressee's identity can be known, hedges could be used as linguistic features for modelling behaviour in dialogue situations. One example is the use of hedges as topic independent features to detect agreement or disagreement in social media dialogue [Misra and Walker, 2013]. Hedges as features contribute to the improvement of a classifier if it built on top of them, therefore this may be improved by using a larger lexicon of hedges.² Similar approaches can be followed to improve this research or to explore new topics related to dialog acts that could be affected by linguistic hedging.

6.2.3 Extending Analysis to Other Domains and Social Media Platforms

This research was carried out over a dataset extracted from a specific web forum. Insights from observations can be generalized to happen in similar contexts, ie. web forums that have similar informal tone to the one addressed in this research. These insights could be contrasted with findings from datasets extracted from such other web forums. This comparative analysis would have a two-fold purpose: of verifying the conclusions from this study are prevailing in other domains, and of extending lexicons of linguistic hedging expressions.

For instance, some hedging expressions belonging to the category Other (cf. Section 3.3.6) have a NCK-like form that is circumscribed to the specific forum domain and some generalizations could be made about their underlying syntactic-semantic pattern. A quick examination in other web forums shows that NCK phrases and NCK-like phrases are also found in posts covering a different topic such as in (212) and (213). Other epistemic expressions have an even more informal style such as in (214) and (215). Other expressions such as in (216) are very domain specific, although, a pattern can be detected there, as a cliché expression for lack of knowledge (e.g **I am a videogames illiterate**, **I am a origami illiterate**, etc.)

(212) **I am clueless**, please help!

(213) **I have NO idea** what to wear with my suit! ³

(214) **I am a COMPLETE noob** at programming.⁴

(215) **I am a newb** at buying hoodies. Advice appreciated.⁵

(216) ... just got a pair of casual nike, **i am a shoe illiterate**. How are these? ⁶

²Personal communication. September, 2013.

³<http://www.styleforum.net/t/235377/i-am-clueless-please-help>

⁴Post subject in <http://forum.arduino.cc/index.php?topic=225484.0;wap2>

⁵<http://www.styleforum.net/t/38645/i-am-a-newb-at-buying-hoodies-advice-appreciated>

⁶<http://www.styleforum.net/t/128230/just-got-a-pair-of-casual-nike-i-am-a-shoe-illiterate-how-are-these>

The collected lexicon could be used to study hedging and opinion analysis in other social media platforms such as microblogs (i.e. Twitter⁷ and Tumblr⁸). Users from these platforms characteristically include hashtags (words preceded by a # symbol) terms or phrases in their posts as keywords that depict a post's content or to provide a subtext, although other uses such as adding metadata to the post's original content and for publicity purposes have been noted [Cunha et al., 2011].

A preliminary search of hedges formatted as hashtags showed interesting results in terms of variety, quantity and quality. For instance, it was possible to find posts (i.e. "tweets") marked with hashtags such as #imho, #maybe, #dunno and #idk to mention but a few. Nonetheless, it was possible to find *tweets* the equivalent of NCK phrases: #iam-clueless, #iamanewbie, #ihope, #ithink, #iamnotsure, #iassume, etc. which have a potential speculative intention.

As these platforms are used as tools to express people's opinions or personal stance and NCK epistemic expressions epitomize first person stance, it seems worth to explore the role of NCK hedges play in this kind of content. However, users' intentions in posts published through social media platforms may differ from the ones in technical web forums and therefore may not necessarily subscribe to the poly-pragmatic model features that served as basis for my research.

6.3 Conclusions

In this chapter, I have outlined two different directions of research that the current study could take. The first path of study involves addressing shortcomings and limitations found in my research in order to build an adaptive method for the identification of hedges in informal language style. The suggested topics may not include the whole range of requirements to create such a method, nonetheless, they are the main ones to be addressed. The second path of future work includes suggestion for applications that encompass either the use of the created lexicons of hedges or extended research of hedging in various domains.

⁷<http://twitter.com>

⁸<https://www.tumblr.com/>

Chapter 7

Conclusion

In this chapter, I will highlight the main original contributions of this research in Section 7.1 and will summarise the findings and conclusions in Section 7.2.

7.1 Contributions

- A new categorization scheme of hedges for informal language style was created. The scheme was created based on empirical observations of in-domain language samples extracted from a particular online web forum, analysis of out-of domain state-of-the-art findings and existing literature around the topic of hedging. The scheme comprises four categories of hedges, so called: Single hedges, Non-claiming-knowledge epistemic phrases, Syntactic hedges and miscellaneous category called Other hedges. The scheme also consider the annotation of the Source and Scope of the hedging expression. (Chapter 3);
- Thorough empirical and theoretical insights about a new category of hedges was provided. I named this new category Not-claiming-knowledge first person epistemic phrases, it is compliant for automatic processing and enriched semantic interpretations in domains where informal language style is used (Section 3.3.4 and Section 4.2.2).
- A new lexicon of hedging words and phrases found in informal online conversations was created. This lexicon of hedges is divided into categories according to the categorization scheme, therefore is built around subsets of Single hedges, Non-claiming-knowledge epistemic phrases, Syntactic hedges, and Other hedges. (Section 4.2);
- Empirical findings of how hedges and the entities associated with them (Source and Scope) occur in the web forum dataset were presented (Chapter 4);
- Pragmatic interpretations of the main hedging categories were described around the concepts of content-orientation and reader-orientation (Section 4.5);

- An annotation procedure for hedges in this domain was created and described (Section 3.4.3).
- Insights of how hedges occur in relation to other meaningful forum post features that aim to identify the characteristics of outstanding post contributions were provided, including statistical models that account for high ratings given to posts according to hedge use (Chapter 5).
- A dataset of web forum posts manually annotated with expressions of hedging, according to the proposed annotation scheme, was created.
- Future paths of work in this particular research topic and suggestions for the automatic processing of hedges in informal domains were described (Chapter 6).

7.2 Summary of concluding remarks and research findings

7.2.1 State of the art in the study of hedging and automatic methods

Besides the original conception proposed by Lakoff of hedges as linguistic devices to define criteria for membership in definitional categories of concepts, subsequent studies have taken two conceptions: one where hedges focus on the commitment expressed in a proposition, and other one where the focus is on expressing some degree of uncertainty.

Hedges have been studied in various particular aspects such as whether they have a sentential or sub-sentential modal meaning, according to their lexical and grammatical categories, studying them as expressions of epistemic modality, and according to their pragmatic functions. Most of these studies have been done on formal language registers such as in academic prose.

Schemes for the annotation of hedges vary from ones based on hedges as minimal units to schemes where various elements involved in the annotation of hedging are considered such as the degree of uncertainty conveyed by hedges and the entity experiencing or originating the uncertainty or modality that was conveyed by a hedge.

A pragmatic taxonomy of hedges was highlighted in this dissertation as it allows to interpret hedges according to reader-orientation and content-orientation.

The emphasis in empirical research of hedging expressions and in most automatic methods for hedging identification has been put into formal academic language style. Nonetheless, automatic method for identification of hedges and their associated entities still present caveats and I have shown the limitations of current automatic methods and in conceptual models of hedging that made them not compatible with hedging in other domains such as web forums. One of the main limitations is that despite it is clear that the Source of hedging is not always the writer, existing automatic methods are not concerned with identifying this element because they focus on the propositional content (content-centered). In other scenarios such as in web forums where prominent users need to be identified ('user-centered'),

it is imperative to determine whether or not a specific user is the one expressing his or her personal stance.

My research considered the limitations in porting current methods and approaches for identification of hedging to informal style of language by addressing the construction of resources such as: an annotation scheme, and lexicons of hedging expressions tailored to informal language style.

7.2.2 Annotation scheme for hedges in informal language style

Automatic identification of hedges creates the need for an annotation scheme comprising all elements involved in hedging, which are: the hedging expression, its source and its scope.

This annotation scheme aims to be useful in both content and user-centered approaches on the study of hedges, and on automatic identification of hedging expressions in informal language style.

Four categories of hedges in informal language were defined: SINGLE-hedges, NOT-CLAIMING-KNOWLEDGE epistemic phrases, SYNTACTIC and OTHER hedges. Guidelines on how to proceed on particular and exceptional cases were described along with each category.

SINGLE-hedges mostly conform to the concept of epistemic modals or traditional hedges such as *may*, *probably* and *likely*.

NOT-CLAIMING-KNOWLEDGE (NCK) epistemic phrases or hedges comprise expressions such as *I think*, *I don't know*, and *I would suggest*, since they have semantic and pragmatic interpretations different from epistemic phrases and other traditional categories of hedges.

SYNTACTIC hedges mainly comprise conditionals. I have followed up the classification of conditionals made by Iatridou: relevance, factual and hypothetical conditionals, which discussed providing representative examples of each kind in cases they could be deemed as signals of hedging.

OTHER hedges comprise hedges that could not be either classified into any of previous categories or that have the structure of a NON-CLAIMING-KNOWLEDGE hedge but are specific-domain such as *I'm a computer illiterate*.

I defined two types of Source: Inner Epistemic Source and Outer Epistemic Source. The Inner Epistemic Source refers to the individual or individuals experiencing a mental state that translates into a hedging expression, while the Outer Epistemic Source is the individual who wrote a proposition (Writer). Some criteria to distinguish these two types of Source have been mainly proposed according to how they occur in the language style under study. The annotation scheme provides the means to annotate distinctively when the Inner epistemic source is not the writer.

In the proposed annotation of the Scope, its constituents are separated from the hedging expressions, in contrast to earlier studies where the hedge was annotated within the scope boundaries. The benefit is that lexical constituents that do not actually form part of the

scope can be left out from being annotated.

The manual annotation included some semi-automatic procedures and it was designed as an iterative procedure where annotations, and the annotation template could be refined and formatted according the language style being addressed. I organized the annotation procedure into these main steps: pre-processing, pilot annotation, pre-annotation and manual annotation.

Pre-processing comprises common pre-processing steps in language processing systems, such as sentence splitting and tokenization. Other steps involved in pre-processing are the ‘cleaning’ of non-textual elements and normalization of extra-linguistic (e.g. images) and pseudo-linguistic textual elements (e.g. emoticons).

The pilot annotation had the purpose of performing a preliminary corpus linguistic study of the dataset, also based on study of the state of the art around hedging. Pre-annotation included an automatic marking in the required annotation tool format of entities according to the initial lexicons of hedges.

The manual annotation itself comprised checking over pre-annotated entities representing hedge occurrences, finding new ones and marking other elements such as the source and scope of the hedging expression. Additionally, some manual annotation strategies were described to improve to some extent the quality of annotations made by a single annotator.

