DE GRUYTER OPEN

ICAME Journal, Volume 42, 2018, DOI: 10.1515/icame-2018-0007

Evaluating stance-annotated sentences from the Brexit Blog Corpus: A quantitative linguistic analysis

Vasiliki Simaki[1, 2], Carita Paradis[2] and Andreas Kerren[3]
Lancaster[1], Lund[2] and Linnaeus[3] Universities

Abstract

This paper offers a formally driven quantitative analysis of stance-annotated sentences in the Brexit Blog Corpus (BBC). Our goal is to identify features that determine the formal profiles of six stance categories (CONTRARIETY, HYPOTHETICALITY, NECESSITY, PREDICTION, SOURCE OF KNOWLEDGE and UNCERTAINTY) in a subset of the BBC. The study has two parts: firstly, it examines a large number of formal linguistic features, such as punctuation, words and grammatical categories that occur in the sentences in order to describe the specific characteristics of each category, and secondly, it compares characteristics in the entire data set in order to determine stance similarities in the data set. We show that among the six stance categories in the corpus, CONTRARIETY and NECESSITY are the most discriminative ones, with the former using longer sentences, more conjunctions, more repetitions and shorter forms than the sentences expressing other stances. NECESSITY has longer lexical forms but shorter sentences, which are syntactically more complex. We show that stance in our data set is expressed in sentences with around 21 words per sentence. The sentences consist mainly of alphabetical characters forming a varied vocabulary without special forms, such as digits or special characters.

1 Introduction

Stance-taking is grounded in the communicative situation and stance is the speaker's expression of his or her attitudes towards a specific topic, an event or an idea (Du Bois 2007). In Simaki et al. (2017c), we proposed a framework based on notional criteria of ten stance categories: AGREEMENT/DISAGREEMENT, CERTAINTY, CONTRARIETY, HYPOTHETICALITY, NECESSITY, PREDICTION, SOURCE OF KNOWLEDGE, TACT/RUDENESS, UNCERTAINTY, and VOLITION. The term notional refers their meanings and functions in discourse. We compiled a corpus of social

media text, and more specifically posts and comments from political blogs about the 2016 UK referendum, the Brexit Blog Corpus (BBC).[1]

We analysed the data with respect to how the speakers positioned themselves on the basis of the above notional categories, and annotated these data at sentence level. This work resulted in the stance-annotated BBC, which we here use for an exploratory study focusing on formal aspects of stance-taking in this study. By formal aspects, we mean the identification of frequent linguistic ways (e.g., punctuation, specific words, grammatical categories) that authors used in order to express a specific stance. As explained in Section 3, we make use of six of these notional categories annotated in BBC, namely CONTRARIETY, HYPOTHETICALITY, NECESSITY, PREDICTION, SOURCE OF KNOWLEDGE and UNCERTAINTY in order to investigate the possibility of approaching stance-taking from a radically different perspective, namely from the point of view of form, in order to determine whether it is at all possible to identify the stance categories based on purely formal grounds. We carry out simple basic counts at character, word and sentence level of the sentences that are expressive of the above stance types. Our main goal is to identify the corpus-based formal characteristics that are significant for each stance class in order to be able to use them for future computational investigations, such as classification experiments. We use the term *stanced* to designate the textual chunks where stance is detected and annotated in BBC. Our second goal is to determine the formal profile of stance-taking in our data set. Two research questions are at the heart of this study:

(1) Are the six stance categories different in terms of formal clues?

(2) What are the salient features of the stanced language of our data set?

The article is organised as follows. In Section 2, we present an overview of previous studies on speaker stance. In Section 3, we describe the methodology of the present study and the corpus used. In Section 4, we present the experimental procedure of this study. In Section 5, we evaluate and discuss the experimental results and the findings of the current study. Finally, Section 6 concludes this work.

2 Speaker stance and stance identification in social media

In this section, we give an overview of theoretical and computational studies of stance in order to present the different methodologies and perspectives of each discipline, both in the areas of speaker stance and automatic stance identification.

Speaker stance is firmly grounded in the speech situation and its participants. It is crucial for the social construction of meaning in different discourses. Stance may be seen as a psychological state involving speaker beliefs, evaluative ability and attitudes, and stance-taking is the performance by humans in communication – actions taken by speakers to express their beliefs, evaluation assessments and attitudes towards their interlocutors. Expressions of stance are the form-meaning pairings that are used to take stance. Stance has been studied from different perspectives under different communicative conditions. There are studies aiming at a more general understanding of the stance concept in human communication (Hunston and Thompson 2000; Berman et al. 2002; Du Bois 2007; Englebretson 2007), and there are others focusing on stance phenomena in specific text types or discourses (Hyland 2005; Biber 2006), or specific stance expressions (Conrad and Biber 2000; Hunston and Thomson 2000; Downing 2001; Kärkkäinen 2003; Paradis 2003; Gray and Biber 2014; Põldvere et al. 2016; Jiang 2017). Apart from these studies, there is work on similar, but more restricted concepts labelled as evidentiality (Precht 2003; Ekberg and Paradis 2009; Gu 2014), modality (Facchinetti et al. 2003; Kanté 2010), subjectivity/intersubjectivity (Benveniste 1971; White 2003; Verhagen 2005; Glynn and Sjölin 2015), evaluation/appraisal (Martin and White 2003; Read and Carrol, 2010), and sentiment (Wiebe et al. 2006; Van de Kauter et al. 2015; Taboada 2016). Discourse data from academic writing (Jiang 2017), political debates (Cabrejas-Peñuelas and Díez-Prados 2014), product reviews (Fuoli 2012), courtroom testimonies (Tracy 2011; Chaemsaithong 2012), public discourse (Bassiouney 2012; Paterson et al. 2015) and social media (Chiluwa and Ifukor 2015) have been studied in order to identify the speakers' attitudes when evaluating, opposing or supporting a topic/idea/event. In another study, Saurí and Pustejovsky (2009) attempted to detect event factuality as a marker of speakers' positioning to a specific topic, and they created the FactBank corpus, in which the different markers of event factuality were observed, and represented.

