Vasiliki Simaki*, Carita Paradis, Maria Skeppstedt,
Magnus Sahlgren, Kostiantyn Kucher and Andreas Kerren

Annotating Speaker Stance in Discourse: The Brexit Blog Corpus

https://doi.org/10.1515/cllt-2016-0060

Abstract: The aim of this study is to explore the possibility of identifying speaker stance in discourse, provide an analytical resource for it and an evaluation of the level of agreement across speakers. We also explore to what extent language users agree about what kind of stances are expressed in natural language use or whether their interpretations diverge. In order to perform this task, a comprehensive cognitive-functional framework of ten stance categories was developed based on previous work on speaker stance in the literature. A corpus of opinionated texts was compiled, the Brexit Blog Corpus (BBC). An analytical protocol and interface (Active Learning and Visual Analytics) for the annotations was set up and the data were independently annotated by two annotators. The annotation procedure, the annotation agreements and the co-occurrence of more than one stance in the utterances are described and discussed. The careful, analytical annotation process has returned satisfactory inter- and intra-annotation agreement scores, resulting in a gold standard corpus, the final version of the BBC.

Keywords: text annotation, blog post texts, modality, evaluation, positioning

1 Introduction

Communication between humans is never completely neutral in the sense that no particular perspective or selection of information is employed. Human interaction is

***Corresponding author: Vasiliki Simaki,** Centre for Languages and Literature, Lund University, Lund, Sweden; Derartment of Computer Science, Linnaeus University, Växjö, Sweden, E-mail: vasiliki.simaki@englund.lu.se
Carita Paradis, Centre for Languages and Literature, Lund University, Lund, Sweden, E-mail: carita.paradis@englund.lu.se
Maria Skeppstedt, Derartment of Computer Science, Linnaeus University, Växjö, Sweden; Gavagai AB, Stockholm, Sweden, E-mail: maria.skeppstedt@lnu.se
Magnus Sahlgren, Gavagai AB, Stockholm, Sweden, E-mail: mange@gavagai.se
Kostiantyn Kucher: E-mail: kostiantyn.kucher@lnu.se, **Andreas Kerren:** E-mail: andreas.kerren@lnu.se, Derartment of Computer Science, Linnaeus University, Växjö, Sweden

Open Access. © 2020 Simaki et al., published by De Gruyter. BY-NC-ND This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

always view-pointed in one way or another, and meanings of lexical items are dynamic and adaptable in relation to the situational and linguistic context where they are used (Langacker 1987; Talmy 2000; Croft and Cruse 2004; Gärdenfors 2014a, 2014b; Dancygier and Sweetser 2012; Paradis 2015). Speakers take stance when interacting with other people. They make assessments and they position themselves in relation to other interlocutors to mark their standpoint (Englebretson 2007).

Speaker contributions come with stances expressed in a range of different ways that may convey more than one stance in different contexts, and there may be more than one stance expression in the same sentence or utterance. There are cases of expressions of stance such as *might, definitely, I am sure that* that are treated as expressions of stance in the literature, but there are also types of constructions that might or might not be included in the category of speaker stance. For instance, *It's about time you sold your car* where the sequence starting *it's about time* + past tense of *sell* expresses a rather direct and tactless recommendation to somebody else about what the speaker thinks is beneficial for the addressee. Utterances of this kind are pervasive in language use and make the study of stance in real communication intriguing, but also methodologically challenging. A challenge that lies ahead of us concerns how we can go about identifying the meanings and the forms of stance in order to be able to train a computational model to automatically identify stance in discourse. For that purpose, we need a solid theoretical ground to start from, consisting of robust criteria for detection and annotation with acceptable inter-coder reliability scores.

A comprehensive cognitive-functional framework consisting of ten notional categories for annotating stance was set up, where the basic units of analysis are utterances. The term *utterance* in the present study is defined as the chunk between full stops. The utterances were holistically analyzed to determine whether they expressed speaker stance or not by two expert analysts, who also identified what type(s) of stance were expressed in each utterance. The main principle for assigning speaker stance to an utterance was that stance-taking should be identifiable through chunks of form-meaning pairings of varying size in language, which we refer to as constructions (Fillmore et al. 1988; Langacker 1987; Goldberg 1995, 2006; Croft 2001). This procedure contrasts with a great deal of work both in quantitative corpus linguistics (Biber 2006) and sentiment analysis in computational linguistics (Pang and Lee 2008), where the starting point is a preconceived list of words that the researcher assume generally express a given sentiment type. In contrast to these works, we offer an utterance-based approach to the analysis of speaker stance in communication based on the identification of constructions that actually express stance on the occasion of use. The prospective usefulness of this study, couched in a usage-based theoretical approach to meaning in language (Paradis 2015), is automatic identification beyond mere form, i.e. lists of words.

In order to come to grips with how speaker stance is expressed in language, and how different stance categories interact in discourse, ten core speaker stance categories have been defined on the basis of the rich literature on stance in language (for a comprehensive, annotated stance bibliography, see Glynn and Sjölin 2015). We compiled a corpus of social media text from blogs, the Brexit Blog Corpus (BBC), consisting of blog posts commenting on political issues related to the 2016 UK referendum. BBC was annotated by two annotators, and the annotation results are evaluated and discussed.

The paper is organized as follows. Section 2 presents an overview of the basic stance-taking concepts and studies in the field. In Section 3, the ten categories of stance are defined and exemplified. In Section 4, the BBC is introduced. It is described, and the annotation process and the results are presented. Section 5 evaluates the annotation results and discusses the cases of multiple stance occurrences in the corpus data. Finally, Section 6 is a summary of the findings and the implication of this study.

2 Background

In this section, an overview of different concepts that are used in the literature to talk about stance and studies from different academic traditions, theoretical as well as computational approaches, are considered and compared.

2.1 What is speaker stance?

Stance-taking in verbal communication among people can be described as the expression of the speaker's assessment of an object, an event or a proposition vis-á-vis his or her interlocutors. Stance is defined as the way speakers position themselves in relation to their own or other people's beliefs, opinions and statements about things or ideas in ongoing communicative interaction with other speakers. Speaker stance is firmly grounded in the speech situation and as such, stance-taking is crucial for the social construction of meaning in different discourses. In this study, we follow Du Bois' (2007) definition of stance.

> One of the most important things we do with words is to take a stance. Stance has the power to assign value to objects of interest, to position social actors with respect to those objects, to calibrate alignment between stance takers and to invoke systems of socio-cultural value.
>
> (Du Bois 2007: 139)

Rather than providing a catalogue of stance expressions, Du Bois aims at a general understanding of the phenomenon as such. For that purpose, he finds it is necessary to pinpoint the foundational principles of taking stance and negotiating meanings (Figure 1).

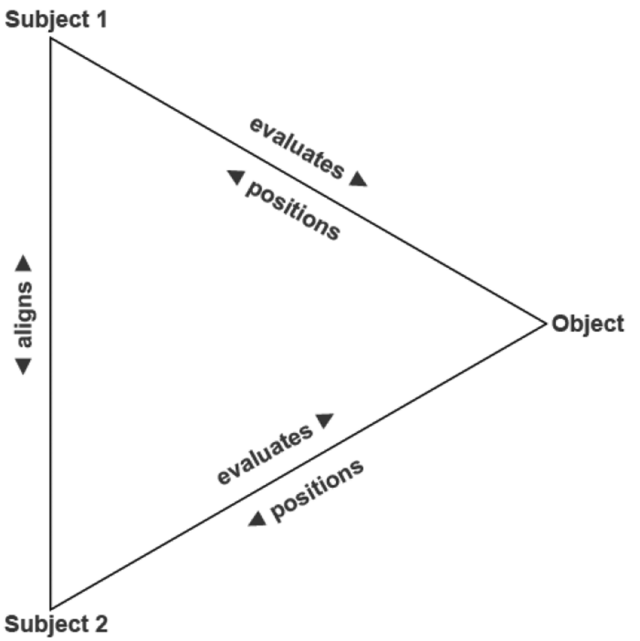

Figure 1: The stance act represented in the form of a stance triangle in Du Bois (2007: 163).

As Figure 1 shows, evaluation, positioning and alignment are three different aspects of a single stance act, where each aspect is distinguishable from the others through the consequences it has in the act. In the stance act, the stance taker evaluates an object, thereby positions himself or herself, and aligns with other subject(s). Du Bois' way of representing stance is important for our understanding of the phenomenon as such. However, to complete the picture we also add some more specifications about its nature.

(a) *Stance* is a psychological state involving speaker beliefs, evaluative ability and attitudes.
(b) *Stance-taking* is the performance by humans in communication, actions taken by speakers to express their beliefs, evaluations and attitudes toward (i) objects, scenes and events, and (ii) toward propositions, and speakers' viewpoints on what is talked about.

(c) *Expressions of stance* are constructions that are used to take stance and express meanings relating to speakers' beliefs, evaluations and attitudes.

Even though stance as a psychological state, as in (a) is important for our understanding of what stance is, it is beyond the scope of this study and therefore not addressed at all. Rather, it is considered a psychological prerequisite for this investigation. Stance-taking (b) and expressions of stance (c), on the other hand, are both ingredients of stance constructions and important for our analysis. They are the meaning side and the form side of what we find in the utterances.