7.2.3 Not-Claiming-knowledge expressions of hedging

I have provided linguistic support for the consideration of Not-Claiming-knowledge epistemic phrases as an important category of hedges in informal language style that have different qualities from hedges from the epistemic modality tradition. I have described some subjective and objective distinctions in this category to account for some cases where it seems that categorical assertions are done in hedging expressions. I have also presented a strong linguistic foundation of the subjective role in first person epistemic phrases and distinction between subjective and objective uses of epistemic modality. These distinction shows that the group of Not-claiming-knowledge hedge comprises epistemic phrases expressing weak commitment and epistemic phrases expressing lack of commitment to the claim of knowledge. One of the main advantages is that in Not-Claiming-Knowledge hedges the source is enclosed within the hedging expression.

I have continued the discussion started in Section 3.3.4 about the distinction between subjective and objective epistemic modality, by showing with lexical findings of what seems to be a claiming-knowledge component in Not-Claiming-knowledge phrases such as *I don't know*, and comparing them overall to distinctions between categorical and hedged assertions. With respect to this point, I conclude that in NCK phrases, the focus of the interpretation of hedging is divided between the source and what is being hedged, in contrast to Single-hedges where only what is being hedged is under scrutiny. This particular feature of NCK phrases emphasizes what has been suggested in the literature about their difference from other types of epistemic expressions. Moreover, empirical findings reinforce the idea

that first person epistemic phrases is a distinctive semantic category of hedges.

I have shown that NCK phrases have less ambiguous occurrences compared to Single-hedges which suggests they can be used as improved types of hedges that convey less ambiguity and can be used in datasets from other non-explored domains since they would require less automatic natural language processing resources such as parsers. Parsing and similar linguistic tasks are quite accurate in formal styles of language but they still have challenges to overcome in language styles that are noisy.

Overall, I have observed that the use of NCK can lend itself to richer interpretations than other kinds of hedges taking into account other posts features than from the use of Single hedges, since NCK is a more complex category that takes into account the user's involvement, intentions and mental state. For instance, the fact that NCK epistemic phrases are written in the first person makes it straightforward to interpret them according to features in distinct user categories e.g. employees are less likely to use this kind of expression because the characteristics of their position in the community.

7.2.4 Lexicons of hedging expressions

I have presented and described a lexicon of hedges comprising words and phrases used for speculation and other hedging functions. Lexical hedging types belonging to the following four categories: Single hedges, Not-Claiming-Knowledge first person epistemic phrases, Syntactic and Other hedges.

Overall I found 790 unique types of hedges, 272 of them belong to the Single hedge category, 300 to NCK phrases, 8 to Syntactic and 209 to Other hedges.

I have mainly described some normalization techniques for Single hedges and NCK epistemic phrases, potentially useful for detecting hedges in user generated content particularly.

Some normalization techniques applied to Single and NCK hedge categories cause grouping of equivalent types that were lexically different because of typos, tense and number variations, abbreviations, non-standard forms and colloquialisms. Single hedges were normalized from 270 to 189 types and NCK were normalized from 303 to 138.

Single hedge types reflect mostly what is found in literature and in previous hedging studies. NCK epistemic phrases has a wider set of lexical realisations, for instance informal expressions such as *I don't know*, acronyms such as *IMO* and elliptical cases (eg. *hope, not sure*), these lexical items are frequently used in social media content.

The two remaining category of hedges were less extensively addressed since either their types are quite regular (Syntactic) or quite heterogeneous (Other). The group of Other-hedges is mainly composed by NCK-like epistemic phrases but whose content is tailored to the domain of the dataset under study, for instance *I'm not really techie enough* and other miscellaneous types. They could be built into hedging patterns taking into account terminology from a specific domain.

Lexical types that would potentially convey a hedging meaning but were not actually

being used as such (ie. non-hedges) were also described.

7.2.5 Other important elements in hedging

As expected it was found that most NCK hedges have an explicit Source (89.9% of occurrences) in comparison to Single-hedges (13.14%). As expected, the source for 99.99% of all NCK occurrences is the writer, while for Single hedges, 2.53% of explicit Inner epistemic Sources is attributed to another individual that is not the writer, and in 16.51% of the cases this source is explicit in the post. The most frequent hedge types whose source is not the writer are variations of *suggest*: [*suggestions, suggested and suggestion*]

Further, I have pointed out some possible caveats in the manual identification of the scope of hedges, such as when the scope is not evident in the sentence or when there is the possibility of attributing a hedge a scope that actually does not correspond to it. I have found that 18.26% of hedge occurrences do not have a scope in-sentence, being Single-hedges the ones that have the highest frequency, for instance *based on, somebody, and strangely*. I have also identified cases where a hedge scope comprises another hedge. These findings could be further explored in the sense of studying interactions between hedges subordinated to other hedge types.

I have provided numerical descriptions of source and scope that illustrate the variety of hedge realisations in this informal style of language. Regarding the scope of hedges, it was shown and discussed how this is not solely determined by syntactic features in-sentence but by the semantic of certain hedge types.

7.2.6 Use of hedges in web forums

Co-occurrence of hedges in each sentence was measured: 10,27% of the sentences have at least two hedges within their boundaries. The most frequent co-occurrence is of two Single-hedges per sentence and the most frequent combination of two hedging categories is where one Single-hedge and one NCK appear in one sentence.

I have discussed using linguistic examples how the pragmatic categories proposed by Hyland match the function of hedges in the domain under study. I compared the intentions of academic writers with the ones from forum contributors, emphasizing these are different from occasional visitors.

The main categories of hedges analysed are: Attribute, Reliability, Writer-oriented and Reader-oriented hedges. I have described frequent specific situations where hedges are used and could be matched to these categories. For instance, attribute hedges are frequently used to make accurate descriptions of problems that make users seek answers in the forum. Some types of NCK phrases are used in ways that could match reliability and reader-oriented hedges, particularly in the latter when they look for reader's acceptance. One striking difference that leads hedges into particular categorization is that in research writing, authors are discouraged from overusing first person with the intent that focus remains on the research topics; such limitation do not exist in web forums, so it could not be said that in

web forum writer-oriented hedges are used as often as in academic articles. I believe that individuals in the web forum use NCK hedges to prevent criticism, and therefore these are representative of reader-oriented hedges. However, in some senses they are not equivalent since many users seeking advice in the forum are not afraid of admitting lack of knowledge.

I have also emphasized the user of hedges in comparison to categorical assertions in cooperative tasks between Advice-seekers and Advice-givers. I noted as well that individuals using hedges in some cases make imprecise descriptions as they do not think the situation needs to be accurately described.

Besides analysis on the use of hedging regarding their occurrence in posts and pragmatic uses, their use with respect to other pseudo and non-linguistic features was also explored

With the motivation of exploring these user's characteristics in posts from the web forum community under study, I proposed posts categorizations based on its author's user category, ratings given to it and polarity of sentiment in the post to analyse their correlations to hedges occurring in posts.

I have described posts according to each of these categorizations and based on them, I proposed some statistical models with an explanatory intent, since the main purpose is related to the question of how hedges are used in this domain and how they correlate with other web posts features, and whether the use of hedges may aid in distinctively characterising users playing particular roles in the web forum community.

There are two types of criteria used in web communities for promoting users to higher ranks: de facto qualities which are descriptive according to pre-defined roles of users, and dynamic qualities that emerge from the user's participation in the forum such as frequency of visits, and the quality of their posts.

Considering the domain under study, four user categories were defined: *employees*, *gurus*, *ranked*, and *unranked* users. Each post is labelled according to category its author holds. *gurus* are the most prestigious users in the community, *ranked* users are individuals actively contributing to the community but they still not qualified to be *gurus*, while *unranked* users are the ones who have shown only a little amount of contribution.

I chose ratings given to posts as a feature that measures the post quality by the community peers and categorize them according to it. A user assigns high ratings ('kudos') to posts that he or she considers as being insightful, useful, or because, all things being equal, he or she just likes the post. Nonetheless, the reasons for a user to give kudos to posts are essentially subjective.

I found that only 7% of posts in the Annset and 9% in the RTD have been given kudos at least once (*kudoer*), considering that each user may give kudos to a post only once. Nonetheless, this is congruent with the idea, that kudoer posts are outstanding contributions.

The polarity of sentiment in posts is important because the trust that makes users participate in online communities comes from their perception of expertise and benevolence in members of such a community. I chose a set of emoticons from earlier research that signal a particular polarity of sentiment. I categorized each post as being either positive, negative or neutral. This feature also proved to be sparse in the datasets; 8% and 9% of posts have

at least one emoticon in the Annset and RTD respectively. I found out that *ranked* users include emoticons most frequently in their posts.

I have formulated four hypotheses that led the analysis of the proposed statistical models: a) Individuals from different categories do not use hedges according to the same qualitative and quantitative patterns, b) including hedges in a post increases its likelihood of receiving kudos, c) NCK epistemic phrases in a post lead to an increased likelihood of being assigned kudos, and d) the use of hedges in posts alongside other features increases the likelihood of more users giving kudos to the post.

Posts from *gurus* and *ranked* users are the ones that receive kudos more frequently, while *unranked* users have the least proportion of kudo-receiving posts. Therefore, kudos was deemed as a proxy category in the statistical models proposed as the higher the proportion of kudos in posts, the higher expertise they appear to be signalling.

I found out that the sole occurrence of hedges is not particularly a good criterion to distinguish between posts from *gurus* and *unranked* users. Notwithstanding, there are some useful qualitative observations such as: on average, *employees* use fewer NCK-hedges in their posts, probably due to the fact their communicative style is less likely to include expressions in the first person. Posts by *gurus* include Other-hedges less frequently, since is not likely they are going to utter domain-tailored expressions admitting lack of knowledge.

The statistical models I proposed were modelled over the RTD as it has representative features in a higher frequency than in the Annset.

Based on three logistic regression models that have as predictor a variable representing the occurrence the different categories of hedges in posts, a Tukey test for pairwise comparison of means was applied to identify relevant statistical differences between groups of posts categorized according to hedges, and their likelihood of being associated with kudoer posts.

These models showed that posts containing exclusively one category of hedges share the same likelihood of having kudos. Most of the relevant groups of posts that have a combination of the main types of hedges have more likelihood of having kudos attributed to them than UNHEDGED posts. However, there is not significant evidence to support the hypothesis that posts exclusively containing NCK hedges are more likely to have kudos awarded than UNHEDGED posts. More heterogeneous posts according to the hedging categories they contain are more likely to get kudos than posts with less variety of hedge types. These models highly suggest that hedges included in posts increase their probabilities of getting kudos assigned. However, the contribution of NCK hedges to this likelihood could not be proven.

Further, I have proposed a logistic regression model whose dependent variable is a continuous representation of kudos, to account for the scenario where a post can be assigned kudos repeatedly, as each user from the community can potentially give kudos to this post.

Since only a minimal percentage of posts in the dataset were assigned kudos at all, this scenario was modelled as a Zero-inflated Negative Binomial distribution where this lack of kudos comes from two possible sources: either an individual has not viewed a post and therefore not given kudos to it, or the individual has viewed the posts but deemed it not

worthy of kudos. Besides the occurrence of hedges, additional predictors are the number of days a posts has been online, the number of words and the number of views a post has been given.