Simaki et al. (2017c) proposed a broad and comprehensive cognitive-functional framework for the analysis of speaker stance. In this framework, ten notional categories were identified, and BBC was annotated with these stances. The categories are described with definitions and examples in Table 1:

Table 1: The stance categories proposed in Simaki et al. (2017c), their definition and examples

Stance Category	Definition	Example
AGREEMENT/ DISAGREEMENT	the expression of a similar or different opinion	*Hmm, yes, I would also do the same.*
CERTAINTY	the expression of confidence as to what the speaker is saying	*Without a doubt, you will be there before 6 o'clock*
CONTRARIETY	the expression of a compromising or a contrastive/comparative opinion	*Despite the weather, I took him for a walk.*
HYPOTHETICALITY	the expression of a possible consequence of a condition	*If it's nice tomorrow, we will go.*
NECESSITY	the expression of a request, recommendation, instruction or an obligation	*You have to leave before noon.*
PREDICTION	the expression of a guess/conjecture about an event	*My guess is that the guests have already arrived*
SOURCE OF KNOWLEDGE	the expression of the origin of what he or she says	*According to the news, the rate of interest is not going up.*
TACT/RUDENESS	the expression of pleasantries and unpleasantries	*Please, do give my love to him.*
UNCERTAINTY	the expression of doubt as to the likelihood or truth of what she or he is saying	*There might be a few things left to do.*
VOLITION	the expression of wishes or refusals, inclinations of disinclinations	*If only I could remember his name.*

In Table 1, the notional stance categories are presented. The table is important for our work as six out of ten categories of this framework are used in this study.

Stance identification has also been the subject of several studies in Text Mining in order to take a step forward in Opinion Mining. Stance classification is connected to the fields of Subjective Language Identification (Wiebe et al. 2004), Opinion Mining and Sentiment Analysis (Pang and Lee 2008), where new information about the speaker's attitude in a given communicative situation is derived. These topics make use of similar methodologies as studies in Text Mining, but the purpose differs in each case. For instance, researchers in the field of automatic stance detection may investigate whether a speaker is *for* or *against* a topic/idea/event. The majority of these studies addresses the automatic stance identification as a binary issue of the for-or-against positioning of the speaker vis-à-vis a topic, idea or event. In many cases, the data used in these studies are extracted from online forum debates or other social media sources

such as blogs and Twitter. In most of these studies, the data are automatically annotated by the researchers according to the information in the title of the thread, e.g., *Supporting the woman's right to abortion, All together against guns, Abolish the death penalty now!,* or to indices like the hashtags that are mostly used in Twitter but in other networks too, i.e., *#not, #pro, #pride*.

Table 2 gives a summary of studies in stance classification, the data used in each one of them, and the best classification accuracy they achieved in their classification experiments.

Table 2: Studies in stance classification and their performance in chronological order

Authors	**Data**	**Classification accuracy**
Somasundaran and Wiebe (2010)	4-topic ideological online debates	63.93%
Anand et al. (2011)	14-topic two-sided debates	69.00%
Walker et al. (2012b)	14-topic two-sided debates	75.00%
Walker et al. (2012a)	Internet Argument Corpus (Walker et al. 2012c)	88.00%
Hasan and Ng (2013a)	4 different data sets	75.00%
Hasan and Ng (2013b)	4 different data sets	75.90%
Hasan and Ng (2013c)	4 different data sets	75.40%
Hasan and Ng (2014)	4 different data sets of ideological debates	69.00%
Faulkner (2014)	International Corpus for Learner English (ICLE: Granger 2003)	82.00%
Ferreira and Vlachos (2016)	Emergent data set	73.00%
Mohammad et al. (2016)	Twitter data set	69.00%

All the studies in Table 2 implement stance classification methods using machine learning techniques and data extracted from social media sources from 2010 to 2016. The column on the right hand side presents the best result in terms of classification accuracy achieved in each study. More precisely, this value shows the percentage of the correctly classified texts to the corresponding stance category depending on the classification features and algorithms used in each case. The studies included in Table 2 show the research interest of the Text Mining community in the stance classification task. The high classification accuracies prove the efficacy of the classification algorithms and the features used to

resolve this problem. Some of these methods and features were used in the present study. While most studies address stance classification as a binary issue (*for/against*), Persing and Ng (2016) annotated student essays with six stance values (*Agree Strongly, Agree Somewhat, Neutral, Disagree Somewhat,* and *Disagree Strongly*), and they proposed two sets of novel, stance-taking, path-related features and knowledge-based features. The implementation of their feature sets with n-grams and Faulkner's (2014) features outperformed previous baselines, and reduced the stance identification error to 11.3% and 5.3% (micro and macro F-score respectively).