Formally, stance markers are notoriously difficult to specify in advance because unlike some other categories, they are not confined to traditional areas of grammar, morphology or vocabulary, but to all of these as well as to longer chunks or even whole sentences. As has already been mentioned, some stance expressions are easily identifiable as such (*perhaps, surely, must, I don't know*) because they are elements that always and unambiguously express speakers' assessments. There are other constructions, however, that may not be thought of as stance markers in the first place since they are parts of lexical items such as *-able/ible* as in *doable, drinkable, possible*, but all of them come with a possibility judgement. The utterances in (1)–(4) below are all examples of advice of some sort, which can be expressed in a number of different ways. From the point of view of the words used to express the recommendations, some such utterances contain words that express a 'request' or 'recommendation', such as *I suggest* (1) and (2) and *I recommend* in (4), while (3) has no clear indication in terms of the individual words. In addition, the use of the past tense in (1) and (3) is an indication of distancing or toning down the speaker's accountability.

(1) *I would suggest that you left your house before the end of this month.*

(2) *I suggest that you leave your house before the end of the month.*

(3) *This wine should drink well within the next couple of years.*

(4) *I recommend that you drink this wine within the next couple of years.*

The use of the past tense, as with *would* in (1), evokes a distancing speaker stance. The speaker stands back and makes a tentative and polite recommendation to the addressee, while the opposite holds true of the speaker stance in (2), which is direct and therefore also much more of an urgent and potentially rude order. In (3), the speaker hides behind the wine. As the subject of a transitive verb in a middle construction, the wine assumes animate powers with the effect of backgrounding

speaker accountability as an assessor. In addition, the use of *should* in the construction adds speaker tentativeness and potential uncertainty on behalf of the speaker. The active first-person construction in (4), on the other hand, forces the speaker to stand up for his or her assessments (Paradis 2009; Hommerberg and Paradis 2014). As our approach to stance-taking is cognitive-functional and usage-based, we are interested in how speaker stance is conveyed in real utterances. For this reason, we explore stance from the point of view of how we interpret the meaning and the illocutionary force of the utterances in our corpus.

2.2 Previous studies of stance

What makes any description of stance research problematic is the fact that it is studied under a range of different names in different research traditions using different methods, and in addition to that, the sheer quantity of research is considerable. There are many works with an explicit mention of stance in the title (e.g. Conrad and Biber 2000; Hunston and Thompson 2000; Mushin 2001; Berman et al. 2002; Kärkkäinen 2003; Precht 2003; Hyland 2005; Englebretson 2007; Gray and Biber 2014), but, stance overlaps with and is closely related to notions such as modality, evidentiality, grounding, subjectivity/intersubjectivity, evaluation/appraisal, opinion/sentiment, as listed in Table 1 with examples

Table 1: List of stance category terms and examples of references.

Stance category terms	Examples of references
Modality	Palmer 1986; van der Auwera and Plungian 1998; Krug 2000; Nuyts 2000; Dendale and van der Auwera 2001 Facchinetti et al. 2003; Usonienė 2004; Boye 2012; Patard and Brisard 2011; Marín-Arrese et al. 2014; Fernández-Montraveta and Vázquez 2014
Evidentiality	Aijmer 1980; Chafe and Nichols 1986; Aikhenvald 2004; Cornillie 2007; Malmström 2008; Ekberg and Paradis 2009
Grounding	Langacker 2009; Verhagen 2005
Subjectivity/ intersubjectivity	Benveniste 1958; Langacker 1991; Traugott and Dasher 2002; Paradis 2003; Verhagen 2005; Tomasello 2010; Boye and Harder 2014; Gärdenfors 2014b
Evaluation/appraisal	Hunston and Thompson 2000; Martin and White 2003; Hunston 2008a, 2008b, 2011; Bednarek and Caple 2012; Fuoli 2012; Fuoli and Paradis 2014; Põldvere et al. 2016
Opinion/sentiment	Pang et al. 2002; Turney 2002; Kim and Hovy 2004; Pang and Lee 2004; Esuli and Sebastiani 2006; Pang and Lee 2008; Socher et al. 2013; Liu 2015; Taboada 2016

of references. Table 1 lists a number of representative examples of work on particular types of stance.

As we have already seen from the examples given above, the notion of modality is central to the notion of stance because the main contribution of items in that category is to express speaker attitude, either epistemic or deontic. Evidentiality is related to stance-taking because it involves the mentioning of where the information given comes from, and the reliability of the source in turn relates to the marking of how reliable the information provided by the speaker is. According to the reliability hierarchy of evidentiality, information given by a speaker who saw or heard something is taken to be more trustworthy than say second-hand information (reported information) or when speakers draw conclusions based on clues that they obtained from elsewhere. Thus, evidentiality is also about how the speaker obtained the information, and on the basis of that source, we infer to what extent this information can be trusted by the interlocutors.

Grounding is a term mainly used in Cognitive Linguistics. It relates to the speech event, the interlocutors, the time, the place, the situational context, previous discourse and shared knowledge of the speech-act participants. In other words, grounding is the process that links an entity or an event to the ground and thereby establishes mental contact between the event and the speech act situation. In nominal constructions, grounding is effected by determiners, demonstratives and sometimes by quantifiers, and at the clause level by markers of tense and modal verbs. Those markers are subjective because they describe the speaker's point of view. The speaker decides whether the situation described is real or potential and whether the hearer has previous knowledge of the information offered in the sentence. This is how the notion of grounding is linked to stance-taking.

Subjectivity and intersubjectivity are part and parcel of all the notions and research approaches in the list above. They are broad notions with a long research tradition and have played a role in linguistics at least since Benveniste (1958). Both notions concern the human experience of being a mental agent. Subjectivity refers the experience of oneself as a mental agent, while intersubjectivity refers the experience of others (Verhagen 2005: 4–8). Fifth, both evaluation and appraisal coincide with subjectivity and intersubjectivity in that the meanings and language resources they are related to are subjective and endorsed by the speaker, but they are also intersubjective in that they take the communicative situation and its agents into account (see Hunston for an insightful comparison of evaluation, appraisal and stance, Hunston 2011: 10–24). The study of stance in linguistics is mainly concerned either with theoretical approaches to different expressions of stance, i.e. what they mean, how they are used and how they differ, or with more descriptive types of

corpus work focusing on their use in discourse. The starting point is typically a word or a group of words that is known to express stance. There are also studies where the goal is to explore how a notional category such as epistemic modality or degree is expressed in a given language or in different genres.

Opinion and sentiment research aim at distinguishing different opinions, sentiments and information in relation to the speaker's stance-taking toward an idea or topic using various computational techniques. This is a research field that has grown immensely thanks to the availability of opinionated texts on the web. Much work is devoted to the development of information technologies that specialize in the development of information systems *per se* in order to describe people's behavior and thereby get a better understanding of opinions, their motivations, mechanisms and effects. We expand on this in Section 2.3 since our study is useful for computational purposes.

Our take on stance in this study is broad and including. All the above notions play a role for the scope of stance that we employ for our annotations and the analysis presented in Section 3. We make use of opinionated data, like studies on opinions and sentiments. However, our focus is *always* on the speaker's positioning, alignment and evaluation of what is talked about, and not whether a text is positive, negative or neutral.

2.3 Computational approaches to stance identification

The identification of speaker stance from a computational perspective is of topical research interest in Text Mining and Computational Linguistics. The methodologies of stance detection, stance identification and classification are grounded in similar principles, and in most cases, they are conducted using Data Mining and Machine Learning approaches, where the steps followed are concrete: data extraction and preprocessing, feature extraction, data training and testing, and evaluation of experimental results. Stance Classification is strongly connected to the fields of Subjective Language Identification (Wiebe et al. 2004), Opinion Mining and Sentiment Analysis (Pang and Lee 2008). The detection of stance-taking in discourse is important for our understanding of speaker attitude, response and favorability toward a topic, an idea and a situation.

In one of the early studies in the field, Whitelaw et al. (2005) used functional taxonomies based within the framework of APPRAISAL (Martin and White 2003) to perform Sentiment Analysis experiments. They used Pang and Lee's (2004) movie review corpus,[1] and a lexicon of appraising expressions. They stated that

1 See also: https://www.cs.cornell.edu/people/pabo/movie-review-data/

more fine-grained semantic information such as the appraisal categories improves sentiment classification. In another study, Saurí and Pustejovsky (2009) attempted to detect event factuality as a marker of speakers' positioning in relation to a specific topic. They defined linguistic clues that match an event as *factual, counterfactual, not totally certain, underspecified*, and created a text collection of 9,488 manually annotated events, the FactBank corpus.

The first studies on stance detection and classification from a computational perspective used data derived from ideological online debates (Somasundaran and Wiebe 2010). They created a lexicon with positive and negative entries, and showed that the sentiment- and argument-based systems outperform the baseline ones in overall accuracy (63.93 %). Anand et al. (2011) attempted to carry out a stance classification task on a corpus of 1,113 two-sided debates across 14 various topics. They used different feature sets in their approach: n-grams, cue words, post information, punctuation, etc., achieving classification accuracy up to 69 %.

Furthermore, Hasan and Ng (2013a, 2013b, 2013c) made advances in stance classification by testing various feature sets based on syntactic dependencies and information related to the preceding post of the thread examined. They tested these features in a corpus containing four data sets on different topics, achieving, in some cases, accuracies around 75 %. In a subsequent study, Hasan and Ng (2014) used the same corpus (each corpus entry was annotated as pro or con) in order to go beyond stance classification in order to detect the reasoning of each author's stance taking to a specific post.