The Zero-inflated negative binomial model fits the data better than similar Zero-inflated Poisson and regular Negative binomial models accounts for zeros in the kudos-given variable, that means that they were not caused because users did not deem it worthy of kudos. Also, this model was shown to fit the data better than a model where the variable representing hedges is not included.

Finally, I built two logistic regression models that have a mixed predictor variable whose values come form the interaction of the occurrence of hedges and polarity of sentiment in posts. One has a binomial representation of hedges in a post (hedged, unhedged) and the second one has values assigned according to the main categories of hedges Single-hedges and NCK phrases.

These models showed that whether a post is HEDGED or UNHEDGED, it is more likely to be awarded kudos if it expresses positive sentiment in comparison to showing negative polarity of sentiment. Overall, posts that are UNHEDGED and have no emoticons within are the least likely to have kudos assigned, alongside with posts that are hedged and express negative sentiment. I have also shown that when a post has positive polarity of sentiment, its chances of having kudos assigned increases if it has hedges.

7.2.7 Future research paths

Future paths of research described in this dissertation comprised two groups of insights. I one, suggestions for improving methods on automatic detection of hedges in informal styles of language are provided. Many of these suggestion emerge from observed limitations of the research described in this document. The second group of insights address the potential use of hedges as features for natural language processing tasks in online communities, for instance for sentiment analysis in micro-blog platforms.

Bibliography

Santa barbara corpus of spoken american english, part 1, 2000. URL <http://www.linguistics.ucsb.edu/research/santa-barbara-corpus>.

in my opinion, 2002. URL <http://idioms.thefreedictionary.com/In+My+Opinion>.

Lisa C Abrams, Rob Cross, Eric Lesser, and Daniel Z Levin. Nurturing interpersonal trust in knowledge-sharing networks. *The Academy of Management Executive*, 17(4):64–, 2003. URL <http://proquest.umi.com/pqdweb?did=923892131&Fmt=7&clientId=4574&RQT=309&VName=PQD>.

Alfred V. Aho and Margaret J. Corasick. Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June 1975. ISSN 0001-0782. doi: 10.1145/360825.360855. URL <http://doi.acm.org/10.1145/360825.360855>.

Karin Aijmer. Epistemic predicates in contrast. In Stig Johansson and Signe Oksefjell, editors, *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, pages 277–295. Rodopi, Amsterdam, 2000.

Alexandra Aikhenvald. *Evidentiality*. Oxford University Press, Oxford, 2004.

Aristotle. *The Art of Rhetoric*. Penguin Classics, London, trans. 1991. Translated with an Introduction and Notes by H.C. Lawson-Tancred.

R. H. Baayen. *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge, U.K., 2008.

Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics*, page 37–47. Springer, 2013. URL http://link.springer.com/chapter/10.1007/978-3-319-03689-2_3.

Albert Bifet, Geoffrey Holmes, and Bernhard Pfahringer. Moa-tweetreader: real-time analysis in twitter streaming data. In *Proceedings of the 14th international conference on Discovery science, DS'11*, pages 46–60, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-24476-6. URL <http://dl.acm.org/citation.cfm?id=2050236.2050243>.

Ramona Bongelli, Carla Canestrari, Ilaria Riccioni, Andrzej Zuczkowski, Cinzia Buldorini, Ricardo Pietrobon, Alberto Lavelli, and Bernardo Magnini. A corpus of scientific biomedical texts spanning over 168 years annotated for uncertainty. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Virginia Mary Brennan. *Root and epistemic modal auxiliary verbs*. PhD thesis, University of Massachusetts Amherst, Amherst, January 1993.

P. Brown. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics, 4. Cambridge University Press, 1987. ISBN 9780521313551.

Ran Cheng and Julita Vassileva. User motivation and persuasion strategy for peer-to-peer communities. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences - Volume 07*, HICSS '05, pages 193.1–, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2268-8-7. doi: 10.1109/HICSS.2005.653. URL <http://dx.doi.org/10.1109/HICSS.2005.653>.

Maarten Clements, Arjen P. De Vries, and Marcel J. T. Reinders. The task-dependent effect of tags and ratings on social media access. *ACM Trans. Inf. Syst.*, 28(4):21:1–21:42, November 2010. ISSN 1046-8188. doi: 10.1145/1852102.1852107. URL <http://doi.acm.org/10.1145/1852102.1852107>.

Peter Crompton. Hedging in academic writing: Some theoretical problems. *English for Specific Purposes*, 16(4):271 – 287, 1997. ISSN 0889-4906. doi: [http://dx.doi.org/10.1016/S0889-4906\(97\)00007-0](http://dx.doi.org/10.1016/S0889-4906(97)00007-0). URL <http://www.sciencedirect.com/science/article/pii/S0889490697000070>.

Evandro Cunha, Gabriel Magno, Giovanni Comarella, Virgilio Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: A language-based approach. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 58–65, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-96-1. URL <http://dl.acm.org/citation.cfm?id=2021109.2021117>.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. University of Sheffield, Department of Computer Science, 2011. ISBN 978-0956599315. URL <http://tinyurl.com/gatebook>.

- Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. TiMBL: Tilburg Memory—Based Learner. Version 6.3. Reference guide. Technical Report ILK 10-01, Induction of Linguistic Knowledge Research Group, Tilberg University, The Netherlands, 2010. URL http://ilk.uvt.nl/downloads/pub/papers/Timbl_6.3_Manual.pdf.
- William M. Darling, Michael J. Paul, and Fei Song. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 1–9, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=2389969.2389970>.
- Paul Egré and Denis Bonnay. Vagueness, uncertainty and degrees of clarity. *Synthese*, 174(1):47–78, 2010. URL <http://link.springer.com/article/10.1007/s11229-009-9684-8>.
- Paul Egre. Reliability, margin for error and self-knowledge. *New waves in epistemology*, page 215–250, 2008.
- Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, page 359–369, 2013. URL <https://www.aclweb.org/anthology/N13/N13-1037.pdf>.
- Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W10-3001>.
- C. Fellbaum. *Wordnet: An Electronic Lexical Database*. Cambridge, US: The MIT Press, 1998.
- Peter Filzmoser, Robert G. Garrett, and Clemens Reimann. Multivariate outlier detection in exploration geochemistry. *Comput. Geosci.*, 31(5):579–587, June 2005. ISSN 0098-3004. doi: 10.1016/j.cageo.2004.11.013. URL <http://dx.doi.org/10.1016/j.cageo.2004.11.013>.
- Edward Finegan. *Language : its structure and use*. Harcourt Brace Jovanovich, Sydney :, australian ed. edition, 1992. ISBN 0729512681 0729512681.
- Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef VanGenabith. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proceedings of IJCNLP*, Chiang Mai, Thailand, 2011. URL <http://doras.dcu.ie/16854/>.

- S. E. Francis and H. Kucera. *Manual of Information to accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers*. Brown University, Providence, Rhode Island, Revised 1989.
- Bruce Fraser. Pragmatic competence: The case of hedging. In *New approaches to hedging*, page 15–34. 2010.
- Grégory Furmaniak. On the emergence of the epistemic use of must. *SKY Journal of Linguistics*, 24:41–73, 2011. URL http://www.linguistics.fi/julkaisut/SKY2011/Furmaniak_netti.pdf.
- Viola Ganter and Michael Strube. Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176, Suntec, Singapore, 2009. URL <http://www.aclweb.org/anthology/P/P09/P09-2044>.
- Maria Georgescu. A hedgehop over a Max-Margin framework using hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, page 26–31, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W10-3004>.
- Nikolas Gisborne and Jasper Holmes. A history of english evidential verbs of appearance. *English Language and Linguistics*, 11(01):1–29, 2007. doi: 10.1017/S1360674306002097. URL <http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=933448&fulltextType=RA&fileId=S1360674306002097>.
- Michel Génèreux, Iris Hendrickx, and Amália Mendes. Introducing the reference corpus of contemporary portuguese online. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- Erving Goffman. *The Presentation of Self in Everyday Life*. New York: Doubleday, 1956.
- Valentine Hacquard. *Aspects of Modality*. PhD thesis, Massachusetts Institute of Technology, 2006. URL <http://dspace.mit.edu/handle/1721.1/37420>.
- M. A. K. Halliday. Functional diversity in language as seen from a consideration of modality and mood in english. *Foundations of Language*, 6(3):pp. 322–361, 1970. ISSN 0015900X. URL <http://www.jstor.org/stable/25000463>.
- Iris Hendrickx, Amália Mendes, and Silvia Mencarelli. Modality in text: a proposal for corpus annotation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry

- Declerck, Mehmet U?ur Do?an, Bente Maegaard, Joseph Mariani, Jan Odiijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- Aurelie Herbelot and Ann Copestake. Annotating genericity: How do humans decide? (a case study in ontology extraction). In Sam Featherston and Winkler, editors, *The Fruits of Empirical Linguistics I*, volume Volume 101 of *Studies in Generative Grammar*, pages 103–122. Mouton de Gruyter, Berlin, 2008. ISBN 978-3-11-021338-6. URL www.cl.cam.ac.uk/~ah433/Annotating_Genericity.pdf.
- H. Hersh and A. Carmazza. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, 105(3):254–276, 1976.
- Nguyen C. Ho and Huynh V. Nam. An algebraic approach to linguistic hedges in zadeh's fuzzy logic. *Fuzzy Sets and Systems*, 129(2):229–254, 2002.
- Janet Holmes. Doubt and certainty in ESL textbooks. *Applied Linguistics*, 9(1):21–44, 1988. doi: 10.1093/applin/9.1.21. URL <http://applij.oxfordjournals.org/content/9/1/21.abstract>.
- Paul J. Hopper. On some principles of grammaticization. In *Approaches to Grammaticalization*, *Approaches to Grammaticalization*, pages 17–36. John Benjamins, 1991. ISBN 9789027228956.
- Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.
- K. Hyland. *Hedging in Scientific Research Articles*. Pragmatics & beyond. John Benjamins Publishing Company, 1998. ISBN 9781556198168.
- Ken Hyland. Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics*, 17(4):433–454, 1996. doi: 10.1093/applin/17.4.433. URL <http://applij.oxfordjournals.org/content/17/4/433.abstract>.
- Sabine Iatridou. *Topics in Conditionals*. PhD thesis, MIT, Cambridge, Massachusetts, 1991. Distributed by MIT Working Papers in Linguistics.
- Shaili Jain, Yiling Chen, and David C. Parkes. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM Conference on Electronic Commerce, EC '09*, pages 129–138, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-458-4. doi: 10.1145/1566374.1566393. URL <http://doi.acm.org/10.1145/1566374.1566393>.
- J.F. Janssen and C. Vogel. Politics makes the swedish :-) and the italians:-(. In *Proceedings of LREC – EMOT 2008, Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology*, page 53, 2008.