In addition to the studies listed in Table 2, many researchers have worked on stance identification without following a classification methodology. For instance, Sridar et al. (2014) investigated the performance of linguistic and relational features in a subset of the Internet Argument Corpus (Walker et al. 2012c), which is a text collection based on online debates on various topics. They showed the significance of features that reflect more complex interactions among writers, and between writers and their posts. In a study of social media text, Rajadesingan and Liu (2014) identified different types of stance-taking, for or against a topic, in a collection of tweets from more than 100,000 different Twitter users, using their ReLP (Retweet-based Label Propagation) framework. From a different perspective, Kucher et al. (2016b) created the uVSAT tool for visual stance analysis to support interactive exploration of time-series data associated with online social media documents. The uVSAT tool contains multiple approaches for analysing text data and identifying stance markers in order to prepare a stance-oriented training data set.

3 Methodology and data description

3.1 Methodology

In the present study, a formally driven linguistic analysis of stanced sentences from blog sources is performed. The first goal is to derive linguistic clues of differentiation among the sentences that are annotated with different stances in order to determine the formal profile for each stance. To this end, we need to understand the nature of this text type, and its characteristics. In social media text, and more specifically in blog text, features of informal language are common, e.g., special characters, emoticons, and expressions or structures that are not frequent in formal written discourse. Another characteristic is the use of dialogic interaction that makes social media text an informal *written oral* discourse type (Simaki 2015). Similar features been used the analysis of social media text for a variety of purposes, such as authorship attribution (e.g., Stamatatos et al.

2000, 2001; Stamatatos 2009), genre detection (e.g., Kessler et al. 1997), topic or trend detection (e.g., Cataldi et al. 2010; Mathioudakis and Koudas 2010), gender, age and/or personality identification (e.g., Mukherjee and Liu 2010; Peersman et al. 2011; Nguyen et al. 2013; Simaki et al. 2015a, 2015b, 2017a), author's profiling (e.g., Schwartz et al. 2013), opinion mining and sentiment analysis (e.g., Pang and Lee 2008; Pak and Paroubek 2010).

The second goal is to highlight the similarities – in terms of linguistic characteristics – among all sentences and create a general profile of the BBC subset. For that purpose, a manually annotated text collection with ten core stance categories (Simaki et al. 2017c) extracted from political blogs was used. The six most frequent categories, out of the ten in the full data set, were selected and explored. Formal characteristics at character, word and sentence level were derived, and their appearance and frequency in the data set were examined and quantified. In Table 3, we present the characteristics investigated in our study. Similar features were used in studies that implement various text classification tasks, such as genre, gender and age identification, authorship attribution and other aspects (Zheng et al. 2006; Simaki et al. 2015a, 2015). They proved to be important clues for the purpose of this study. We used simple linguistic features for this first analytical attempt at a formal investigation of these stance categories. After evaluating the corpus at character, lexical and sentence level, we can continue with analyses that are more refined.

Table 3: The linguistic characteristics used in the present study

Linguistic features
Character level
Number of special characters/ total number of characters[2]
Number of punctuation symbols/ total number of characters[3]
Number of spaces/ total number of characters
Number of upper case characters/ total number of characters
Number of alphabetical characters/ total number of characters
Number of digit characters/ total number of characters
Average sentence length in terms of characters
Lexical level
Average word length
Number of short words (less than four characters)/ total number of characters
Average sentence length in terms of words

Number of different words/ total number of words[4]	
Hapax legomena/ total number of words[5]	
Hapax dislegomena/ total number of words[6]	
Sentence level	
Comma frequency	
Full stop frequency	
Exclamation mark frequency	
Colon frequency	
Semicolon frequency	
Quotation mark frequency	
Frequency of 10 different POS tags (10 features)[7]	

In Table 3, the features are grouped into three linguistic levels. The rationale for this grouping is not based on the type of the metric that is used for the calculation of the feature, but is based on the information that each feature provides at different levels of linguistic analysis. More specifically, the character level features inform us about basic metrics at character level such as the length of the sentence in terms of characters, the frequency of alphabetical letters, or digits in the data. These items are not *per se* informative entities from which we can derive safe conclusions about language patterns in our data set. However, the features in this group support further conclusions about the lexical or sentence level of analysis when combined with findings from the other two feature groups. For instance, the number of the spaces provides insights about the speaker's lexical choices when combined with the short-word and the word-length parameters. The lexical level features link to a deeper level of linguistic analysis, and provide useful insights about the lexical choices that people make when expressing different stances, and the vocabulary variation of the stanced discourse. Finally, the third feature group consists of metrics that support the conclusions related to the syntactic structure of stanced discourse. In this category, we have included even simple character-based metrics, such as the frequency of various punctuation marks, which is informative in the light of how a sentence is structured, how syntactically complex it is, whether subclauses are part of the sentence, and its orientation regarding sentence type (declarative, exclamation, etc.). Also, the calculations of the frequency of the grammatical categories support conclusions about syntactic patterns that can be detected in the sentences. There are 29 features in total on which we perform statistical

tests. The statistical findings are evaluated and discussed in terms of the formal profile of the stanced language in our data set and the profiles of its individual stance categories.