Sridar et al. (2014) examined how both linguistic features and relations between authors and posts influence stance classification in a subset of the Internet Argument Corpus (Walker et al. 2012a, 2012b). They showed that a content-based approach can be significantly improved when information about interactions between authors and posts are incorporated in the methodology. Also, Faulkner (2014) set out to detect and classify stance in a different text type, namely student essays. He used the International Corpus of Learner English (ICLE: Granger 2003), and created a data set of 1,135 essays, 564 annotated as *for* and 571 as *against*. He used a stance lexicon and various other features, and performed experiments using different classification algorithms, achieving up to 82 % accuracy. Ferreira and Vlachos (2016) presented the Emergent data set for stance classification containing 300 rumoured claims and 2,595 associated news articles. This corpus was annotated with *for, against* and *observing* labels, and can be also used for other natural language processing (NLP) tasks.

With the expansion of the social media, the research interest shifted to the investigation of this text type. Rajadesingan and Liu (2014) tried to identify different types of stance (for or against a topic) in a collection of more 543,404

tweets from 116,033 different Twitter users. The SemEval-2016 Task 6[2] Evaluation Competition was dedicated to the stance classification with promising results (Mohammad et al. 2016; Zarella and Marsh 2016; Wojatzki and Zesch 2016). The participants were free to create their own data set on predefined topics and outcrop from the existing categories, the annotation value attributed to each tweet *in favour, against, none*. Mohammad et al. (2016) implemented a stance and sentiment investigation, and they achieved an accuracy level for the stance part of up to 69 %. They observed that sentiment features improve the stance classification results, which indicates that knowledge of the sentiment in a tweet facilitates the identification of stance. In their study, Kucher et al. (2016b) described an approach for stance analysis based on sentiment or certainty considerations and presented the uVSAT tool for visual stance analysis, the analysis of temporal and textual data, and the exportation of stance markers in order to prepare a stance-oriented training data set.

In the studies presented, we observe that most researchers deal with the issue of the automatic stance detection as a pro/con binary problem, while more often than not, stance is not an either-or phenomenon, but a matter of degree of the force of the expression in the context where it is used. Stance classification focuses on the detection of the author's positioning to a given topic, which usually is controversial in terms of opinion making. The data are in many cases pre-labelled utterances into the pro/con classes, as in Twitter with the hashtag labels (i.e. #not, #pro, #pride, etc.).

Our starting point differs from these studies. We adopt a cognitive-functional, holistic view of stance-taking that is gradient and context sensitive rather than a simple lexically driven approach. Our view of human communication is that meanings in general and opinions in particular are negotiated all the time, and our annotation schema is designed to reflect this view. The categories in the BBC are described in Section 3.

3 Stance categories

For the purpose of this study, we have identified a broad framework of ten notional stance categories from the literature reviewed in Section 2. In alphabetical order, the categories are listed as follows: AGREEMENT/DISAGREEMENT, CERTAINTY, CONTRARIETY, HYPOTHETICALITY, NECESSITY, PREDICTION, SOURCE OF KNOWLEDGE, TACT/RUDENESS, UNCERTAINTY and VOLITION, as shown in

2 SemEval-2016 Task 6: http://alt.qcri.org/semeval2016/task6/

Table 2. This list comprises a wide spectrum with the most important stance-related concepts according to the research described in Section 2.

Table 2: Stance categories.

Stance category	Description	Examples of utterances
AGREEMENT/ DISAGREEMENT	The speaker expresses a similar or different opinion.	I couldn't agree more to what you are saying. No, please don't do that. In contrast to you, my opinion is that we should try.
CERTAINTY	The speaker expresses confidence as to what she or he is saying	I am sure they will fight about it. Of course it is true. Without a doubt, you will be there before 6 o'clock.
CONTRARIETY	The speaker expresses a compromising or a contrastive/comparative opinion.	While these are kind of notes to myself, you might still find them useful. The result is fairly good, but it could be better. Despite the weather, I took him for a walk.
HYPOTHETICALITY	The speaker expresses a possible consequence of a condition.	If it's nice tomorrow, we will go. I will be happy, if Mike visits Granny tomorrow. How am I going to be able to catch the plane, if I can't use the car?
NECESSITY	The speaker expresses a request, recommendation, instruction or an obligation.	I must hand back all the books by tomorrow. You have to leave before noon. This wine should drink well for two more decades.
PREDICTION	The speaker expresses a guess/ conjecture about a future event or an event in the future of the past.	My guess is that the guests have already arrived. The meeting should not last longer than 2 hours. That ought to be fine.
SOURCE OF KNOWLEDGE	The speaker expresses the origin of what he or she says.	I saw Mary talking to Elena yesterday. According to the news, the rate of interest is not going up. It was obvious that she didn't want to talk.

(*continued*)

Table 2: (*continued*)

Stance category	Description	Examples of utterances
TACT/RUDENESS	The speaker expresses pleasantries and unpleasantries.	Please, do give my love to him. You lazy bastard. Get lost. Don't you think it might be a good idea to postpone the meeting until tomorrow.
UNCERTAINTY	The speaker expresses doubt as to the likelihood or truth of what she or he is saying.	Surely we have enough time. We have enough time, haven't we? There might be a few things left to do.
VOLITION	The speaker expresses wishes or refusals, inclinations of disinclinations.	We wanted him to sell the house. I wish I could join you next summer. I prefer to stay in a cheap hotel.

In Table 2, each category is followed by a brief description of the type of stance-taking associated with each notional category. Examples are given in the third column. The stance-taking elements are constructions of varying length. For instance, *Don't you think it might be a good idea to postpone the meeting until tomorrow* is formed in a tactful and polite way while the word *wanted* in *We wanted him to sell the house* is solely responsible for the volitional reading. As already pointed out, stance constructions include individual stance-marking items such as *must, could, possibly* as well as longer stretches such as *It's about time you sold your car* where the sequence starting *it's about time* + past tense of *sell* are indicators of an interpretation that expresses a rather direct and rather tactless recommendation to somebody else about what the speaker thinks is beneficial for the addressee.

The descriptions of the stance categories in Table 2 were used as instructions for the annotators. The categories are not mutually exclusive but co-occur in the utterances. For instance, the category SOURCE OF KNOWLEDGE includes evidence from the five senses (sight, touch, hear, taste, smell) as well as expressions referring to inferential reasoning such as *it seems*. The coding for an utterance with *it seems* is not only an expression of knowledge source but also of UNCERTAINTY and maybe also other stances depending on the utterance as a whole, as we will see in Section 4. Furthermore, some categories encompass two properties. For instance, hypothetical constructions *per definition* involve a prediction. A decision was, however, made not to place them in two categories, but to code them as HYPOTHETICALITY only since such utterance will always also be prediction. Two of our categories comprise opposing

notions within the same meaning-function domain: AGREEMENT and DISAGREEMENT are concerned with alignment *versus* disalignment with the addressee at the functional level, while TACT and RUDENESS are contrastive notions on the dimension of politeness. In the case of CONTRARIETY and NECESSITY, the naming of the categories contains two similar representations. CONTRARIETY involves both CONCESSION and CONTRARINESS, which are both contrastive construals (Paradis and Willners 2011; Jones et al. 2012), and NECESSITY includes NEED and REQUIREMENT, e.g. obligation, requests for the benefit of the speaker and recommendations which are issued by the speaker for the benefit of the addressee (Paradis 2009).

4 The Brexit Blog Corpus

In this section, we describe the design of the BBC and how the collection of the data was carried out. The rationale for the compilation of the corpus is summarized in (i)–(iv).

(i) This social media text type was expected to contain subjective language and thus stance-taking information.
(ii) The BBC serves as the benchmark to test the validity, applicability and usage patterns of the ten stance categories.
(iii) The BBC enables us to carry out further computational research on stance in order to arrive at a more integrated and plausible analysis of speakers' positioning, alignment and evaluation in online social media discussions.
(iv) The annotated gold standard corpus is a linguistic resource, which will be publicly available to the research community for further study and analysis through SND.[3]

In Section 4.1, we describe the data collection process and in Section 4.2 the annotation.

4.1 Corpus description and annotation process

The BBC is a collection of texts from blog sources. In line with principles governing traditional corpora (Gilquin and Gries 2009), it consists of a time-bound data set of networked assemblages of streaming data (Zappavigna 2012). In the literature, blogs are defined as a distinct text type differing from social media texts

3 The BBC in the Swedish National Data Service: https://doi.org/10.5878/002924

from other sources in terms of structure and size (Baldwin et al. 2013; Myers and Hamilton 2015; Berger et al. 2015). The blog post texts can be considered as a text type that shares many characteristics with impromptu spoken conversation in that the contributions are relatively spontaneous and very often not supervised, and not revised or according to the norms of formal writing. However, blogging suffers less than microblogging, e.g. Twitter posts, from features that impose difficulties in analytical tasks such as non-standard orthography, omitted characters and other spelling issues, special characters and XML tags.

The corpus texts are thematically related to the 2016 UK referendum concerning whether the UK should remain members of the European Union or not. The texts were extracted from the Internet from June to August 2015. With the Gavagai API,[4] the texts were detected using seed words such as *Brexit, EU referendum*, etc. We retrieved and filtered only URLs ending in wordpress.com, blogger.com, blogspot.*. Only entries described as blogs in English were selected. Each document was split into sentential utterances, from which 2,200 utterances were randomly selected. The final size of the corpus is 1,682 utterances, 35,492 words (169,762 characters without spaces). Each utterance contains from 3 to 40 words with a mean length of 21 words.