- Thorsten Joachims. Making large-scale support vector machine learning practical. pages 169–184, 1999.
- Thorsten Joachims, Thomas Finley, and Chun-Nam Yu. Cutting-plane training of structural svms. *Machine Learning*, 77:27–59, 2009. ISSN 0885-6125. URL <http://dx.doi.org/10.1007/s10994-009-5108-8>. 10.1007/s10994-009-5108-8.
- Joanne Scheibman. Local patterns of subjectivity in person and verb type in. In Joan L. Bybee and Paul Hopper, editors, *Frequency and the Emergence of Linguistic Structure*, volume 45 of *Typological studies in language*, pages 61–89. John Benjamins Publishing Company, Amsterdam; Philadelphia, 2001. ISBN 9789027229489.
- Skelton John. How to tell the truth in the british medical journal: Patterns of judgement in the 19th and 20th centuries. In *Hedging and Discourse*, volume Volume 24 of *Research in Text Theory*, pages 42–63. DE GRUYTER, March 1997. ISBN 978-3-11-015591-4. URL <http://dx.doi.org/10.1515/9783110807332.42.0>.
- John Lyons. *Semantics*, volume 2 of *Semantics*. Cambridge University Press, 1977. ISBN 9780521291866.
- E. Kärkkäinen. *Epistemic Stance in English Conversation: A Description of Its Interactional Functions, with a Focus on I Think*. Epistemic Stance in English Conversation: A Description of Its Interactional Functions, with a Focus on I Think. John Benjamins Publishing Company, 2003. ISBN 9781588114440.
- Elise Kärkkäinen. Position and scope of epistemic phrases in planned and unplanned american english. In *New approaches to hedging*, pages 207–241. Elsevier, Amsterdam, 2010.
- Paul Kay. Pragmatic aspects of grammatical constructions. In L.R. Horn and G.L. Ward, editors, *The handbook of pragmatics*, Blackwell handbooks in linguistics, pages 675–700. Blackwell Publishing, 2005. URL <http://www.icsi.berkeley.edu/~kay/cg.prag.pdf>.
- Graeme D. Kennedy. Quantification and the use of english: A case study of one aspect of the learner’s task. *Applied Linguistics*, 8(3):264–286, 1987a. doi: 10.1093/applin/8.3.264. URL <http://applij.oxfordjournals.org/content/8/3/264.abstract>.
- Graeme D. Kennedy. Expressing temporal frequency in academic english. *TESOL Quarterly*, 21(1):pp. 69–86, 1987b. ISSN 00398322. URL <http://www.jstor.org/stable/3586355>.
- Halil Kilicoglu and Sabine Bergler. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP ’08,

pages 46–53, Morristown, NJ, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-11-4. URL <http://www.aclweb.org/anthology/W/W08/W08-0607.pdf>.

Halil Kilicoglu and Sabine Bergler. A high-precision approach to detecting hedges and their scopes. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 70–77, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W10-3010>.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: <http://dx.doi.org/10.3115/1075096.1075150>. URL <http://dx.doi.org/10.3115/1075096.1075150>.

Manfred Kochen and Albert N. Badre. On the precision of adjectives which denote fuzzy sets. *Journal of Cybernetics*, 4(1):49–59, 1974. ISSN 0022-0280. URL <http://www.informaworld.com/10.1080/01969727408546055>.

Natalia Konstantinova and Sheila C. M. de Sousa. Annotating negation and speculation: Annotation guidelines. Guidelines, Research Group in Computational Linguistics, University of Wolverhampton, Stafford Street, Wolverhampton, WV1 1SB, UK, 2012. URL http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. A review corpus annotated for negation, speculation and their scope. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Angelika Kratzer. The notional category of modality. In Paul Portner and Barbara H. Partee, editors, *Formal Semantics*, pages 289–323. Blackwell Publishers Ltd, 2008. ISBN 9780470758335. doi: [10.1002/9780470758335.ch12](https://doi.org/10.1002/9780470758335.ch12). URL <http://dx.doi.org/10.1002/9780470758335.ch12>.

Angelika Kratzer. *Modals and Conditionals: New and Revised Perspectives*. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford, 2012. ISBN 9780199234684. URL <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199234684.001.0001/acprof-9780199234684>.

- Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*, pages 633–642, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124371. URL <http://doi.acm.org/10.1145/2124295.2124371>.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL <http://portal.acm.org/citation.cfm?id=645530.655813>.
- George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2(4):458–508, 1973. doi: 10.1007/BF00262952. URL <http://dx.doi.org/10.1007/BF00262952>.
- George Lakoff and Mark Johnson. *Metaphors we Live by*. University of Chicago Press, Chicago, 1980. ISBN 978-0-226-46800-6.
- Daniel Lassiter. Gradable epistemic modals, probability, and scale structure. In *Proceedings of SALT*, volume 20, pages 197–215, 2011. URL <http://elanguage.net/journals/salt/article/view/20.197/1280>.
- M. Light, X. T. Qui, and P. Srinivasan. The language of bioscience: Facts, speculations, and statements in between. *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, pages 17 – 24, 2004. URL <http://pages.cs.brandeis.edu/~jamesp/biolink2004/papers/pdf/BIO003.pdf>.
- Liliana Mamani Sanchez and Carl Vogel. Imho: An exploratory study of hedging in web forums. In *Proceedings of the SIGDIAL 2013 Conference*, pages 309–313, Metz, France, August 2013. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W13-4046>.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330, 1993.
- Alexa T. McCray, Suresh Srinivasan, and Allen C. Browne. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care (SCAMC)*, pages 235–239, Washington (US), 1994.
- Ben Medlock and Ted Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of*

- Computational Linguistics*, pages 992–999, Prague, Czech Republic, 2007. URL <http://www.aclweb.org/anthology/P07-1125>.
- Amita Misra and Marilyn Walker. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France, August 2013. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W/W13/W13-4006>.
- Roser Morante and Walter Daelemans. Learning the scope of hedge cues in biomedical texts. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 28–36, Morristown, NJ, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-30-5.
- Roser Morante and Caroline Sporleder. Modality and Negation: An Introduction to the Special Issue. *Computational Linguistics*, 38(2):223–260, February 2012. ISSN 0891-2017. doi: 10.1162/COLI.a_00095. URL http://dx.doi.org/10.1162/COLI_a_00095.
- Kevin Kyung Nam, Mark S. Ackerman, and Lada A. Adamic. Questions in, knowledge in?: A study of naver’s question answering community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 779–788, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518821. URL <http://doi.acm.org/10.1145/1518701.1518821>.
- Shigeko Nariyama. Subject ellipsis in english. *Journal of Pragmatics*, 36(2):237–264, February 2004. ISSN 03782166. doi: 10.1016/S0378-2166(03)00099-7. URL <http://linkinghub.elsevier.com/retrieve/pii/S0378216603000997>.
- J. Nuyts. *Epistemic Modality, Language, and Conceptualization: A Cognitive-pragmatic Perspective*. Human cognitive processing. J. Benjamins, 2001. ISBN 9789027223579.
- Oxford English Dictionary. “weasel, n.”, April 2000. URL <http://dictionary.oed.com/cgi/entry/50282064>.
- Arzucan Özgür and Dragomir R. Radev. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 3 of *EMNLP '09*, pages 1398–1407, Morristown, NJ, USA, 2009. Association for Computational Linguistics. URL <http://portal.acm.org/citation.cfm?id=1699648.1699686>.
- Sotirios Paroutis and Alya Al Saleh. Determinants of Knowledge Sharing Using Web 2.0 Technologies. *Journal of Knowledge Management*, 13(4):52–63, 2009.
- Paul Portner. Imperatives and modals. *Natural Language Semantics*, 15(4):351–383, 2007. URL <http://dx.doi.org/10.1007/s11050-007-9022-y>.

- Paul Portner. *Modality*. Oxford Surveys in Semantics and Pragmatics. Oxford University Press, USA, New York, 2009. ISBN 9780199292431.
- Hong Liang Qiao and Renje Huang. Design and implementation of agts probabilistic tagger. *ICAME Journal*, 22:23–48, 1998. URL <http://helmer.aksis.uib.no/icame/ij22/hong.pdf>.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Longman, London, 1985.
- Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=1628960.1628969>.
- David Rosenwasser and Jill Stephen. *Writing Analytically*. Cengage Learning, sixth edition, 2011. ISBN 9781133715085.
- V. Rubin, E. Liddy, and N. Kando. Certainty identification in texts: Categorization model and manual tagging results. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text*. Springer, 2005a. URL http://publish.uwo.ca/~vrubin/Publications/RubinLiddyKando_CertaintyIdentification_Ch7.pdf.
- Victoria Rubin, E. Liddy, and N. Kando. Certainty identification in texts: Categorization model and manual tagging results. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text*. Springer, 2005b. URL http://publish.uwo.ca/~vrubin/Publications/RubinLiddyKando_CertaintyIdentification_Ch7.pdf.
- Victoria L. Rubin. *Identifying Certainty in Texts*. PhD thesis, Syracuse University, Syracuse, NY, 2006.
- Victoria L. Rubin. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46(5):533–540, 2010.
- Magdalena Schwager. Conditionalized imperatives. In Christopher Tancredi, Makoto Kanazawa, Ikumi Imani, and Kiyomi Kusumoto, editors, *Proceedings of SALT XVI*. Cornell University Linguistics Department: CLC Publications, 2007. URL <http://research.nii.ac.jp/salt16/proceedings/schwagerCORR.pdf>.
- Phillip Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, and Lotfi A. Zadeh. *Semantic Computing*. John Wiley & Sons, 2011. ISBN 9780470920879.