3.2 Data description

In the present study, we use a subset of BBC (the rationale of the stance framework described in Section 2). The data were manually annotated according to the total semantic information of the sentence (for the annotation protocol that was followed, see Appendix 1) – each one of them was considered as a whole construction – and the annotation agreement results were tested and evaluated (see Appendix 2). In many cases, more than one of these notional categories could be identified in the same sentence, and in that case, the annotators attributed them to the sentence. BBC consists of 1,682 annotated sentences. In Table 4, we present the distribution of these sentences in relation to the stance categories, and the average number of words per sentence for each stance.

Table 4: The ten stance categories, the number of sentences/stance, and the average word number/sentence for each stance in BBC

Stance category	**Number of sentences**	**Average number of words/sentence**
CONTRARIETY	352	23.46
SOURCE OF KNOWLEDGE	287	22.76
PREDICTION	252	19.57
NECESSITY	204	18.22
UNCERTAINTY	196	21.48
HYPOTHETICALITY	171	22.07
CERTAINTY	84	20.17
AGREEMENT/DISAGREEMENT	50	19.06
TACT/RUDENESS	44	16.72
VOLITION	42	17.71
Total:	1,682	21.12

Table 4 shows the ten stance categories in BBC from the most frequent stance type to the least frequent one, the total number of sentences annotated, and the

average number of words per sentence for each stance category. The data set used in this study focuses on six categories. They are CONTRARIETY, SOURCE OF KNOWLEDGE, PREDICTION, NECESSITY, UNCERTAINTY, and HYPOTHETICALITY. These categories were the most frequently used categories in BBC. They were also the ones with the highest scores of inter-annotator agreement, which means that the annotators agreed on the annotation decision to a high degree (see Appendix 2). These categories constitute a data set of 1,462 sentences (31,331 words; 150,190 characters). The data were retrieved from June to August 2015 using the Gavagai API.[8] The blog post texts were detected using seed words such as *Brexit*, *EU referendum, pro-Europe, europhiles, eurosceptics, United States of Europe*, *David Cameron, Downing Street.* We created a list of about 50 seed words based on our judgement about the upcoming referendum and its key figures, and we then searched for the URLs referring to any of them. The URLs were retrieved and filtered so that only links ending in wordpress.com, blogger.com, blogspot.* or similar sources from http://www.lobbyplanet.eu/links/eu-blog were selected. The texts were segmented into sentences, and sentences in quotes were excluded from these data because our focus is on stance-taking by the speakers only (not reported stance-taking). Next, the data set that was subsequently annotated consisted of randomly selected sentences. We did not keep any author-, time-, source-, context-related information in the final data set for the annotators. The basic idea was to annotate the sentences based on the semantic information that the sentences themselves provided (in terms of stance-taking) without metadata information. Questions were not included in the corpus, which was a decision made at a previous stage of our research, when we set the requirements for the data collection and processing of BBC. Our scope was restricted to affirmation as the main expression type for speaker stance.

4 Linguistic analysis of stance-annotated data

In this section, we present the occurrences of our data. We used the NLTK[9] toolkit for the estimation of the features (feature extraction process). In Table 5, we show the mean values of the features for each stance category. All our features (as shown in Table 3) are normalised in order to have values in a scale from 0 to 1. Table 5 presents the mean value of each feature in the sentences of each stance category, and henceforth we use the feature names as presented in the first column of the table.

Table 5: The mean values for the linguistic clues in each category

Stance categories / Linguistic features	CONTRARIETY	HYPOTHETICALITY	NECESSITY	PREDICTION	SOURCE OF KNOWLEDGE	UNCERTAINTY
Special characters	0.002	0.001	0.002	0.001	0.001	0.001
Punctuation	0.020	0.021	0.025	0.020	0.019	0.019
Spaces	0.167	0.171	0.163	0.166	0.163	0.166
Upper case characters	0.026	0.026	0.035	0.033	0.030	0.028
Alphabetical characters	0.810	0.808	0.810	0.812	0.814	0.814
Digit characters	0.004	0.002	0.002	0.003	0.005	0.002
Short words	0.433	0.437	0.419	0.413	0.403	0.417
Average word length	0.040	0.041	0.061	0.050	0.041	0.044
Average sentence length/ characters	0.479	0.500	0.417	0.462	0.503	0.463
Average sentence length/ words	0.554	0.536	0.433	0.498	0.540	0.507
Different words	0.910	0.920	0.936	0.932	0.932	0.931
Hapax legomena	0.835	0.852	0.883	0.874	0.876	0.874
Hapax dislegomena	0.062	0.057	0.045	0.050	0.045	0.047
Comma frequency	0.383	0.343	0.248	0.263	0.320	0.280
Full stop frequency	0.510	0.539	0.643	0.624	0.556	0.597
Exclamation mark frequency	0.006	0.005	0.012	0.010	0.011	0.010
Colon frequency	0.008	0.005	0.008	0.006	0.009	0.002
Semicolon frequency	0.010	0.003	0.011	0.008	0.009	0.019
Quotation mark frequency	0.080	0.101	0.075	0.081	0.091	0.089
Noun frequency	0.248	0.233	0.249	0.247	0.273	0.234
Pronoun frequency	0.062	0.072	0.075	0.058	0.053	0.065
Adjective frequency	0.083	0.071	0.074	0.089	0.084	0.091
Verb frequency	0.169	0.181	0.191	0.162	0.181	0.175