For the data annotation process the Active Learning and Visual Analytics (ALVA) system implemented by Kucher et al. (2016a, 2017) was used (see Figure 2). At the top of the annotation interface page, the utterances to be annotated appear in a box one at a time and beneath the utterance box are the stance categories and three options: (i) *Submit annotation*, (ii) *Mark as neutral* and (iii) *Mark as irrelevant*. The first option was used in cases where at least one of the stance categories is present. The second one when the utterance did not contain any explicitly worded stance-taking expressions. The third option was used for incomplete utterances and boilerplate text chunks such as *Posted by [name]*, i.e. where no statement, question, promise or order is issued. As mentioned above, annotators could attribute more than one category in the same utterance as in (5)–(7).

(5) *Different political contexts, I know, but the core principle of the Eurozone that the bondholders must be paid every cent due to them – will be defended to the hilt and regardless of the human consequences*. → AGREEMENT, CONTRARIETY, PREDICTION

(6) *Regrettably I believe that Farage is going to prove to be a negative influence that may even guarantee an unsuccessful outcome*. → UNCERTAINTY, PREDICTION

4 Gavagai API: https://developer.gavagai.se

(7) *Once again, this is not the talk of doom-mongers, but 'respected' institutions everywhere.* → SOURCE OF KNOWLEDGE, CONTRARIETY

Neutral is a problematic notion, which we make use of in a simplistic and practical way. Utterances were categorized as neutral in cases where none of our ten stance categories was identified, like in the examples (8)–(11).

(8) *In the very centre of the lake stood a mighty water wheel, known as the Wheel of the Cycle of Life.*

(9) *A family trip is different from a bike trip with friends.*

(10) *He does anything at all, he is held up as a monster and evil.*

(11) *So we will see ... Aussie comes with daily chart.*

All sentences that were found to be irrelevant by at least one of the annotators were removed. In Figure 2, we present an instance of the ALVA annotation system used for the annotation of an utterance.

In the study reported here, the annotations were self-paced. When the annotator had classified the utterance, he/she ticked the submit box and a new utterance appeared in the box. It was not possible to go back to previous utterances to reconsider and revise the annotation. Two annotators, one with a Licentiate degree in English Linguistics and the other with a PhD in Computational Linguistics, carried out the annotations independently of one another. They followed the protocol that is schematically presented in Figure 3.

A manual with information about the annotation process, basic knowledge about the stance framework and the annotation tool was distributed to the two annotators. After studying the stance annotation manual carefully, the annotators participated in a seminar given by a senior linguist expert, where questions about the categorization and the process were addressed, and all outstanding issues resolved. The rationale of the annotation scheme was explained in detail on the basis of the information in Table 2 and the examples were discussed. Before the annotation session, there were five pilot rounds, each consisting of 100 utterances from our data set. After the pilot round, the annotators were given the opportunity to discuss their annotations with the senior linguist expert in order to ascertain that they had understood the task correctly and were able to work in accordance with the categorization scheme. The annotation procedure included two rounds of annotation by each of the annotators. The outcome of this procedure is evaluated and discussed in Section 4.2.

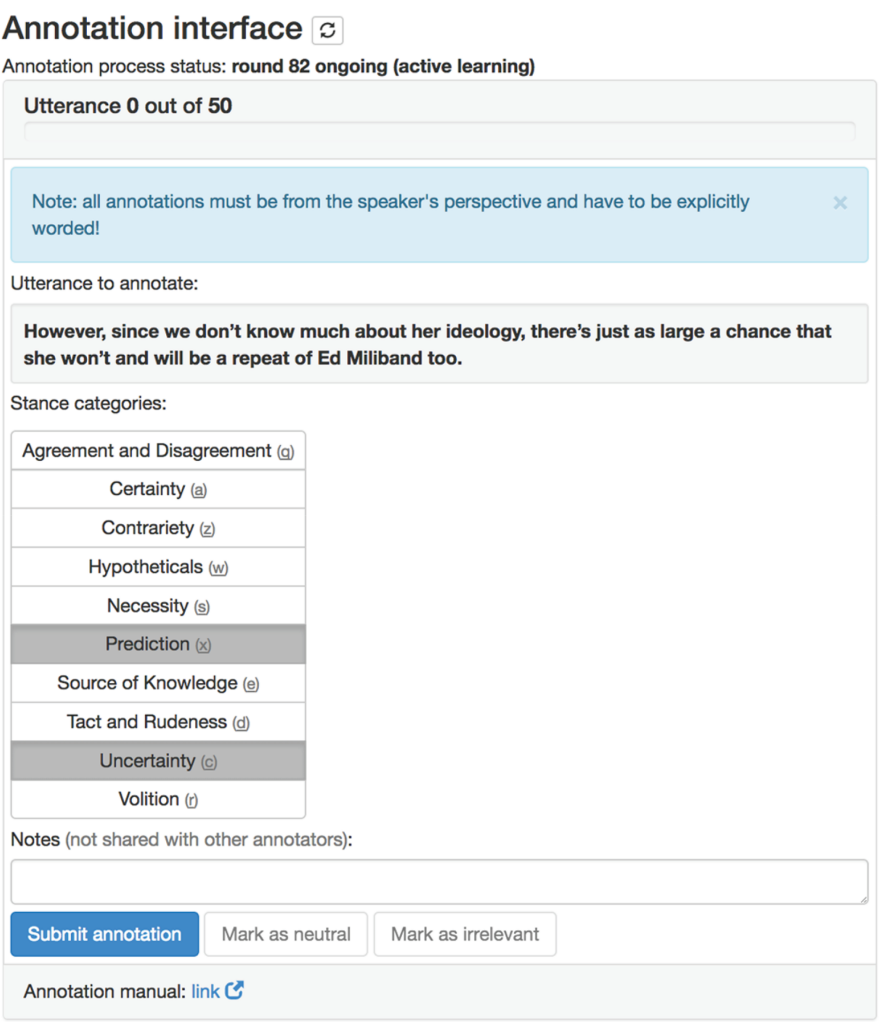

Figure 2: Annotating with ALVA: the annotator's decision about the utterance *However, since we don't know much about her ideology, there's just as large a chance that she won't and will be a repeat to Ed Miliband too.* is PREDICTION and UNCERTAINTY.

4.2 Annotation results

In order to evaluate the annotation results, we used two metrics: the *inter*-annotator agreement, which calculates the agreement between the annotators' decision about an utterance, and the *intra*-annotator agreement, which tests the annotator's second decision in relation to the first one about the same utterance.

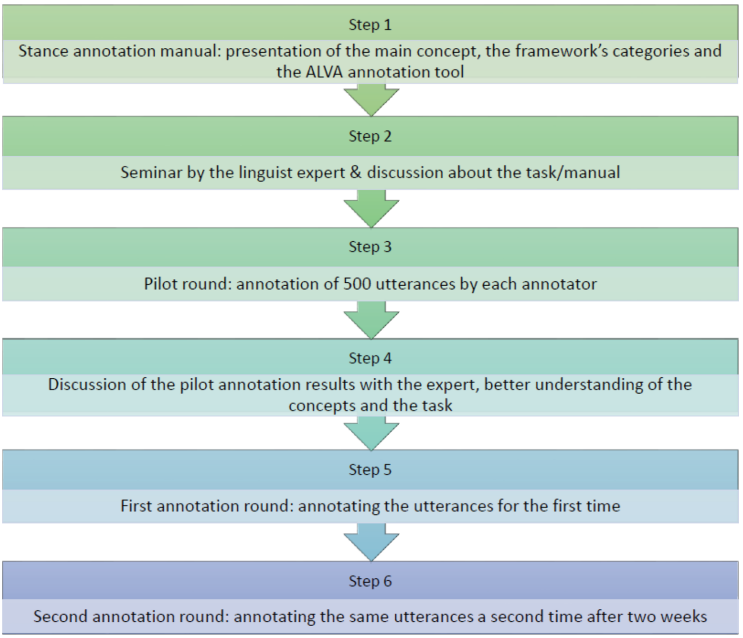

Figure 3: The protocol followed in the annotation process of the BBC.

In the final version of the annotated corpus, the annotation disagreements between the two sets were discussed and problems resolved, and one of the two conflicting annotations was chosen, resulting in a gold standard version of the annotated corpus. In Table 3, the inter-annotator and the intra-annotator agreement sets are presented, as well as the gold standard corpus. All results reported in this section are based on the gold standard data.

In Table 3, the output of the annotation process is presented. The first column shows the number of utterances annotated for each of the categories by each annotator. The second column presents the number of utterances attributed to the stance categories after the first and after the second rounds by the same annotator. Finally, the third column shows the final gold standard corpus after the annotation process, namely the number of annotations attributed to each stance category. In Figure 4, we illustrate the distribution of the stances in the BBC according to their annotation.

Figure 4 shows the distribution in terms of frequency of the different stances in the BBC. As discussed in Section 4.1, the utterances may express more than one stance type. In the gold standard corpus, 874 utterances out of 1,682 utterances in total were annotated with two stance categories. In 252 cases,

Table 3: The inter- and intra-annotator agreement sets.