- Anne-Marie Simon-Vandenberg. I think and its dutch equivalents in parliamentary debates. *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, pages 297–317. Rodopi, 1998.
- Vibha Singhal Sinha, Senthil Mani, and Monika Gupta. Exploring activeness of users in qa forums. In *Proceedings of the 10th Working Conference on Mining Software Repositories, MSR '13*, pages 77–80, Piscataway, NJ, USA, 2013. IEEE Press. ISBN 978-1-4673-2936-1. URL <http://dl.acm.org/citation.cfm?id=2487085.2487104>.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April 2012. Association for Computational Linguistics.
- Chengjie Sun, Yi Guan, Xiaolong Wang, and Lei Lin. Rich features based conditional random fields for biological named entities recognition. *Comput. Biol. Med.*, 37:1327–1333, September 2007. ISSN 0010-4825. doi: 10.1016/j.combiomed.2006.12.002. URL <http://portal.acm.org/citation.cfm?id=1276516.1276560>.
- György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio, 2008. URL <http://www.aclweb.org/anthology/W/W08/W08-0606>.
- Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 13–17, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W10-3002>.
- Sandra A. Thompson and Anthony Mulac. A quantitative perspective on the grammaticalization of epistemic parentheticals in english. In *Approaches to Grammaticalization*, pages 314–329. John Benjamins, 1991. ISBN 9789027228994.
- Tzong-han Tsai, Chia-wei Wu, and Wen-lian Hsu. Using maximum entropy to extract biomedical named entities without dictionaries. In *Proceedings of IJCNLP2005*, pages 270–275, 2005.
- Amy B.M. Tsui. The pragmatic functions of I don't know. *Text - Interdisciplinary Journal for the Study of Discourse*, 11:607, 2009. ISSN 16134117. doi: 10.1515/text.1.1991.11.4.607. 4.
- Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust part-of-speech tagger for biomedical

- text. In Panayiotis Bozanis and Elias N. Houstis, editors, *Advances in Informatics*, pages 382–392. Springer, 2005.
- Teppo Varttala. Remarks on the communicative functions of hedging in popular scientific and specialist research articles on medicine. *English for Specific Purposes*, 18(2):177 – 200, 1999. ISSN 0889-4906. doi: [http://dx.doi.org/10.1016/S0889-4906\(98\)00007-6](http://dx.doi.org/10.1016/S0889-4906(98)00007-6). URL <http://www.sciencedirect.com/science/article/pii/S0889490698000076>.
- Veronika Vincze. Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 383–391, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL <http://www.aclweb.org/anthology/I13-1044>.
- Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9, 2008. URL <http://www.biomedcentral.com/1471-2105/9/S11/S9>.
- Veronika Vincze, Gyorgy Szarvas, Gyorgy Mora, Tomoko Ohta, and Richard Farkas. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and genia event corpora. *Journal of Biomedical Semantics*, 2(Suppl 5):S8, 2011. URL <http://www.jbiomedsem.com/content/2/S5/S8>.
- Carl Vogel and Jerom Janssen. Emoticonsciousness. In Maria Marinaro Anna Esposito, Amir Hussain and Raffaele Martone, editors, *Multimodal Signals: Cognitive and Algorithmic Issues*, pages 271–287. Springer Berlin / Heidelberg, 2009. COST Action 2102 and euCognition International School Vietri sul Mare, Italy, April 21-26, 2008. Revised Selected and Invited Papers.
- Carl Vogel and Liliana Mamani Sanchez. Epistemic signals and emoticons affect kudos. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pages 517 –522, Dec. 2012. doi: 10.1109/CogInfoCom.2012.6422036. Best Paper Award.
- Quang H. Vuong. Likelihood ratio tests for selection and non-nested hypotheses. *Econometrica*, 57(2):397–333, 1989.
- Jun Wang, Arjen P. de Vries, and Marcel J. T. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06*, pages 501–508, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148257. URL <http://doi.acm.org/10.1145/1148170.1148257>.

- Don Watson. *Watson's dictionary of weasel words, contemporary clichés, Cant & management jargon / Don Watson*. Knopf, Milsons Point, N.S.W. :, 2004. ISBN 1740513215.
- Andrew Weir. Left-edge deletion in english and subject omission in diaries. *English Language and Linguistics*, 16:105–129, 3 2012. ISSN 1469-4379. doi: 10.1017/S136067431100030X. URL http://journals.cambridge.org/article_S136067431100030X.
- Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. A corpus study of evaluative and speculative language. In *Proceedings of the Second SIG-dial Workshop on Discourse and Dialogue - Volume 16*, pages 1–10, Morristown, NJ, USA, 2001. Association for Computational Linguistics. doi: <http://dx.doi.org/10.3115/1118078.1118104>. URL <http://dx.doi.org/10.3115/1118078.1118104>.
- Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language ANN. *Language Resources and Evaluation*, 39(2/3): 164–210, 2005. URL <http://www.cs.pitt.edu/~wiebe/pubs/papers/lre05withappendix.pdf>.
- A. Wierzbicka. *English: meaning and culture*. Oxford University Press, USA, 2006.
- Wikipedia. Wikipedia:writing better articles, 2014. URL http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles.
- Theresa Wilson and Janyce Wiebe. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 53–60, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=1608829.1608837>.
- T. Winograd. *Language as a Cognitive Process Vol. 1: Syntax*. Language as a Cognitive Process. Addison-Wesley, 1983. ISBN 9780201085716.
- Seth Yalcin. Probability operators. *Philosophy Compass*, 5(11):916–937, 2010. ISSN 1747-9991. doi: 10.1111/j.1747-9991.2010.00360.x. URL <http://dx.doi.org/10.1111/j.1747-9991.2010.00360.x>.
- Lofti A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965. URL <http://www-bisc.cs.berkeley.edu/Zadeh-1965.pdf>.
- Csaba Zainkó, Tamás Csapó, and Géza Németh. Special speech synthesis for social network websites. In *Text, Speech and Dialogue*, volume 6231 of *Lecture Notes in Computer Science*, pages 455–463–463. Springer Berlin / Heidelberg, 2010. URL http://dx.doi.org/10.1007/978-3-642-15760-8_58.

Appendix A

My relevant publications

- (A). (Submitted) Mamani Sanchez, L. and C. Vogel (2015). **A hedging annotation scheme focused on epistemic phrases for informal language**. *MOdels for Modality Annotation 2015 Workshop co-located with IWCS 2015*.
- (B). Mamani Sanchez, L. and C. Vogel (2013). **IMHO: An Exploratory Study of Hedging in Web Forums**. In: *Proceedings of the SIGDIAL 2013 Conference*. Metz, France: Association for Computational Linguistics.
- (C). Vogel, C. and L. Mamani Sanchez (2012). **Epistemic signals and emoticons affect kudos**. In: *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*. Best Paper Award, pp.517-522.
- (D). Mamani Sanchez, L. and C. Vogel (2012). **Emoticons Signal Expertise in Technical Web Forums**. In: *Proceedings of the 22nd Italian Workshop on Neural Networks, Smart Innovation, Systems and Technologies*. Springer Berlin Heidelberg, pp.415-425.
- (E). Mamani Sanchez, L., B. Li, and C. Vogel (2010). **Exploiting CCG structures with tree kernels for speculation detection**. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning Shared Task*. CoNLL'10: Shared Task. Uppsala, Sweden: Association for Computational Linguistics.

Appendix B

Forum information

Table B.1 – Number of posts and users for full and training datasets.

	Specific ranks	Number of posts		Number of users
		Full dataset	Full training dataset	
1	[Company_name] Employee	6.95% (16016)	6.95% (12012)	367
2	Volunteer	2.78% (6407)	2.78% (4805)	3
3	Moderator	1.17% (2690)	1.17% (2017)	10
4	Administrator	3.15% (7273)	3.15% (5454)	4
5	[Product_name] Fighter	17.58% (40542)	17.58% (30407)	9
6	Super Trojan Terminator	0% (0)	0% (0)	0
7	Trojan Terminator	2.33% (5366)	2.33% (4024)	1
8	Super Worm Eliminator	0% (0)	0% (0)	0
9	Worm Eliminator	0% (0)	0% (0)	0
10	Super RootKit Eradicator	0% (0)	0% (0)	0
11	Rootkit Eradicator	2.12% (4896)	2.12% (3672)	1
12	Super Bot Obliterator	1.77% (4090)	1.77% (3068)	1
13	Bot Obliterator	0% (0)	0% (0)	0
14	Super Virus Trouncer	0.81% (1877)	0.81% (1407)	1
15	Virus Trouncer	1.25% (2891)	1.25% (2168)	1
16	Super Phishing Phryer	3.82% (8818)	3.82% (6614)	7
17	Phishing Phryer	0.31% (709)	0.31% (531)	2
18	Super Spam Squasher	3.82% (8807)	3.82% (6606)	13
19	Spam Squasher	0.78% (1804)	0.78% (1353)	4
20	Super Keylogger Crusher	0.78% (1801)	0.78% (1352)	7
21	Keylogger Crusher	0.75% (1720)	0.75% (1289)	9
22	Super Spyware Scolder	1.27% (2929)	1.27% (2197)	16
23	Spyware Scolder	0.95% (2195)	0.95% (1646)	12
24	Super Contributor	0.97% (2241)	0.97% (1679)	21
25	Regular Contributor	9.52% (21948)	9.52% (16463)	204
26	Contributor	15.17% (34974)	15.17% (26232)	1976
27	Regular Visitor	1.8% (4147)	1.79% (3103)	784
28	Visitor	12.53% (28899)	12.53% (21673)	7909
29	Newbie	7.6% (17530)	7.6% (13148)	8588

Table C.1 – Lakoff’s hedges list [Lakoff, 1973].

sort of	in a real sense
kind of	in an important sense
loosely speaking	in a way
more or less	mutatis mutandis
on the ____ side (tall, fat, etc.)	in a manner of speaking
roughly	details aside
pretty (much)	so to say
relatively	a veritable
somewhat	a true
rather	a real
mostly	a regular
technically	virtually
strictly speaking	all but technically
essentially	practically
in essence	all but a
basically	anything but a
principally	a self-styled
particularly	nominally
par excellence	he calls himself a ...
largely	in name only
for the most part	actually
very	really
especially	(he) as much as ...
exceptionally	-like
quintessential(ly)	-ish
literally	can be looked upon as
often	can be viewed as
more of a ____ than anything else	pseudo-
almost	crypto-
typically/ typical	(he’s) another (Caruso/ Lincoln/ Babe Ruth ...)
as it were	____ is the ____ of ____
in a sense	(e.g., America is the Roman Empire of the modern world. Chomsky is the DeGaulle of Linguistics, etc.)
nearly	
in one sense	

Table C.2 – Keywords expressing absolute certainty proposed by Rubin.

all	always	are doomed to	condemn
demand	deny	ever	everything
extraordinary	immediate	impossible	is doomed to
must	never	no	no one
nobody	none	none of	not a single
not since	nothing	obligated to	obligation
obsession	only	over somebody’s dead body	overwhelming
refuse	the	triumph	unprecedented
urgent			

Table C.3 – Keywords expressing high certainty proposed by Rubin.

a lot of	almost	apparent	apparently
are committed to	are considered	are due	are scheduled
are to	as many	as much as	be considered
become	begin	believe	big
can't	cannot	certain	certainly
clear	clearly	commitment	conclude
conclusion	confirm	continue	convinced
could not	crucial	decide	decision
decisively	demonstrate	do	evidence
evident	extremely	firmly	fundamental
good	great	hard	have to
high	hundreds of	important	in fact
indeed	is committed to	is considered	is due
is scheduled	is to	it is the first time	it is the second time
it is who	it was who	knew	know
known	likely	long	lot of
major	many of	millions of	more than
most	most of	much	nearly
necessarily	necessary	need	not only
not to say that	not until	not with	obviously
often	once again	one of	ought to
particularly	powerful	probably	profound
real	really	really a part of	remain
remarkably	repeatedly	require	routinely
serious	seriously	should	should
should not	show	significant	significantly
some of	staggering	strong	strongly
substantial	surely	the most	thousands of
too	totally	truly	very
very much	want	was committed to	were committed to

Appendix D

Lexical items found in this research

D.1 Single hedges

Table D.1 – Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
1	would	would	441	491
		'd	48	
		world	1	
		wuuld	1	
2	try	try	175	450
		tried	147	
		trying	111	
		tries	17	
3	some	some	396	396
4	other	other	305	357
		others	52	
5	may	may	155	319
		maybe	93	
		may be	71	
6	can	can	226	226
7	seem	seems	113	196
		seem	57	
		seemed	23	
		seemed like	1	
		seem like	1	
		seems like	1	
8	could	could	169	169