Adverb frequency	0.079	0.061	0.060	0.072	0.062	0.074
Preposition frequency	0.146	0.152	0.149	0.139	0.159	0.131
Conjunction frequency	0.046	0.028	0.025	0.032	0.023	0.031
Interjection frequency	0.000	0.000	0.000	0.000	0.000	0.000
Determiner frequency	0.109	0.119	0.111	0.115	0.109	0.117
Particle frequency	0.003	0.006	0.004	0.004	0.003	0.003

We performed a two-step statistical analysis in order to identify the significant difference for each feature among the six stance categories. In a first step, the one-way Analysis of Variance (ANOVA) was used. The one-way ANOVA determines whether there are significant differences between the means of three or more (in the present study six) independent groups. More specifically, we wanted to test the means of 29 features for the six stance categories. We tested the null hypothesis (H_0), and if it was rejected there were at least two group means that were significantly different from each other.

The ANOVA F-value is estimated with the commonly used a=0.05 (5% significance level, 95% confidence interval). The corresponding p-value was also estimated, and when $p<0.05$ there is a significant difference. In Tables 6 and 7, we show the results for the linguistic features whose results confirmed the null hypothesis (Table 6) and the features that rejected H_0 (Table 7).

Table 6: The features with no significant difference among the stance categories

Linguistic features	**F-value**	**P-value**
Exclamation mark frequency	0.306	0.909
Colon frequency	0.402	0.847
Quotation mark frequency	0.518	0.762
Particle frequency	0.846	0.516
Determiner frequency	1.063	0.378
Semicolon frequency	1.140	0.336
Special characters	1.227	0.293
Interjection frequency	1.341	0.244
Upper case characters	1.568	0.165
Alphabetical characters	1.642	0.145

Table 6 presents the features that confirmed H_0, i.e., their p-value is greater than 0.05, and the F-value is smaller than 2.220, which is the critical value (F_{crit}) of the ANOVA test (F_{critA}) according to the confidence interval used in this case. We observe that ten out of 29 features confirm the null hypothesis. In this case, no significant differences are observed among the group means, which means that these features are similar across the six stance categories. If we look closer at the results of Table 6, we see that three out of seven character features and seven out of 16 sentence features are the same among the six stances, and that none of the lexical features confirmed the H_0.

Table 7: The features with significant difference among the stance categories

Linguistic features	**F-value**	**P-value**
Average word length	20.438	0.000
Conjunction frequency	15.612	0.000
Average sentence length/words	10.600	0.000
Comma frequency	9.422	0.000
Full stop frequency	7.671	0.000
Hapax legomena	6.750	0.000
Different words	6.633	0.000
Average sentence length/characters	5.647	0.000
Punctuation	5.466	0.000
Hapax dislegomena	5.504	0.000
Preposition frequency	5.199	0.000
Noun frequency	5.165	0.000
Spaces	4.883	0.000
Verb frequency	4.538	0.000
Adverb frequency	4.331	0.000
Pronoun frequency	4.272	0.000
Short words	3.413	0.004
Adjective frequency	3.085	0.008
Digit characters	2.876	0.013

Table 7 presents the 19 out of 29 features that have rejected the H_0. Their p-value is smaller than 0.05 and the F-value is greater than 2.220, which is the F_{crit} according to the confidence interval used in this case. We see that all word-

based characteristics proved to be clues of differentiation among the six stance categories. From the other two groups, four character and nine sentence level features have different group means. Our first important finding is that all word-based features (six out of six) are significantly different, as they rejected the H_0.

The findings in Table 7 presents information about the exploration of the whole set of formal features which may or may not be significant for the identification of different stances. Although many of these features are clues that may not make any sense one by one, a key question for this exploratory study is whether there is useful 'hidden' information that becomes evident when they are combined with other characteristics. Table 7 features three main aspects of our data set: the length of words and sentences, the syntactic structure of the sentences, and the lexical variation in the sentences. The combination of the average word length, average sentence length/characters, average sentence length/words, spaces and short words shows that the length of the sentences and the forms in the data set are important factors for the differentiation of the sentences expressing different stances. This means that our findings with regard to sentence length and word length as clues of differentiation in stanced sentences are motivated by more than one feature. For instance, if we combine the spaces feature with short words and average sentence length/characters features, we see that their highest values are observed for the sentences that express HYPOTHETICALITY; see Table 5. The features related to the syntactic structure of stanced sentences are the combination of the frequencies of conjunctions, commas, full stops, punctuation, prepositions, pronouns, and verbs. For instance, the comma feature can be combined with the conjunctions features, and their highest values are observed for CONTRARIETY. This highlights a frequent pattern observed in contrastive sentences, namely the use of comma before or after a contrastive form, e.g., '*but'*, '*however,'*, etc. Furthermore, if we combine the highest frequencies of verbs and pronouns, we find them for expressions of NECESSITY. The NECESSITY sentences feature a salient syntactic pattern of *pronoun + must/need/have to* in sentences expressing recommendations and instructions, i.e., what is necessary to do. Finally, the features that inform us about the lexical choices that speakers expressing different stances have made are the frequencies of hapax legomena, different words, hapax dislegomena, nouns, adverbs, and adjectives. These characteristics provide formal insights not only about salient grammatical categories in the data of each stance category, but in some cases, they are also indications of lexical variation. For instance, in the case of CONTRARIETY, different words and hapax legomena show the lowest values among all other stances, while the highest frequency of hapax dislegomena (forms repeated twice in the sentence) is observed. This highlights that in this

stance category, the vocabulary is quite limited in terms of different forms variation, and speakers tend to repeat the same forms in their sentences.