Stance categories	Number of annotated utterances		
	Inter-annotator agreement set (first/second annotator)	Intra-annotator agreement set (first/second decision)	Number of utterances in gold standard
AGREEMENT/DISAGREEMENT	28/8	12/25	50
CERTAINTY	33/42	47/45	84
CONTRARIETY	109/127	151/214	352
HYPOTHETICALITY	85/75	81/85	171
NECESSITY	91/97	94/98	204
PREDICTION	139/83	135/141	252
SOURCE OF KNOWLEDGE	100/114	148/113	287
TACT/RUDENESS	16/27	14/14	44
UNCERTAINTY	121/78	89/86	196
VOLITION	24/25	12/13	42

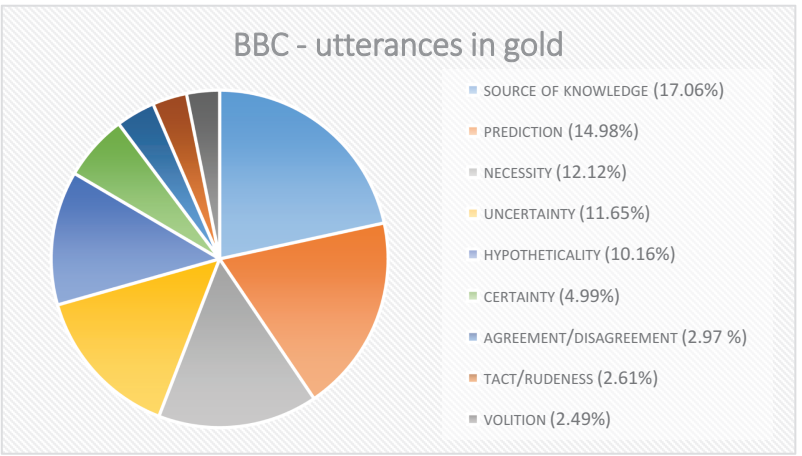

Figure 4: The distribution of the stance categories in the BBC in descending order.

three categories were attributed to one utterance. In 61 cases, utterances were annotated with four different stance categories, and 11 cases with five categories. In Figure 5, we show the category combinations of annotated utterances where more than one stance category was attributed.

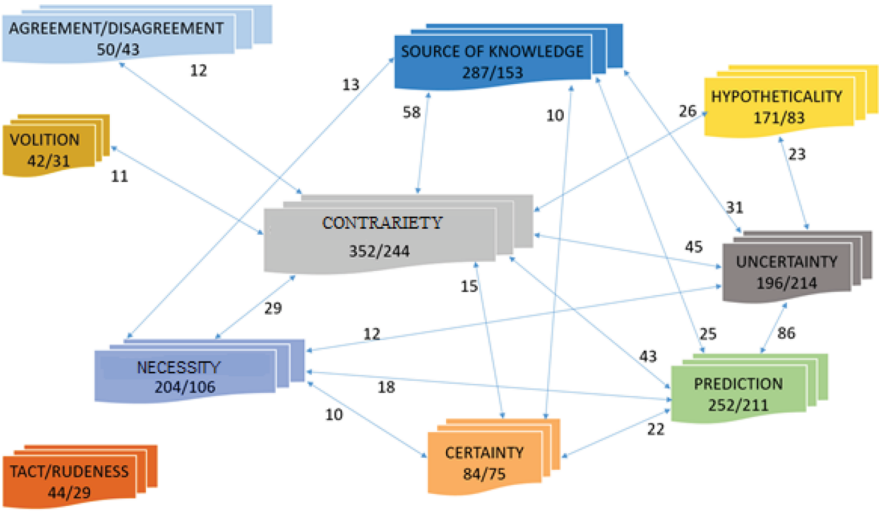

Figure 5: The different categories co-occurrences, after the annotation process.

Figure 5 shows the results for the cases where the combination occurred ten times or more. The numbers in the stance category text boxes refer to the total number of utterances annotated as such (the first one) and to the number of the additional stance annotations (the second one). For instance, in 244 cases out of the 352 utterances annotated as CONTRARIETY, more stances are detected and annotated. We observe that the CONTRARIETY category co-occurs frequently (58 utterances) with SOURCE OF KNOWLEDGE, and less often with VOLITION (11 cases). The high number in some cases of these co-occurrences, as in UNCERTAINTY (214 categories co-assigned) shows that more than two categories were attributed to these utterances, and consequently this is an indication of the relation or interaction among the stance categories. The labelled arrows show how the stance annotations combine and in how many utterances this co-occurrence was attributed.

Our next step was the evaluation of the annotation results. The metrics were used in order to calculate the agreement level of the annotations of the different categories in the BBC. For these experiments we used two agreement measures, the *F*-score (van Rijsbergen 1979) and Cohen's kappa (Artstein and Poesio 2008). As confidence interval (95 %) calculations for both measures, a bootstrap resampling based on percentiles using a 10,000-fold resampling was used (Myers and Hsueh 2001). The mean *F*-score and kappa were based on the mean result from the 10,000 resampled folds. In Table 4, the results of these measures are presented for both inter- and intra-annotator agreement sets.

Table 4: The *F*-score and the kappa for the inter- and intra-annotator agreement sets.

Stance categories	Inter-annotator agreement set		Intra-annotator agreement	
	Mean *F*-score	Mean kappa	Mean *F*-score	Mean kappa
AGREEMENT/DISAGREEMENT	0.21	0.20	0.58	0.58
CERTAINTY	0.45	0.42	0.67	0.65
CONTRARIETY	**0.78**	**0.76**	0.76	0.71
HYPOTHETICALITY	**0.78**	**0.76**	0.79	0.77
NECESSITY	0.77	0.75	0.79	0.76
PREDICTION	0.57	0.52	0.78	0.75
SOURCE OF KNOWLEDGE	0.53	0.47	0.72	0.68
TACT/RUDENESS	0.55	0.54	0.78	0.77
UNCERTAINTY	0.62	0.58	**0.81**	**0.79**
VOLITION	0.44	0.43	0.71	0.70

In Table 4, we observe that for the inter-annotator agreement set, the highest evaluation score appears in the cases of CONTRARIETY and HYPOTHETICALITY annotations. Their *F*-score is up to 0.78, which is a level of high agreement among the annotators for these two categories. The lowest score, on the other hand, is observed for the AGREEMENT/DISAGREEMENT category (0.21). For the intra-annotator agreement set, the best score appears in the UNCERTAINTY category, achieving 0.81 as *F*-score. Except for the lowest score, again, for the AGREEMENT/DISAGREEMENT category (0.58), all other categories have high agreement scores. These results are discussed and assessed in Section 5.

5 Discussion

In this study, we annotated stance in utterances extracted from blog posts. Based on the semantic information of the utterance, the annotators determined which stance/stances was/were expressed among ten different categories. In many cases, the speaker's positioning was complex, expressing multiple stances in the same utterance. In other cases, no stance was expressed according to our stance categorization scheme. The final gold standard corpus shows that it is a possible task to determine speaker stance in texts from social media sources, in this case blog posts, for stance based on notional, cognitive-functional categories, and that the stance category framework was sufficiently clear and intuitive to be used as a practical protocol for annotations. No major divergences between the annotators' decisions about the same utterances were observed.

This gives credibility to our annotated data and enables us to use them for multiple purposes in future research.

In this section, we discuss and evaluate the annotation process and the results. Concerning the distribution of the annotated utterances of the BBC, we showed in Section 4 that there was a difference between the stance categories with regard to how often they occur. CONTRARIETY is clearly the most often used stance category by the political blog text authors. This suggests that contrast is an appropriate rhetorical device for expressing opinions in a more balanced way as in (12).

(12) *I don't disagree that the YES campaign made mistakes, but they must be contextualised within the truly immense opposition they had.*

In (12), the speaker has a clear opinion, expressing it using a contrastive sequence, starting with *I don't disagree*, followed by a clause, starting with the adversative conjunction *but*, creating a balance, resulting in a less blunt statement.

The second most frequently occurring stance category is SOURCE OF KNOWLEDGE. We confirm in our data the common practice of people to refer to the opinions expressed by others, which is a typical pattern in conversations about political issues. In (13), we can see how referring to others is a means used to back up arguments with facts from an authoritative source.

(13) *All the polls indicate it would happen again tomorrow if the process were to be repeated.*

Prediction is the third most frequently occurring stance category in these data, and it appears to be an important component in talking about future elections, EU and politics in general, as in (14).

(14) *The current crisis could easily lead to the country leaving the euro and eventually the union itself.*

Less frequent, but still occurring relatively frequently, are the three categories NECESSITY, UNCERTAINTY and HYPOTHETICALITY. Utterances belonging to the NECESSITY category often take the form of recommendations or as a rhetorical device in which the reader is directly addressed. Expressions of UNCERTAINTY are used to convey doubt. HYPOTHETICALITY is mainly expressed in conditional sentences, and after a closer look at the annotation data, these forms/structures confirm their hypothetical meaning and function.

The remaining four categories appear less frequently in the data set. CERTAINTY is only half as common as its opposite category UNCERTAINTY. AGREEMENT/DISAGREEMENT and TACT/RUDENESS are more likely to occur rarely due to the non-dialogic nature of the text type in this study. Expressions that indicate that these two categories are more likely to occur often when the text is directed toward a specific reader (in an online dialogue) and less often when speaker interaction is limited, as in the case of blog posts. This is, for instance, shown when we compare the occurrence frequencies obtained here to those in previous works (Skeppstedt et al. 2016), in which 31% of the sentences extracted from debate forum posts contained expressions conveying agreement or disagreement. The final category, VOLITION, might be infrequent in texts on political topics of this kind, and many of the utterances annotated as VOLITION were wishes.

Moreover, it is important to note that due to utterance complexity more than one stance category were attributed to the 52% of the BBC (874 utterances). In Figures 6–8, we show schematically the co-occurrences of stance categories in utterances. In Figure 6, we can see that AGREEMENT/DISAGREEMENT frequently co-occurs with CONTRARIETY (in 12 cases out of 42 co-occurrences). In (15), we give an example showing that in utterances where the speaker agrees or disagrees he/she expresses a contradiction too.