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
9	something	something	144	162
		something like	9	
		something else	7	
		somethink	1	
		somethinge	1	
10	might	might	133	133
11	question	question	75	125
		questions	50	
12	many	many	101	101
13	a few	a few	98	99
		a few others	1	
14	several	several	98	98
15	about	about	88	88
16	should	should	83	83
17	suggestion	suggestions	48	80
		suggestion	30	
		sergestion	1	
		sergestions	1	
18	probably	probably	80	80
19	appear	appears	51	79
		appear	19	
		appeared	9	
20	most	most	45	64
		most of	19	
21	attempt	attempt	27	63
		attempted	15	
		attempting	11	
		attempts	8	
		attempting	1	
		attempteing	1	
22	sometimes	sometimes	48	61
		sometime	10	
		some times	3	
23	suggest	suggested	49	60
		suggests	5	
		suggest	3	

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
		suggesting	3	
24	perhaps	perhaps	57	57
25	someone	someone	53	53
26	similar	similar	46	48
		similiar	2	
27	possible	possible	45	45
28	another	another	43	43
29	like	like	40	40
30	a bit	a bit	35	39
		a bit of	4	
31	look like	looks like	26	35
		look like	6	
		looked like	3	
32	one of	one of	34	35
		one of those	1	
33	a lot	a lot of	25	34
		a lot	9	
34	think	think	16	33
		thought	13	
		thinks	4	
35	a little	a little	32	32
36	anyone	anyone	29	29
37	hopefully	hopefully	29	29
38	certain	certain	27	28
		cerain	1	
39	almost	almost	27	27
40	strange	strange	24	27
		very strange	3	
41	likely	likely	27	27
42	a couple	a couple of	24	26
		couple of	1	
		a couple	1	
43	apparently	apparently	25	26
		apprently	1	

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
44	sound	sounds like	14	24
		sounds	7	
		sound exactly like	1	
		sounded like	1	
		sound like	1	
45	may not	may not	23	24
		may never	1	
46	around	around	24	24
47	a while	a while	19	19
48	must	must	19	19
49	unknown	unknown	18	18
50	few	few	17	17
51	claim	claimed	9	16
		claims	5	
		claim	2	
52	various	various	16	16
53	chance	chance	13	16
		chances	3	
54	based on	based on	15	15
55	temporarily	temporarily	11	14
		temporary	3	
56	sort of	sort of	12	13
		sorta	1	
57	potential	potential	13	13
58	suspect	suspected	9	13
		suspect	4	
59	confusing	confusing	12	12
60	possibly	possibly	12	12
61	not always	not always	7	12
		n't always	5	
62	somehow	somehow	12	12
63	confused	confused	11	11
64	possibility	possibility	10	11

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
		possibilities	1	
65	somewhere	somewhere	11	11
66	plan	plan planning	8 2	10
67	odd	odd	10	10
68	unlikely	unlikely	10	10
69	random	random	10	10
70	confusion	confusion confusion	9 1	10
71	typically	typically	10	10
72	curious	curious	10	10
73	multiple	multiple	9	9
74	might not	might not	8	8
75	somebody	somebody	8	8
76	effort	efforts effort	4 4	8
77	a long	a long	8	8
78	suppose	supposed suppose	6 2	8
79	potentially	potentially	7	7
80	someone else	someone else	7	7
81	generally	generally	7	7
82	intend	intend intended	4 3	7
83	wonder	wonder wondering	5 2	7
84	amount of	amount of	7	7
85	would not	would n't would not	6 1	7
86	part	part part of	4 2	6
87	do not know	did n't know	6	6

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
88	somewhat	somewhat some what	5 1	6
89	mostly	mostly	6	6
90	occasionally	occasionally	6	6
91	numerous	numerous	6	6
92	supposedly	supposedly	5	5
93	assume	assuming assume	3 2	5
94	kind of	kind of kinda	3 2	5
95	normally	normally	5	5
96	not necessarily	not necessarily	4	4
97	often	often	4	4
98	tend	tend tends	3 1	4
99	not sure	not sure	4	4
100	randomly	randomly	4	4
101	strangely	strangely	4	4
102	test	test	4	4
103	slightly	slightly	4	4
104	in part	in part	3	3
105	proposed	proposed	3	3
106	seemingly	seemingly seemngly	2 1	3
107	approximately	approx . approx approximately	1 1 1	3
108	feel	feel	3	3
109	usually	usually	3	3
110	not appear	n't appear	3	3
111	weird	weird wierd	2 1	3

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
112	not many	not many	2	3
		not too many	1	
113	partly	partly	3	3
114	relatively	relatively	3	3
115	promised	promised	3	3
116	guess	guess	1	3
		guesses	1	
		guessing	1	
117	for the most part	for the most part	3	3
118	at times	at times	2	2
119	hope	hope	1	2
		hopes	1	
120	speculation	speculation	2	2
121	largely	largely	2	2
122	obscure	obscure	2	2
123	shortly	shortly	2	2
124	optional	optional	2	2
125	occasional	occasional	2	2
126	clueless	clueless	2	2
127	promises	promises	2	2
128	beginning	beginning	2	2
129	doubt	doubt	2	2
130	alleged	alleged	2	2
131	suspicious	suspicious	2	2
132	not knowing	not knowing	2	2
133	every other	every other	2	2
134	questionable	questionable	2	2
135	alleviate	alleviate	2	2
136	unsure	unsure	1	1
137	technically	technically	1	1
138	surprisingly	surprisingly	1	1

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
139	vague	very vague	1	1
140	secretly	secretly	1	1
141	confusingly	confusingly	1	1
142	liable	liable	1	1
143	virtually	virtually	1	1
144	figure	figured	1	1
145	reservations	reservations	1	1
146	over	over	1	1
147	mysteriously	mysteriously	1	1
148	misunderstanding	misunderstanding	1	1
149	unbeknownst	unbenknownst	1	1
150	rare	rare	1	1
151	unexplained	unexplained	1	1
152	inconclusive	inconclusive	1	1
153	practically	practically	1	1
154	according to	according to	1	1
155	unclear	unclear	1	1
156	for ages	for ages	1	1
157	half	half	1	1
158	lot	lots	1	1
159	confuse	confuses	1	1
160	assumption	assumptions	1	1
161	estimate	estimate	1	1
162	uncategorized	uncategorized	1	1
163	for times	for times	1	1
164	unaware	unaware	1	1
165	too much	too much	1	1
166	coming	coming	1	1
167	hidden	hidden	1	1
168	larger	larger	1	1

Continued on Next Page...

Table D.1 – Continued: Raw frequencies for Single-hedge types and subtotal raw frequencies for their normalised types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
169	whatever	whatever	1	1
170	do not appear	do n't appear	1	1
171	puzzled	puzzled	1	1
172	curiosity	curiosity	1	1
173	perceptible	perceptible	1	1
174	no specific	no specific	1	1
175	putative	putative	1	1
176	personally	personally	1	1
177	intention	intentions	1	1
178	whoever	whoever	1	1
179	slight	slight	1	1
180	thousands of	thousands of	1	1
181	variations	variations	1	1
182	probable	probable	1	1
183	not clear	not clear	1	1
184	elsewhere	elsewhere	1	1
185	everything	everything	1	1
186	not permanent	not permanent	1	1
187	misunderstand	misunderstand	1	1
188	resembling	resembling	1	1
189	conflicting	conflicting	1	1

D.2 NCK hedges

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
1	i think	i think	157	210
		i thought	35	
		i 'm thinking	6	
		i * think	5	
		i thing	1	

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

Normalized keyword	Original keyword	Freq.	Subtotal
	i was thinking	1	
	i still think	1	
	i am thinking	1	
	i now think	1	
	think	1	
	i thinh	1	
2	i hope	59	125
	hope	49	
	i was hoping	4	
	i 'm hoping	3	
	i sure hope	2	
	i do hope	2	
	i just hope	2	
	i hoped	1	
	i am hoping	1	
	i had hoped	1	
	i am hopeful	1	
3	i do not know	43	89
	i do n't know	8	
	i dont know	8	
	do n't know	6	
	i did n't know	5	
	did n't know	3	
	i really do n't know	2	
	do not know	2	
	i idd not konw	1	
	dont know	1	
	i dont even know	1	
	dunno	1	
	i do n't know for sure	1	
	i did not know	1	
	i do n't have a way of knowing	1	
	i know nothing	1	
	i really do not know	1	
	donno	1	
	i do notknow	1	
	i * do n't know	1	
4	i am not sure	30	75
	i 'm not sure	19	
	not sure	16	

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
		i am just not sure	2	
		im not sure	2	
		never been sure	1	
		i was not sure	1	
		i 'm not 100 % sure	1	
		i * am not sure	1	
		i am not quite sure	1	
		i 'm not quite sure	1	
5	i believe	i believe	46	48
		i beleive	1	
		i believed	1	
6	i wonder	i wonder	12	43
		i was wondering	6	
		wonder	4	
		i am wondering	4	
		i 'm wondering	3	
		i wondered	2	
		i 'm just wondering	2	
		i did wonder	1	
		i * wonder	1	
		i am starting to wonder	1	
		wondering	1	
		i 'm really wondering	1	
		i also wonder	1	
		wonders	1	
		i 'm still wondering	1	
		i am just a tad wondering	1	
		i * am wondering	1	
7	i guess	i guess	30	36
		guess	2	
		i 'm guessing	2	
		i am just making educated guesses	1	
		i guest	1	
8	i do not think	i do n't think	21	29
		i dont think	3	
		i do not think	2	
		i did not think	1	
		i also do n't think	1	
		i don think	1	

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
9	i assume	i assume	11	21
		i 'm assuming	6	
		i assumed	3	
		i am assuming	1	
10	i do not understand	i do n't understand	5	20
		i can not understand	3	
		i ca n't understand	2	
		i dont understand	1	
		i do n't really understand	1	
		i do not understand	1	
		do not undrstand	1	
		still ca n't understand	1	
		i personally do n't understand	1	
		i still do n't understand	1	
		dont understand	1	
		i * do not truly understand	1	
		i still dont understant	1	
		11	i suspect	
i suspected	1			
i 'm suspicious	1			
i now suspect	1			
12	i have no idea	i have no idea	10	14
		no idea	2	
		i had no idea	1	
		i have not the faintest idea	1	
13	i suggest	i suggest	8	13
		i also suggest	2	
		i respectfully suggest	1	
		i suggested	1	
		i do suggest	1	
14	i am confused	i 'm a bit confused	3	13
		i am confused	1	
		i 'm a little confused	1	
		i am so confused	1	
		i am a little confused	1	
		i am a bit confused	1	
		i am totally confused	1	
		i was totally confused	1	
		i 'm confused	1	