Another interesting observation is that the punctuation feature (that measures the frequency of all punctuation marks) is among the discriminatory features, while only the comma and full stop features turned out to be discriminatory too, and none of the other punctuation marks are significant. This may be due to the high frequency of commas and full stops in the data, which influences the ratio of the punctuation marks feature, resulting in its significance. The rest of the punctuation marks show relatively low frequencies in our data, which means that they appear rarely in the corpus. A more detailed interpretation of the feature combinations and the information that can be extracted from their values are presented in Section 5.

The one-way ANOVA was the first step in our study. In order to determine which specific groups differed from each other, we performed the Scheffé *post-hoc* test (Scheffé 1959). The Scheffé test is a conservative single-step multiple comparison procedure, which applies to the set of estimates of all possible contrasts among the factor level means, not just the pairwise differences that are considered by the Tukey-Kramer method (Tukey 1949). In this test, the group means and the ANOVA findings were used in order to see if there were differences or not among the combinations of two groups. In our case, the combinations were 15. To find the F-value of the Scheffé test (F_S), we compared all groups two by two. We calculated the F_{crit} value of the Scheffé test (F_{critS}), which compared to the F_S tells us if there is a significant difference between the means of two groups. This value is equal to 11.101 in our case. If $F_{S>}F_{critS}$ then there is a significant difference between the means of the two categories. In Table 8, we show the Scheffé post-hoc test results for the 19 features that rejected the ANOVA's null hypothesis for the 15 paired combinations of the six stance categories. The values highlighted in yellow are the ones that proved to be significant and the values in orange are the ones that are not significant, but their value is close to the Scheffé critical value.

Table 8: The Scheffé post-hoc test results. In yellow the significant values and in orange the values that are close to the Scheffé critical value

Stance A	Stance B	Linguistic features	Average word length	Conjunction frequency	Average sentence length/words	Comma frequency	Full stop frequency	Hapax legomena	Different words	Average sentence length/chars	Punctuation/chars	Hapax dislegomena	Preposition frequency	Noun frequency	Spaces/chars	Verb frequency	Adverb frequency	Pronoun frequency	Short words	Adjective frequency	Digits/chars
CONTRARIETY	HYPOTHETICALITY		0,18	28,0	0,9	2.4	1.0	2.2	2,5	1,3	0.4	1.1	0.9	2.5	4.0	2.8	9.5	3.2	0.1	3.6	3.6
CONTRARIETY	NECESSITY		79,9	42,8	44,7	30.9	25.4	20.6	20,6	12,7	16.3	15.0	0.1	0.0	5.4	11.7	11.5	5.8	2.0	2.2	3.4
CONTRARIETY	PREDICTION		18,8	21,8	10,8	27.5	21.4	15.5	16,7	1,0	0.1	8.1	2.0	0.0	0.9	1.3	1.6	0.6	4.9	1.5	0.
CONTRARIETY	SOURCE OF KNOWLEDGE		0,3	61,5	0,7	8.1	3.8	18.8	17,6	2,5	0.4	17.8	6.0	10.1	8.5	3.8	11.7	3.3	11.7	0.0	1.4
CONTRARIETY	UNCERTAINTY		3,0	20,8	6,6	17.6	10.5	13.5	13,1	0,8	0.7	11.4	6.3	2.4	0.7	0.8	0.6	0.2	2.8	2.4	2.7
HYPOTHETICALITY	NECESSITY		51,9	0,6	23,2	11.0	11.1	6.2	5,8	16,5	7.9	5.5	0.2	2.4	14.3	1.9	0.0	0.1	2.4	0.1	0.0
HYPOTHETICALITY	PREDICTION		10,3	1,1	3,4	8.4	8.2	3.5	3,6	3,6	0.9	1.9	4.3	2.1	7.3	6.5	3.3	5.5	4.8	7.9	1.0
HYPOTHETICALITY	SOURCE OF KNOWLEDGE		0,0	1,8	0,0	0.7	0.3	4.5	3,6	0,0	1.4	6.1	1.2	17.4	18.9	0.0	0.0	10.5	10.1	4.0	8.0
HYPOTHETICALITY	UNCERTAINTY		1,2	0,6	1,8	4.8	3.3	3.2	2,7	3,1	1.8	3.7	9.0	0.0	6.4	0.4	4.2	1.3	3.1	9.1	0.0
NECESSITY	PREDICTION		20,7	4,0	11,3	0.3	0.4	0.6	0,4	6,0	16.8	1.2	2.7	0.0	1.7	17.8	4.1	8.7	0.3	6.1	0.9
NECESSITY	SOURCE OF KNOWLEDGE		65,3	0,2	32,0	8.1	9.8	0.3	0,5	23,1	19.7	0.0	2.9	6.9	0.0	2.5	0.0	15.4	2.5	2.6	8.0
NECESSITY	UNCERTAINTY		39,7	2,8	12,8	1.3	2.3	0.5	0,5	5,5	18.8	0.1	6.9	2.3	1.6	4.7	5.1	2.7	0.0	7.3	0.0
PREDICTION	SOURCE OF KNOWLEDGE		13,0	7,6	5,4	5.6	6.8	0.0	0,0	5,8	0.0	1.3	13.1	8.8	3.0	8.5	3.7	0.8	1.0	0.9	3.9
PREDICTION	UNCERTAINTY		4,4	0,0	0,1	0.3	0.9	0.0	0,0	0,0	0.2	0.4	1.2	2.0	0.0	3.6	0.1	1.3	0.1	0.1	0.5
SOURCE OF KNOWLEDGE	UNCERTAINTY		1,4	5,5	2,9	2.4	2.1	0.0	0,0	4,9	0.0	0,1	20.5	18.0	2.7	0.5	4.6	4.3	1.7	1.7	6.9