(15) *In principle I agree with what he is striving for, but the in practice it is not so simple.*

In the right-hand chart of Figure 6, the co-occurrence of CERTAINTY with other stance categories is presented. We observe that CERTAINTY is primarily

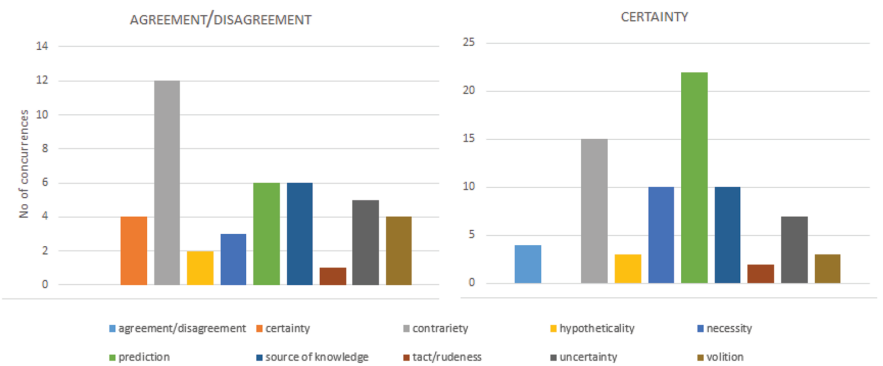

Figure 6: The stance categories that co-occur with AGREEMENT/disagreement (on the left) and with CERTAINTY (on the right).

co-occurring with PREDICTION (in 22 utterances out of 75 co-occurrences), with CONTRARIETY (14 cases) and with SOURCE OF KNOWLEDGE and NECESSITY (10 cases). In the examples below, we show how these combinations may be worded and explain what made the annotators make their decisions.

(16) *as for brexit ... its simply not going to happen.*

(17) *Further, these economic reasons are not just clear, but overwhelming.*

(18) *The bottom line is that native born Scots did indeed vote by a small majority for independence.*

(19) *It is clear that any significant reforms will require a treaty change, which will require later ratification by all 28 EU states.*

In (16), the speaker expresses a prediction that Brexit is not going to happen in an assertive way, and the use of *simply* reinforces his or her position about the prediction he/she makes. In (17), the speaker juxtaposes two properties in a contrastive construction. The first part, including *not just clear*, triggers the readers' expectations of a property that is stronger than 'clear' here expressed by *overwhelming*. This sliding along the scale, the contrastive adversative construction and the final extreme evaluation are indications of the speaker's strongly assertive stance. In (18), the speaker alludes a reported source of knowledge reinforced by the assertive expressions *the bottom line* and *did indeed*. And in (19), in the first part of the utterance, CERTAINTY is expressed through *it is clear*, which takes scope over the rather assertive and imperative consequence expressed in the subclause. CERTAINTY does not occur often with AGREEMENT/DISAGREEMENT or HYPOTHETICALITY. In the first case, we could assume that speakers, when expressing their agreement or disagreement do not use forms or structures of certainty, so they agree or disagree with something or someone in a gentler or milder way. Additionally, conditionals in HYPOTHETICALITY express a low level of certainty by their semantic nature as being irrealis expressions.

In Figure 7, the stance categories that co-occur with CONTRARIETY and HYPOTHETICALITY are presented. As we have already said, CONTRARIETY is the most frequent category, and it is also the category that co-occurs often with other stances. Its co-occurrence with SOURCE OF KNOWLEDGE is the most frequent one (in 58 utterances out of 244 co-occurrences), but also appears quite frequently with PREDICTION and UNCERTAINTY (43 and 45 cases respectively), NECESSITY (29 cases) and HYPOTHETICALITY (26 cases), and less frequently with the other categories. CONTRARIETY is the most interactive one with the other

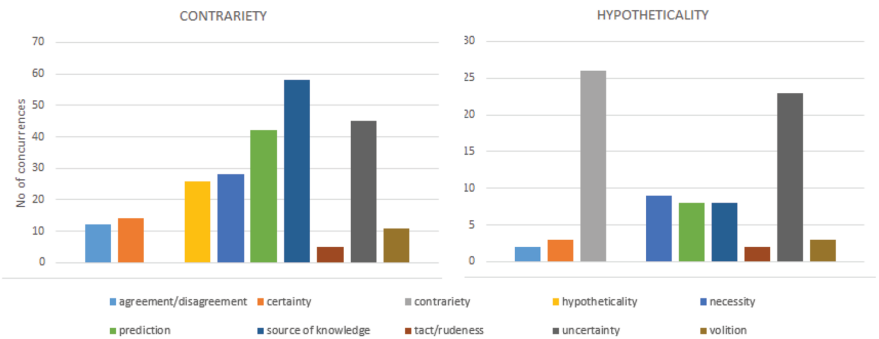

Figure 7: The stance categories that co-occur with the CONTRARIETY (on the left) and HYPOTHETICALITY (on the right).

stances, and it is compatible with all of them. In (20)–(22), we show examples of the most frequent co-occurrences.

(20) *Not this week or next year or even the year after that, but it will come about.*

(21) *Statistics also show, that despite or because of the NHS, no one gets out of here alive!*

(22) *I think it might have happened before the referendum, but for Project Fear and the BBC's slavish promotion of it.*

In (20), PREDICTION is co-occurring with CONTRARIETY. The structure *negation* + *but* expresses contradiction and has also a predictive meaning. In the utterance in (21), we can identify SOURCE OF KNOWLEDGE, where the objective source *statistics* is employed to confirm the statement made by the speaker, which was annotated as CONTRARIETY. In (22), we observe the markers of uncertainty *think* and *might* in an utterance where CONTRARIETY and UNCERTAINTY co-occur.

CONTRARIETY has a wide range of forms through which it is worded (and in addition to that it interacts significantly with other stances and their ways of being worded), and it appears to be a common construction for speaker positioning. The combinations of HYPOTHETICALITY annotations co-occur mostly with CONTRARIETY (26 utterances) and with UNCERTAINTY (23 utterances). Conditional structures are often speculative and as in (24) include a comparative and contrastive, which enhance the role of these other two stances. Apart from the conditional structure that expresses HYPOTHETICALITY, the lexical forms *rather than* and *perhaps* evoke CONTRARIETY (23) and UNCERTAINTY (24) stances.

(23) *If there's a major breakaway, the SNP could end up being helped rather than harmed.*

(24) *Perhaps this is credible if one thinks Britain is as mismanaged at home and ineffectual abroad as Italy.*

In Figure 8, the combinations of NECESSITY and PREDICTION with other stance categories are presented. The NECESSITY category is compatible with other stances, but most often it combines with CONTRARIETY, as in (25).

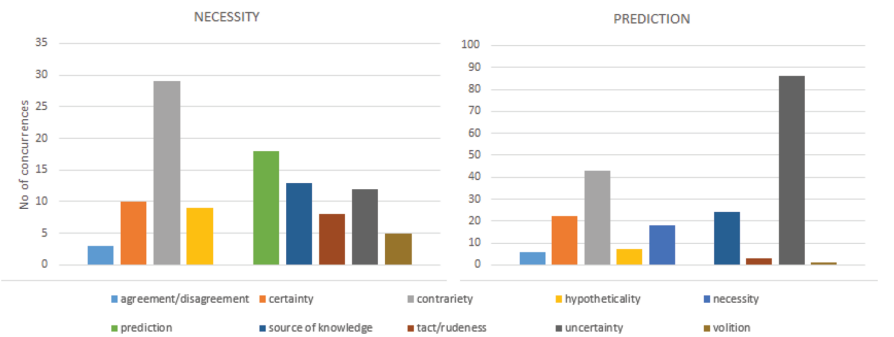

Figure 8: The stance categories that co-occur with the NECESSITY (on the left) and PREDICTION categories (on the right).

(25) *This campaign should focus money on winning, not making staff rich.*

In the case of PREDICTION, the most frequent co-occurrence is UNCERTAINTY (in 86 utterances) given that predictions are not facts or certain events, and cannot be adopted with confidence, see (26).

(26) *On the one hand, as some Tory MPs will clearly cause internal division, this might suggest that Cameron will have a hard time convincing Conservative voters of a 'Bremain'.*

In (26), we find expressions of uncertainty (*might, suggest*) which attenuate the strength of the prediction and the related level of confidence.

As for the categories SOURCE OF KNOWLEDGE and VOLITION, we see a similar distribution. Apart from a second stance that frequently co-occurs with the first category, not many utterances are annotated with co-occurring stances. SOURCE OF KNOWLEDGE co-occurs most often with CONTRARIETY (58 utterances). In 11

cases out of the 42 utterances annotated as VOLITION, the CONTRARIETY category was co-assigned. Not many co-occurrences are annotated with TACT/RUDENESS. It is not possible to draw any conclusions about why this is the case because of sparsity of the data. Finally, the UNCERTAINTY utterances are mostly co-annotated with PREDICTION, CONTRARIETY and SOURCE OF KNOWLEDGE. UNCERTAINTY is the most co-annotated category, where more than one stance is frequently attributed to the same utterance.