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
		i am also somewhat confused	1	
		i am really confused	1	
15	imo	in my opinion	7	12
		imo	5	
16	i would suggest	i would suggest	10	11
		i 'd suggest	1	
17	i doubt	i doubt	7	10
		i honestly doubt	1	
		have doubts * i	1	
		i sincerely doubt	1	
18	i feel	i feel	10	10
19	it seems to me	seems to me	7	10
		seemed to me	2	
		seemed first to me	1	
20	i suppose	i suppose	7	8
		i spose	1	
21	afaik	afaik	4	8
		as far as i know	3	
		as far as i knew	1	
22	i need to know	i need to know	3	6
		i really need to know	2	
		i * need to know	1	
23	i do not recall	i ca n't recall	3	6
		i do n't recall	1	
		i dont recall	1	
		dont recall	1	
24	i do not remember	i do n't remember	1	6
		i really do n't remember	1	
		i do not remember	1	
		do not remember	1	
		i dont remember	1	
		i ca n't remember	1	
25	i would like to know	i would like to know	6	6
26	i expect	i expect	3	6
		i expected	2	
		i 'm expecting	1	

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
27	i am unsure	i 'm unsure i am unsure i * was unsure am unsure	2 1 1 1	5
28	we hope	we hope	5	5
29	to my knowledge	to my knowledge to the best of my knowledge	3 2	5
30	i am not familiar	am not familiar i 'm not familiar i am not totaly familiar i 'm not exactly familiar i am not familiar	1 1 1 1 1	5
31	i would think	i would think	5	5
32	i can not seem	can't seem i can not seem i * can not seem i cant seem	1 1 1 1	4
33	my guess	my guess my guess is	2 2	4
34	i can not say	i can not say for sure i can not say i * can not say for sure	2 1 1	4
35	i hesitate	i am hesitant am hesitant i hesitate	2 1 1	4
36	i gather	i gather	4	4
37	i am not an expert	i am not an expert i 'm not an expert i am far from an expert	2 1 1	4
38	i presume	i presume	4	4
39	we will investigate	we will investigate	3	3
40	we do not know	we need to know we dont know	2 1	3
41	i can not figure out	i ca n't figure out i can not figure it out cant figure out	1 1 1	3

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
42	i would assume	i would assume	3	3
43	i am considering	i am considering	3	3
44	i would not think	i would not think i would n't think	2 1	3
45	we think	we think we thought	1 1	2
46	i have missed	i 've missed	2	2
47	imho	imho	2	2
48	i am unclear	i 'm unclear i am unclear	1 1	2
49	i misunderstand	i completely misunderstood i am misunderstanding	1 1	2
50	i am not certain	i 'm not certain i am not certain	1 1	2
51	i missed something	i missed something	2	2
52	o hope	i really hope	2	2
53	i wish i knew	i wish i knew	2	2
54	as far as i can tell	as as far as i can tell as far as i can tell	1 1	2
55	i can not determine	i cant determine i can not determine	1 1	2
56	i would guess	i would guess	2	2
57	i have no clue	i have no clue have no clue * i	1 1	2
58	i have a question	i have a question	2	2
59	i would expect	i would expect	2	2
60	it looks to me	it looks to me like it looks to me	1 1	2
61	i would hope	i would hope	2	2
62	that is just my opinion	that's just my opinion	1	1
63	one would think	one would think	1	1
64	it would look to me like	it would look to me like	1	1

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
65	we are of the belief	[username] and i both are of the belief	1	1
66	we do not have the full knowledge	we do not have the full knowledge	1	1
67	as far as i saw	as far as i saw	1	1
68	i am lost	i am totally lost	1	1
69	we only know	we only know	1	1
70	i am almost certain	i 'm almost certain	1	1
71	i have run out of ideas	i have run out of ideas	1	1
72	i still have not figured out	i still have n't figured out	1	1
73	my feeling is	my feeling is	1	1
74	so far as i know	so far as i know	1	1
75	i reckon	i reckon	1	1
76	it is my hope	it is my hope	1	1
77	i am unaware	i am unaware	1	1
78	it is my understanding	it 's my understanding	1	1
79	i have no experience	i have no experience	1	1
80	i can not pin it down	i ca n't pin it down	1	1
81	it is not clear to me	it's not clear to me	1	1
82	we are not sure	we ca n't be sure	1	1
83	one would hope	one would hope	1	1
84	we believe	we believe	1	1
85	i can not decide	i can not decide	1	1
86	i can not vouch for	i ca n't vouch for	1	1
87	i am out of ideas	i am out of ideas	1	1
88	it is incomprehensible to me	is incomprehensible to me	1	1
89	my guess would be	my guess would be	1	1
90	i am not clear	i am not clear	1	1
91	i never knew	i never knew	1	1
92	my point of view	my point of view	1	1
93	i imagine	i imagine	1	1

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
94	we gather	we gather	1	1
95	i can not seem to know	i ca n't seem to know	1	1
96	i would presume	i would presume	1	1
97	i seem	i seam	1	1
98	it sounds to me	it sounds to me	1	1
99	i forgot	forgot	1	1
100	i infer	i infer	1	1
101	i am planning	i 'm planning	1	1
102	we feel	we feel	1	1
103	i am intrigued	i am intrigued	1	1
104	i have not thought	i had n't thought	1	1
105	i am inclined to believe	i am inclined to believe	1	1
106	not that i know	not that i know	1	1
107	my doubt is	my doubt was	1	1
108	as far as i can see	as far as i can see	1	1
109	i am uncertain	i 'm uncertain	1	1
110	i am not confident	i am not confident	1	1
111	my assumption is	my assumption is	1	1
112	we have yet to confirm	we 've yet to confirm	1	1
113	i would like to suggest	i would like to suggest	1	1
114	i am enquiring	i 'm enquiring	1	1
115	i mistakenly thought	i mistakenly thought	1	1
116	i have tried	i 've tried	1	1
117	i have no explanation	i have no explanation	1	1
118	my understanding is	my understanding in reading this is	1	1
119	i can not explain	i can not explain	1	1
120	it appears to me	appears to me	1	1
121	i am wrong	i am wrong	1	1
122	i do not have an answer	i do n't have an answer	1	1
123	i am hesitant	i am a little hesitant	1	1

Continued on Next Page...

Table D.2 – Raw frequencies for Non-claiming-Knowledge phrasal types in the annotation dataset.

	Normalized keyword	Original keyword	Freq.	Subtotal
124	my suspicion is	my suspicion is	1	1
125	i can not help wondering	i ca n't help wondering	1	1
126	i would estimate	i would estimate	1	1
127	i do not see	i really do n't see	1	1
128	i am under the impression	i was under the impression	1	1
129	i would be interested to know	i 'd be interested to know	1	1
130	i do not realize	i did n't realize	1	1
131	in my view	in my view	1	1
132	my suggestion is	my suggestion is	1	1
133	i am skeptical	i was slightly skeptical	1	1
134	my impression is	my impression is	1	1
135	i am clueless	i was clueless	1	1
136	i take it	i take it	1	1
137	we would like to know	we would like to know	1	1
138	i 'm still learning	i 'm still learning	1	1

D.3 Other hedges

Table D.3 – Raw frequencies for hedge types in the *Other* category found in the annotation dataset.

N	Tokenized expression	Freq.	Subcat. label
1	so far	30	other
2	or so	22	other
3	or something	9	other
4	to be sure	7	other
5	more or less	6	other
6	or whatever	6	other
7	as * as possible	5	other
8	fingers crossed	5	other
9	at your own risk	4	other
10	it would appear	3	nck.like
11	you can find out	3	other
12	i do n't get	3	nck.like

Continued on Next Page...

Table D.3 – Continued: Raw frequencies for hedge types in the *Other* category found in the annotation dataset.

N	Tokenized expression	Freq.	Subcat. label
13	from time to time	2	other
14	i am looking for a way	2	nck.like
15	in theory	2	other
16	i want to make sure	2	nck.like
17	just a thought	2	other
18	my question is	2	nck.like
19	im new	2	nck.like
20	there are times	2	other
21	10-15	2	other
22	do not know	2	nck.like
23	does not know	1	other
24	i have never done this kind of thing before	1	nck.like
25	being a newbie	1	nck.like
26	i know i sound like	1	nck.like
27	i have n't found an answer	1	nck.like
28	does n't lend itself to a quick or easy diagnosis	1	other
29	you do not know	1	other
30	can not trust it 100 % of the time	1	other
31	i do n't have a specific eta	1	nck.like
32	i need to research	1	nck.like
33	does not provide any references	1	other
34	n't * that i 'm aware of	1	nck.like
35	they do n't understand	1	other
36	without knowing for sure	1	other
37	does not increase confidence	1	other
38	nobody knows	1	other
39	their engineers are researching	1	other
40	we are working with the powerdesk engineers to figure out	1	nck.like
41	ca n't believe	1	nck.like
42	it is only a guess	1	nck.like
43	its hard to know	1	nck.like
44	it is hard to tell	1	nck.like
45	no longer has trustworthy	1	other
46	so far today	1	other
47	keeping fingers crossed	1	other
48	you did not make clear	1	other
49	by this reasoning	1	other
50	of the like	1	other
51	i finally did a google search	1	nck.like
52	i am not a techie	1	nck.like

Continued on Next Page...

Table D.3 – Continued: Raw frequencies for hedge types in the *Other* category found in the annotation dataset.

N	Tokenized expression	Freq.	Subcat. label
53	i need guidance	1	nck.like
54	we 're investigating	1	nck.like
55	we are stuck on how to	1	nck.like
56	i am not that brilliant in technical side	1	nck.like
57	i have n't been able to find any solution	1	nck.like
58	keep my fingers crossed	1	other
59	you do n't say	1	other
60	does n't give me a lot of confidence	1	nck.like
61	i have n't tried this personally	1	nck.like
62	need to explain	1	other
63	nothing however is guaranteed	1	other
64	it's hard to say	1	nck.like
65	no other (known)	1	other
66	cross my fingers	1	other
67	you do n't know	1	other
68	(as a second opinion)	1	other
69	it does n't know	1	other
70	from what you have said	1	other
71	where i need to go	1	other
72	that i know of	1	nck.like
73	under investigation	1	other
74	i am new to ghost	1	nck.like
75	trying to ubderstand	1	nck.like
76	any info would be appreciated	1	other
77	confused the daylight's out of me	1	nck.like
78	not acquainted witht	1	other
79	we are currently in the process of implementing and testing	1	nck.like
80	but it is only as good as your habits	1	other
81	i 'm not that good with computers	1	nck.like
82	or something like that	1	other
83	please explain more	1	other
84	as soon as i can	1	other
85	just to clarify	1	other
86	just my feeling	1	nck.like
87	i am new here	1	nck.like
88	as far as the suggestions	1	other
89	i have some additional questions	1	nck.like
90	i 've already searched and ca n't find an answer	1	nck.like
91	from what limited knowledge i have	1	nck.like
93	from my limited knowledge	1	nck.like

Continued on Next Page...

Table D.3 – Continued: Raw frequencies for hedge types in the *Other* category found in the annotation dataset.