In Table 8, we present the F_S for all 19 distinctive features in each one of the 15 stance categories combinations. In the following section, we analyse these results.

5 Discussion

In this study, we explored the potential of approaching stance from a formal angle in order to be able to get a better grasp of stance and stance-taking in text. We explored the language of stanced sentences in a subset of BBC. Our hypothesis that 'it is possible to detect distinctive linguistic clues where different stances are employed in social media text' was confirmed. We showed that writers' formal choices, even at character-level, are informative clues for the identification of stance-taking in discourse. After retrieving standard formal features from our corpus and performing statistical analyses, we are in a position of making the following statements:

(1) There are linguistic characteristics (10 in total) that are common to all stanced sentences in our data set.

(2) We observed that 19 out of 29 linguistic characteristics are clues of differentiation among the six stance categories.

(3) There are features that are significant in more stance combinations than other characteristics, and there are stance categories where stance can be identified in a more prominent way (in terms of linguistic choices), resulting in a more distinct formal profile.

Concerning the first statement, we observed that the ten linguistic characteristics that are common in stanced sentences in the BBC subset involve the less commonly used punctuation (use of exclamation, and quotation marks, semicolon, colon), special characters, alphabetical characters, the use of upper case characters, the use of interjections, particles and determiners. The distribution of these characteristics among all six stance categories points to the linguistic profile of stanced discourse in the corpus. In this data set, the writers do not make much use of punctuation characters other than full stops and commas (structural punctuation characters). Another interesting observation is that the use of special characters is also low. The stanced sentences of our data are mostly affirmative, and writers make most often use of alphabetical characters (the digit characters feature is a clue of differentiation, but it does not occur often in in the corpus), without using emoticons (which are basically combinations of punctuation marks and special characters sequences) or other non-linguistic clues to express

their stance. The low number of upper-case characters suggests that not many proper names or acronyms are used in the sentences, even though we expected to see some entities thematically related to the political topic discussed, e.g., country names and politicians' names. The use of particles such as the infinitive *to* is also low, and the use of determiners is balanced among all six stance categories.

The existence of common formal features among all stance categories enables us to construct the profile of the stanced sentences in the BBC subset. We showed that bloggers and blog commentators mainly use standard means of expression (lexical items) in stanced expressions relating to the six categories we tested. In contrast to other social media texts, they avoid using special characters, interjections, digits, or less commonly used punctuation in contrast, for instance to Twitter users, who make frequent use of such elements (Park et al. 2014). In future studies, it will be an interesting point to evaluate if this observation only applies to the stanced sentences of BBC, or if it can be generalised to stanced sentences in other corpora. Our findings regarding the non-discriminatory features can only be understood as indications of a pattern in these data. In any case, it seems reasonable to assume that the writers of BBC's sentences explicitly express their opinions with words, without making any non-linguistic forms. If we also take a look at the features that appear to be significantly different among the stances, we see that the length of the stanced sentences in our data set varies, and a quick look at the average sentence length in terms of word number per stance category in the data shows a mean of 21 words per sentence. In order for us to characterise the sentences of BBC as short or long, various sources were used where information about the optimal sentence length in terms of readability and ease of understanding is provided. In the guidelines on the official British e-government platform,[10] the recommended length is up to 25 words per sentence, where in the Plain English Campaign[11] the optimal sentence length is between 15 and 20 words. The writers use many short lexical items (high frequency of short words) and their vocabulary is varied (high number of hapax legomena and different words). In a further step, calculation of stop words, i.e. very common words such as *the, has, a,* would shed more light on the lexical structuring of the sentences of different stances.

Concerning the second statement, we detected 19 features of discriminatory linguistic choices in stance-annotated discourse in the BBC subset among the six stance categories. The majority of these features are word-based, which means that writers make different lexical choices according to the different stances they take. Among the sentences annotated with different stances, we observe significant differences in terms of sentence length, word length, lexical

variation, use of commas, digits, spaces, and full stops. Some of these features can be grouped together as characteristics related to sentence length, form and syntactic structure (full stops, punctuation, commas, spaces, sentence length in terms of words and characters, digits), and to lexical forms (word length, short words, different words, hapax legomena, and dislegomena). This observation about the second group of distinctive linguistic clues calls for a follow-up semantic analysis in order to identify the semantic choices in different stances.