Concerning the annotation agreement results, we provide two potential explanations about the results. The inter-annotation results provide useful information about the efficacy and the reliability of the proposed framework. More specifically, a high degree of disagreement (low inter-annotator scores) between the annotators' decisions in the annotation of one or more stance categories to the utterances may indicate either paucity of this stance in discourse, which may make its recognition a difficult task for the annotators (especially in the absence of explicit, straightforward cues such as *if, so, but*, etc.), or the inefficiency of the framework's notional category for detection. The divergence among the annotator's decisions could also be due to lack of linguistic items that could lead them to the detection of a stance in an utterance. Another clue that may explain the annotators' disagreement can be the annotators' different interpretations as a consequence of their intuitions about the semantic information that each utterance carries. In contrast, the agreement (high inter-annotator scores) between the two annotators in the attribution of stance is an indication of the reliability of the annotation scheme, or there were unambiguous linguistic clues that made the annotation decision easier and more evident for the annotators.

While the CONTRARIETY, HYPOTHETICALITY, NECESSITY and UNCERTAINTY categories showed good inter-annotation agreement score, the remaining six categories did not have high scores. In the intra-annotation agreement, we also observe differences among the different categories' scores, but not to the same extent as the inter-annotation agreement results. The intra-annotation scores are on a similar level with the highest value found in the UNCERTAINTY category (0.81). All of them are over 0.65 with the only exception of the AGREEMENT/DISAGREEMENT category, which is the lowest (0.58).

Regarding the *F*-score and the kappa measurements, the few occurrences of the utterances annotated for CERTAINTY, AGREEMENT/DISAGREEMENT, TACT/RUDENESS and VOLITION led to very wide confidence intervals, and, therefore, no informative measures for intra- or inter-annotator agreement could be derived. Among the other six categories, for which more informative agreement

measures were obtained, it can be concluded that for CONTRARIETY, HYPOTHETICALITY and NECESSITY, reasonable[5] agreement figures were obtained. For these three categories, there were no major differences across the results obtained for the inter- and the intra-annotator agreement. For UNCERTAINTY, PREDICTION, and, in particular, SOURCE OF KNOWLEDGE, reasonable levels were obtained for intra-annotator agreement, but low inter-annotator agreement scores.

A critical comment about the corpus' size could be that the BBC is a small text collection in comparison to automatically generated and annotated corpora. The extraction of blog posts was easily implemented using the Gavagai API, but the annotation of the sentences by human annotators was a laborious and time consuming task to perform. A great deal of previous work on stance annotation of text data has been about a binary decision, whether an utterance is for or against a specific topic. But, in the present study, the goal was to proceed in a way that better mirrors language understanding in natural communication focusing on a relatively free task carried out by language users on the basis of instructions describing cognitive-functional categories. The goal of this was to investigate whether it is at all possible to successfully make use of such categories instead of lexically driven tasks that only make use of form. We thereby took the initial steps to delve deeper into what stances speakers take and the way they are expressed in a sentence and thereby to highlight the dynamics of word meaning in discourse. It is clear that the annotation decisions depend not only on how the annotators perceived the semantic information in the utterances, but also on the speakers' purpose and scope when they chose a specific form/structure for the creation of an utterance. Annotators are of course constrained by their own intuitions and general knowledge about how communication evolves on different occasions and in different contexts. They identified speaker stance in writing in each utterance and decided among ten different categories, which is a demanding task. For instance, the utterance in (27) should not be annotated with the CERTAINTY label as the speaker is trying to present someone else's position about a topic.

(27) *The PM was sure about the referendum's result.*

In an automatic processing of (27), the annotation tool would select the CERTAINTY label mainly based on the word *sure*, while human annotators are

5 Reasonable according to Landis and Koch (1977) who categorize agreements between 0.61 and 0.80 as substantial, while Artstein and Poesio (2008) set 0.67 as threshold for reasonable agreement.

able to distinguish such differences and understand the potential difference between the speaker's actual positioning, which in this case is annotated as SOURCE OF KNOWLEDGE. The annotators had to take into account such cases and decide accordingly, and as a result of the procedure followed in our study, we can be confident that the output, the annotated corpus, is a carefully annotated data set that will be used for further research purposes and in various tasks.

6 Summary and conclusion

This interdisciplinary study is an experiment in annotating a corpus using cognitive-functional categories as the basis for speaker stance identification. We start by reviewing the linguistics and the computational linguistics literature in order to provide a comprehensive background of work on stance in both fields. This review of previous work is meant to promote synergies and to point to the plethora of terms used for stance in general. On the basis of this review, ten different notional stance categories were identified, defined and exemplified, namely AGREEMENT/DISAGREEMENT, CERTAINTY, CONTRARIETY, HYPOTHETICALITY, NECESSITY, PREDICTION, TACT/RUDENESS, SOURCE OF KNOWLEDGE, UNCERTAINTY and VOLITION. This information was then written up as a practical tool for corpus annotation (Table 2). In order to test the viability of annotating a corpus using cognitive-functional categories, we compiled a corpus of political blogs, the BBC, which is a novel linguistic gold standard resource. Using the ALVA system, two experts performed two rounds of annotation, attributing one or more stance labels to the utterances. We measured the annotation agreement scores in order to determine the reliability of the annotations, both inter- and intra-annotation agreements, and obtained good results, in particular for the intra-annotation agreement. The highest score for inter-annotation agreement was found for the CONTRARIETY and HYPOTHETICALITY categories and for the intra-annotator agreement for the UNCERTAINTY category.

Through our methodology, we approach stance in a different way from many studies both in linguistics and computational linguistics, where researchers in most cases start with a set of words that they assume have a given meaning (at least most of the time), which is far from always the case as we have shown in this article. Words and structures through which people take stance are polyfunctional and capable of expressing more than one stance type. In this study, the point of departure for the annotators was notional in the sense that it was their task to identify which stance types were expressed by the speaker in each utterance. The BBC is a gold standard linguistic resource that

can be useful for both theoretical and computational studies. Linguists can perform analyses of real language in use; computer scientists have a data set with predefined classes ready to perform stance classification experiments or other NLP and visual analytics tasks. The next step includes the development of automatic methods in stance identification based on how speakers take stance in utterances and what constructions in language they use to cue different stances.

Acknowledgments: This research is part of the StaViCTA project,[6] supported by the Swedish Research Council (framework grant the Digitized Society Past, Present, and Future, No. 2012-5659). We thank Tom Sköld for the data annotation.

Funding: Vetenskapsrådet, (Grant/Award Number: 'StaViCTA project, framework grant the Digitized So').

References

Aijmer, K. 1980. I think – An English modal particle. In T. Swan & O. J. Westvik (eds.), *Modality in Germanic languages: Historical and comparative perspectives*, 1–48. Berlin: Mouton de Gruyter.

Aikhenvald, A. Y. 2004. *Evidentiality*. Oxford: Oxford University Press.

Anand, P., M. Walker, R. Abbott, J. E. F. Tree, R. Bowmani & M. Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, 1–9. Association for Computational Linguistics.

Artstein, R. & M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596.

Baldwin, T., P. Cook, M. Lui, A. MacKinlay & L. Wang. 2013. How noisy social media text, how diffrnt social media sources? In *IJCNLP*, 356–364.

Bednarek, M. & H. Caple. 2012. *News discourse*, vol. 46. London: A&C Black.

Benveniste, E. 1958. Catégories de pensée et catégories de langue. *Les études philosophiques* 13(4). 419–429.

Berger, P., P. Hennig, M. Schoenberg & C. Meinel. 2015. Blog, forum or newspaper? Web genre detection using SVMs. In *2015 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 3, 64–68. IEEE.

Berman, R., H. Ragnarsdóttir & S. Strömqvist. 2002. Discourse stance: Written and spoken language. *Written Language & Literacy* 5(2). 253–287.

Biber, D. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5(2). 97–116.

6 StaViCTA project: http://cs.lnu.se/stavicta/

Boye, K. 2012. *Epistemic meaning: A crosslinguistic and functional-cognitive study*, vol. 43. Berlin: Walter de Gruyter.
Boye, K. & P. Harder. 2014. (Inter) subjectification in a functional theory of grammaticalization. *Acta Linguistica Hafniensia* 46(1). 7–24.
Chafe, W. L. & J. Nichols. 1986. *Evidentiality: The linguistic coding of epistemology*. Norwood, NJ: Ablex.
Conrad, S. & D. Biber. 2000. Adverbial marking of stance in speech and writing. In Geoff Thompson (ed.), *Evaluation in text: Authorial distance and the construction of discourse*, 56–73. Oxford: Oxford University Press.
Cornillie, B. 2007. *Evidentiality and epistemic modality in Spanish (semi-) auxiliaries: A cognitive-functional approach*, vol. 5. Berlin: Walter de Gruyter.
Croft, W. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Cambridge: Oxford University Press on Demand.
Croft, W. & D. A. Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.
Dancygier, B. & E. Sweetser (eds.). 2012. *Viewpoint in language: A multimodal perspective*. Cambridge: Cambridge University Press.
Dendale, P. & J. van der Auwera. 2001. *Les verbes modaux*, vol. 8. Amsterdam: Rodopi.
Du Bois, J. W. 2007. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction* 164. 139–182.
Ekberg, L. & C. Paradis. 2009. Evidentiality in language and cognition. *Special Issue of Functions of Language* 16(1). 5–7.
Englebretson, R. (ed.). 2007. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, vol. 164. Amsterdam: John Benjamins Publishing.
Esuli, A. & F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC* 6. 417–422.
Facchinetti, R., F. Palmer & M. Krug (eds.). 2003. *Modality in contemporary English*, vol. 44. Berlin: Walter de Gruyter.
Faulkner, A. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. *Science* 376(12). 86.
Fernández-Montraveta, A. & G. Vázquez. 2014. The SenSem Corpus: An annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory* 10(2). 273–288.
Ferreira, W. & A. Vlachos. 2016. Emergent: A novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Sheffield.
Fillmore, C. J., P. Kay & M. C. O'connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language* 64. 501–538.
Fuoli, M. 2012. Assessing social responsibility: A quantitative analysis of Appraisal in BP's and IKEA's social reports. *Discourse & Communication* 6(1). 55–81.
Fuoli, M. & C. Paradis. 2014. A model of trust-repair discourse. *Journal of Pragmatics* 74. 52–69.
Gärdenfors, P. 2014a. *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT Press.
Gärdenfors, P. 2014b. The evolution of sentential structure. *Humana. Mente–Journal of Philosophical Studies* 27. 79–97.
Gilquin, G. & S. T. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1–26.