N	Tokenized expression	Freq.	Subcat. label
93	it will be nice to know whether it works	1	other
94	as far as you saying that	1	other
95	i was looking to find	1	nck.like
96	2-4	1	other
97	maybe or not	1	other
98	a * as possible	1	other
99	this is only a theory on my part	1	nck.like
100	my question at the moment is	1	nck.like
101	this is starting to	1	other
102	according to this website	1	other
103	i 'm not too techy	1	nck.like
104	my other thought was	1	nck.like
105	somebody has any news	1	other
106	all i am asking is	1	nck.like
107	no * is 100 %	1	other
108	x amount of	1	other
109	10%-30 %	1	other
110	i must admit i dont recognise it	1	nck.like
111	you can not say for sure	1	other
112	just wondering	1	nck.like
113	any help is appreciated	1	other
114	i have not tested ghost in this scenario	1	nck.like
115	more than likely	1	other
116	i am not a real techie	1	nck.like
117	15-20	1	other
118	once in a while	1	other
119	in the event of	1	other
120	not being able to figure this all out	1	nck.like
121	before i know for sure	1	nck.like
122	no one has an answer	1	other
123	upward of	1	other
124	i never received any explanation	1	nck.like
125	filled with hope * i was	1	nck.like
126	it is n't obvious	1	other
127	i am not a technician	1	nck.like
128	i remain very interested in trying to find out	1	nck.like
129	n't * that i am aware of	1	nck.like
130	god knows	1	other
131	may or may not	1	other
132	circa	1	other
133	i've googled relentlessly * without success	1	nck.like

Continued on Next Page...

Table D.3 – Continued: Raw frequencies for hedge types in the *Other* category found in the annotation dataset.

N	Tokenized expression	Freq.	Subcat. label
134	my lack of computer knowledge	1	nck.like
135	support people did not know	1	other
136	my windows knowledge is n't that great	1	nck.like
137	leaves me in the dark	1	nck.like
138	i still need help on	1	nck.like
139	need some additional insight from you	1	nck.like
140	no one is 100 % sure	1	other
141	some do not even know	1	other
142	are researching a solution	1	other
143	in all likelihood	1	other
144	a good portion	1	other
145	have not understood	1	other
146	i 'm a complete newbie	1	nck.like
147	the case could be made	1	other
148	i 'm curious	1	nck.like
149	question/problem	1	other
150	i really new	1	nck.like
151	i want to know	1	nck.like
152	one person	1	other
153	i do n't see	1	nck.like
154	i did not realise	1	nck.like
155	goes into investigation	1	other
156	i still have not heard from anyone	1	nck.like
157	not computer-savy enough to know	1	other
158	to get an understanding	1	other
159	still working on this to figure it out	1	nck.like
160	i have to hope for the best	1	nck.like
161	i 'm a novice	1	nck.like
162	at some point	1	other
163	i am a rookie	1	nck.like
164	floating around in my memory bank	1	other
165	as far as i can	1	other
166	did n't even know	1	other
167	i 'm a computer novice	1	nck.like
168	ca n't figure it out	1	nck.like
169	point of view	1	other
170	to be mistaken	1	other
171	makes no warranty that	1	other
172	there is no provision for	1	other
173	cross your fingers !	1	other
174	close to	1	other

Continued on Next Page...

Table D.3 – Continued: Raw frequencies for hedge types in the *Other* category found in the annotation dataset.

N	Tokenized expression	Freq.	Subcat. label
175	i 'm not really techie enough any more	1	nck.like
176	we are currently investigating	1	nck.like
177	i googled	1	nck.like
178	they are n't aware of	1	other
179	he wo n't be so sure	1	other
180	2-5	1	other
181	i am not very computer literate	1	nck.like
182	wish me luck	1	other
183	not realizing	1	other
184	is not sure	1	other
185	i 'm relatively computer-stupid	1	nck.like
186	not all users would know	1	other
187	not being able to figure out	1	other
188	i 'm lost	1	nck.like
189	70-80	1	other
190	for years	1	other
191	no one from [company_name] has been able to figure out	1	other
192	half of the way	1	other
193	i am technically challenged	1	nck.like
194	~ 73gb	1	other
195	20-30	1	other
196	i 'm new	1	other
197	i am a relative newbie	1	nck.like
198	i am investigating	1	nck.like
199	from a techie's point of view	1	other
200	from a normal user's point of view	1	other
201	i was confused and intimidated	1	nck.like
202	i am a neophyte when it comes to	1	nck.like
203	30 - 76	1	other
204	7-10	1	other
205	75-100	1	other
206	10 - 0	1	other
207	midnight-1am	1	other
208	not 100 % sure	1	other
209	no one knows	1	other

Appendix E

Data Preparation

E.1 Substitution by wildcards

Wildchar	Explanation	Example
UUUUUU	Username	
QQQQQQ	Quotations content would make user's discourse spurious as it contains another user's discourse ([username] wrote:). Quotations are enclosed in BLOCKQUOTE tags.	<code><BLOCKQUOTE><HR/>UUUUUU wrote:
No, I have never upgraded beyond 2007.<HR /></BLOCKQUOTE></code>
LLLLLL	The content of link tags are is relevant as linguistic feature for discourse analysis, the sole indication of link presence enough.	<code>http://www.dslreports.com/dummyspage<A></code>
HHHHHH	HKEYS from Windows	
IIIIII	Images	<code></code>
TTTTTT	Timestamp	
SSSSSS	Smilie	
OOOOOO	Emoticon	

Table E.1 – Wildchars used for text standardization in web forum posts.

It happens that some posts are edited by administrators to moderate the interaction in the forum. Textually, this fact is evidenced by some notes:

[Edit: a line has been snipped from this post because it references a post that has been removed. Additionally, The context of the post below may seem a little bit out of place without that message.]

Edit: Clarifying the thread subject

These notes were removed from the post files.

E.2 Skipping files

Some files types may be skipped as they are not necessary, are redundant or their disclosure is restricted to ensure the protection of users and company confidential information.

E.3 Anonymisation

This procedure was done partially in an automatic manner. Manual intervention was needed on editing the list of user names as many of them would case a mismatch of another types of entities that did not correspond to user names.

Appendix F

Description of statistical models

F.1 Confidence intervals

Table F.1 – Confidence intervals for probabilities in pairwise comparisons for model AllCats1 at 95% of confidence.

	Estimate	lwr	upr
unhedged - SingSyntOth	0.32481	0.27858	0.37473
unhedged - SingSynt	0.33176	0.30992	0.35436
unhedged - SingSyntNCKOth	0.34176	0.29352	0.39350
unhedged - SingSyntNCK	0.35007	0.32488	0.37611
SyntNCK - SingSyntOth	0.40489	0.31625	0.50020
Synt - SingSyntOth	0.40761	0.35079	0.46700
unhedged - Synt	0.41147	0.37335	0.45069
SyntNCK - SingSynt	0.41251	0.33486	0.49477
unhedged - SyntNCK	0.41420	0.33462	0.49853
Synt - SingSynt	0.41524	0.38122	0.45009
Synt - SingSyntNCKOth	0.42615	0.36770	0.48673
unhedged - Sing	0.43211	0.40537	0.45924
unhedged - SingNCK	0.43434	0.39487	0.47467
Synt - SingSyntNCK	0.43515	0.39846	0.47257
SingSyntNCK - SingSynt	0.47964	0.45841	0.50095
Sing - NCK	0.55397	0.48093	0.62475
SingSyntNCK - Sing	0.58552	0.56196	0.60870
SingSyntNCK - SingNCK	0.58773	0.55000	0.62446
SingSyntNCKOth - Sing	0.59441	0.54168	0.64504
SingSyntNCKOth - SingNCK	0.59660	0.53619	0.65422
SingSynt - Sing	0.60515	0.58551	0.62446
SingSynt - SingNCK	0.60732	0.57223	0.64134
SingSyntNCK - SingOth	0.60953	0.51124	0.69967
SingSyntOth - Sing	0.61266	0.56134	0.66158
SingSyntOth - SingNCK	0.61482	0.55579	0.67065
SingSyntNCKOth - SingOth	0.61824	0.50948	0.71631
SingSynt - SingOth	0.62874	0.53240	0.71583
SingSyntOth - SingOth	0.63607	0.52891	0.73125
SingSyntNCK - NCK	0.63696	0.56663	0.70189
SingSyntNCKOth - NCK	0.64541	0.56157	0.72118
SingSynt - NCK	0.65559	0.58761	0.71774
SingSyntOth - NCK	0.66267	0.58073	0.73588

Table F.2 – Confidence intervals for probabilities in pairwise comparisons for model AllbutSynt1 at 95% of confidence.

	Estimate	lwr	upr
NCKOth - NCK	0.35645	0.11110	0.71053
Oth - NCK	0.48367	0.36185	0.60747
Sing - NCK	0.58533	0.53613	0.63288
SingNCK - NCK	0.59093	0.54062	0.63941
SingNCKOth - NCK	0.58682	0.52406	0.64689
SingOth - NCK	0.57570	0.51551	0.63372
unhedged - NCK	0.47696	0.42563	0.52878
Oth - NCKOth	0.62843	0.26488	0.88812
Sing - NCKOth	0.71819	0.36805	0.91771
SingNCK - NCKOth	0.72285	0.37329	0.91949
SingNCKOth - NCKOth	0.71943	0.36754	0.91880
SingOth - NCKOth	0.71011	0.35742	0.91517
unhedged - NCKOth	0.62213	0.27319	0.87822
Sing - Oth	0.60109	0.48674	0.70539
SingNCK - Oth	0.60662	0.49197	0.71063
SingNCKOth - Oth	0.60257	0.48177	0.71204
SingOth - Oth	0.59157	0.47183	0.70135
unhedged - Oth	0.49328	0.37913	0.60814
SingNCK - Sing	0.50578	0.48913	0.52243
SingNCKOth - Sing	0.50154	0.46039	0.54267
SingOth - Sing	0.49011	0.45359	0.52674
unhedged - Sing	0.39248	0.37500	0.41024
SingNCKOth - SingNCK	0.49576	0.45313	0.53844
SingOth - SingNCK	0.48433	0.44617	0.52268
unhedged - SingNCK	0.38698	0.36662	0.40774
SingOth - SingNCKOth	0.48857	0.43513	0.54228
unhedged - SingNCKOth	0.39101	0.35045	0.43314
unhedged - SingOth	0.40195	0.36491	0.44015

Table F.3 – Confidence intervals for probabilities in pairwise comparisons for model SinglevsNCK at 95% of confidence.

	Estimate	lwr	upr
Sing - NCK	0.59255	0.55142	0.63242
SingNCK - NCK	0.59399	0.55198	0.63467
unhedged - NCK	0.48416	0.44072	0.52784
SingNCK - Sing	0.50149	0.48786	0.51512
unhedged - Sing	0.39225	0.37722	0.40748
unhedged - SingNCK	0.39082	0.37365	0.40827