Glynn, D. & M. Sjölin. 2015. Subjectivity and epistemicity: Corpus, discourse, and literary approaches to stance. *Lund Studies in English* 117. 360–410.
Goldberg, A. E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.
Goldberg, A. E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Granger, S. 2003. The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly* 37(3). 538–546.
Gray, B. & D. E. Biber. 2014. *Stance markers*. Cambridge: Cambridge University Press.
Hasan, K. S. & V. Ng. 2013a. Stance classification of ideological debates: Data, models, features, and constraints. In *IJCNLP*, 1348–1356.
Hasan, K. S. & V. Ng 2013b. Frame semantics for stance classification. In *CoNLL*, 124–132.
Hasan, K. S. & V. Ng. 2013c. Extra-linguistic constraints on stance recognition in ideological debates. In *ACL*, vol. 2, 816–821.
Hasan, K. S. & V. Ng 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *EMNLP*, 751–762.
Hommerberg, C. & C. Paradis. 2014. Constructing credibility through representations in the discourse of wine: Evidentiality, temporality and epistemic control. In D. Glynn & M. Sjölin (eds.), *Subjectivity and Epistemicity. Corpus, discourse, and literary approaches to stance. Lund Studies in English* 117. 211–238.
Hunston, S. 2008a. Collection strategies and design decisions. *Corpus linguistics: An international handbook* 1. 154–167.
Hunston, S. 2008b. Starting with the small words patterns, lexis and semantic sequences. *International journal of corpus linguistics* 13(3). 271–295.
Hunston, S. 2011. *Corpus approaches to evaluation: Phraseology and evaluative language*. Abingdon: Routledge.
Hunston, S. & G. Thompson (eds.). 2000. *Evaluation in text: Authorial stance and the construction of discourse: Authorial stance and the construction of discourse*. UK: Oxford University Press.
Hyland, K. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse studies* 7(2). 173–192.
Jones, S., M.L. Murphy, C. Paradis & C. Willners. 2012. *Antonyms in English: Construals, constructions and canonicity*. Cambridge: Cambridge University Press.
Kärkkäinen, E. 2003. *Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think*, vol. 115. Amsterdam: John Benjamins Publishing.
Kim, S. M. & E. Hovy 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, 1367. Association for Computational Linguistics.
Krug, M. G. 2000. *Emerging English modals: A corpus-based study of grammaticalization*, vol. 32. Berlin: Walter de Gruyter.
Kucher, K., A. Kerren, C. Paradis & M. Sahlgren. 2016a. Visual analysis of text annotations for stance classification with ALVA. In *EuroVis 2016, The 18th EG/VGTC conference on visualization*, 49–51. Eurographics-European Association for Computer Graphics.
Kucher, K., C. Paradis, M. Sahlgren & A. Kerren. 2017. Active learning and visual analytics for stance classification with ALVA. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7(3). 1–31.

Kucher, K., T. Schamp-Bjerede, A. Kerren, C. Paradis & M. Sahlgren. 2016b. Visual analysis of online social media to open up the investigation of stance phenomena. *Information Visualization* 15(2). 93–116.

Landis, J. R. & G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33. 159–174.

Langacker, R. W. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, vol. 1. Redwood City, CA: Stanford University Press.

Langacker, R. W. 1991. 8. Cognitive grammar. *Linguistic Theory and Grammatical Description: Nine Current Approaches* 75. 275.

Langacker, R. W. 2009. *Investigations in cognitive grammar*, vol. 42. Berlin: Walter de Gruyter.

Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.

Malmström, H. 2008. Knowledge-stating verbs and contexts of accountability in linguistic and literary academic discourse. *Nordic journal of English studies* 7(3). 35–60.

Marín-Arrese, J. I., M. Carretero, J. A. Hita & J. van der Auwera (eds.). 2014. *English modality: Core, periphery and evidentiality*, vol. 81. Berlin: Walter de Gruyter.

Martin, J. R. & P. R. White. 2003. *The language of evaluation*. London: Palgrave Macmillan.

Mohammad, S. M., P. Sobhani & S. Kiritchenko. 2016. Stance and sentiment in tweets. *arXiv preprint arXiv:1605.01655*.

Mushin, I. 2001. *Evidentiality and epistemological stance: Narrative retelling*, vol. 87. Amsterdam: John Benjamins Publishing.

Myers, C. & J. F. Hamilton. 2015. Open genre, new possibilities: Democratizing history via social media. *Rethinking History* 19(2). 222–234.

Myers, L. & Y.-H. Hsueh. 2001. Using nonparametric bootstrapping to assess kappa statistics. In *Annual meeting of the American Statistical Association*. American Statistical Association.

Nuyts, J. 2000. Tensions between discourse structure and conceptual semantics: The syntax of epistemic modal expressions. *Studies in Language* 24(1). 103–135.

Palmer, F. R. 1986. *2001. Mood and Modality*. Cambridge: Cambridge University Press.

Pang, B. & L. Lee 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271. Association for Computational Linguistics.

Pang, B. & L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2(1–2). 1–135.

Pang, B., L. Lee & S. Vaithyanathan 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, vol 10, 79–86. Association for Computational Linguistics.

Paradis, C. 2003. Between epistemic modality and degree: The case of really. *Topics in English Linguistics* 44. 191–222.

Paradis, C. 2009. "This beauty should drink well for 10–12 years": A note on recommendations as semantic middles. *Text & Talk-An Interdisciplinary Journal of Language, Discourse Communication Studies* 29(1). 53–73.

Paradis, C. 2015. Meanings of words: Theory and application. *Handbuch Wort und Wortschatz* 3. 274–293.

Paradis, C. & C. Willners. 2011. Antonymy: From convention to meaning-making. *Review of cognitive linguistics* 9(2). 367–391.

Patard, A. & F. Brisard (eds.). 2011. *Cognitive approaches to tense, aspect, and epistemic modality*, vol. 29. Amsterdam: John Benjamins Publishing.

Põldvere, N., M. Fuoli & C. Paradis. 2016. A study of dialogic expansion and contraction in spoken discourse using corpus and experimental techniques. *Corpora* 11(2). 191–225.

Precht, K. 2003. Stance moods in spoken English: Evidentiality and affect in British and American conversation. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-* 23(2). 239–258.

Rajadesingan, A. & H. Liu 2014. Identifying users with opposing opinions in Twitter debates. In *International conference on social computing, behavioral-cultural modeling, and prediction*, 153–160. Springer International Publishing.

Saurí, R. & J. Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language resources and evaluation* 43(3). 227–268.

Skeppstedt, M., M. Sahlgren, C. Paradis & A. Kerren. 2016. Unshared task: (dis)agreement in online debates. In *Proceedings of the 3rd workshop on argument mining (ArgMining '16) at ACL ' 16*, 154–159.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng & C. Potts 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 1631–1642.

Somasundaran, S. & J. Wiebe 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 116–124. Association for Computational Linguistics.

Sridhar, D., L. Getoor & M. Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 109–117. Baltimore, MD: ACL.

Taboada, M. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2. 325–347.

Talmy, L. 2000. *Toward a cognitive semantics*. Cambridge, MA: The MIT Press.

Tomasello, M. 2010. *Origins of human communication*. Cambridge, MA: MIT Press.

Traugott, E. C. & R. B. Dasher. 2002. *Regularities in semantic change*, vol. 97. Cambridge: Cambridge University Press.

Turney, P. D. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 417–424. Association for Computational Linguistics.

Usonienė, A. 2004. *Modalumas anglų ir lietuvių kalbose: Forma ir reikšmė*. Vilnius: Vilniaus Universiteto Leidykla.

van der Auwera, J. & V. A. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2. 79–124.

van Rijsbergen, C. J. 1979. Information retrieval. Dept. of computer science, University of Glasgow. citeseer. ist. psu. edu/vanrijsbergen79information. html.

Verhagen, A. 2005. *Constructions of intersubjectivity: Discourse, syntax, and cognition*, vol. 102. Oxford: Oxford University Press.

Walker, M. A., P. Anand, R. Abbott & R. Grant. 2012a. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 592–596. Association for Computational Linguistics.

Walker, M. A., J. E. F. Tree, P. Anand, R. Abbott & J. King. 2012b. A corpus for research on deliberation and debate. In *LREC*, 812–817.

Whitelaw, C., N. Garg & S. Argamon 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 625–631. ACM.

Wiebe, J., T. Wilson, R. Bruce, M. Bell & M. Martin. 2004. Learning subjective language. *Computational linguistics* 30(3). 277–308.

Wojatzki, M. & T. Zesch 2016. ltl. uni-due: Stance detection in social media using stacked classifiers.

Zappavigna, M. 2012. *Discourse of Twitter and social media: How we use language to create affiliation on the web*. London: A&C Black.

Zarrella, G. & A. Marsh. 2016. MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection. *arXiv preprint arXiv:1606.03784*.