Regarding the third statement, we evaluated the significant features and their values among the different stance combinations. In 12 out of 15 stance category combinations, we show that there is a statistically significant difference at least on at least one linguistic parameter. The most frequent features that are clues of differentiation are word length, frequency of conjunctions, and sentence length in terms of words, which are statistically important in five to seven stance category combinations. We note that these three discriminatory characteristics differ in the data for the different stance categories. It is more frequent for expressions that involve CONTRARIETY or NECESSITY. The reason for this may be that the values of word length, conjunction frequency and sentence length in terms of words in these two stance categories show the highest deviation from one another. The average word length in CONTRARIETY shows the lowest value and in NECESSITY the highest value (0.040 and 0.61 respectively), whereas the average sentence length in CONTRARIETY shows the highest values and in NECESSITY the lowest one (0.55 and 0.43 respectively). From another point of view, CONTRARIETY sentences show the second highest values of spaces and short words (0.167 and 0.433), while NECESSITY sentences appear to have lower values in these two features (0.163 in spaces, and 0.419 in short words). The sentences annotated for CONTRARIETY contain more and shorter words than the sentences annotated for NECESSITY. The frequency of conjunctions, finally, shows the highest value in CONTRARIETY and the lowest in NECESSITY among all other stances (0.046 for CONTRARIETY and 0.025 for NECESSITY).

Hapax legomena and the different words (in terms of different types) are discriminative clues for four stance combinations, and they appear most often where CONTRARIETY combines with another stance. These two features have the lowest means in this stance category (0.835 for the hapax legomena and 0.910 for the different words) and we observe that for CONTRARIETY there is limited lexical variation, and a less varied vocabulary as can be seen in Table 4. Punctuation is also a clue to differentiation in four combinations. In all of them NECESSITY has the highest mean value (0.025) among all stance categories. In three stance category combinations involving CONTRARIETY, full stops, commas, and hapax dislegomena appear to be important features of differentiation. Full

stops for CONTRARIETY have the lowest mean (0.510) among all the categories and commas and hapax dislegomena the highest ones (0.383 and 0.062 accordingly).

Besides simple linguistic clues of differentiation at character, lexical and sentence level, we demonstrated that seven sentence features, labelled syntactic features, are discriminative across stances. Conjunctions appeared to be one of them, and CONTRARIETY sentences show the highest frequency of conjunctions (0.046) in relation to all the other five stances. This is due to contrastive forms (*apart from, but*) used when contrariness is adopted by the speaker, as shown in Examples (1)–(2).

(1) *FeuroEveryone can see that - apart from the SNP.*

(2) *NZDUSD and AUDUSD increased as well, but the bearish are fighting back in these pairs.*

CONTRARIETY also differs from NECESSITY, SOURCE OF KNOWLEDGE and HYPOTHETICALITY, in terms of adverb frequency use. The former has the highest mean value (0.079) and the latter three the lowest mean values (0.060 – 0.62).The frequent use of prepositions and nouns is a characteristic of sentences annotated as SOURCE OF KNOWLEDGE. This stance category demonstrates the highest mean value of preposition use (0.159) and shows a significant difference from PREDICTION and UNCERTAINTY that show the lowest values (0.139 and 0.131 respectively). Concerning the noun frequency, we have a similar picture for SOURCE OF KNOWLEDGE showing the highest mean value (0.273), while HYPOTHETICALITY and UNCERTAINTY show the lowest values (0.233 and 0.234 respectively). The frequency of verbs is a clue of differentiation in sentences where NECESSITY (0.191) is compared with CONTRARIETY (0.169) and with PREDICTION (0.162). The high number of verbs and pronouns per sentence is a characteristic of NECESSITY, and pronoun frequency appears to be a discriminative clue between NECESSITY (0.075) and SOURCE OF KNOWLEDGE (0.053), which are the two categories where the two extreme values of these features are observed. The adjective feature appears to be among the least useful clue of stance differentiation in terms of the ANOVA results. It did not have any significant differences in the Scheffé test.

We also show that in three out of 15 stance combinations the features do not have any significant differences. The three stance combinations that do not exhibit any discriminatory characteristics are HYPOTHETICALITY – PREDICTION, HYPOTHETICALITY – UNCERTAINTY, and PREDICTION – UNCERTAINTY. Our findings so far do not reveal any indications regarding the relation of the two first pairs. Further analyses of the lexical forms used in HYPOTHETICALITY, and comparison with

the lexical forms used in the other two stances, may point to constructions that HYPOTHETICALITY shares with PREDICTION and UNCERTAINTY. The lack of significant differences for the pair PREDICTION – UNCERTAINTY may have different reasons. One reason may be that UNCERTAINTY and PREDICTION are difficult notions to identify and annotate (low inter-annotator agreement scores, 0.57 for the PREDICTION and 0.62 for the UNCERTAINTY). This is a likely explanation because PREDICTION always involves some level of UNCERTAINTY since it is about the future, and hence something that is uncertain. Another explanation based on our previous study of sentence-level stance annotation of the BBC (Simaki et al. 2017c) is that the stance categories whose combinations have similar feature mean values and co-occur frequently in BBC's sentences share common constructions and forms in many cases, more precisely the combinations HYPOTHETICALITY – PREDICTION and PREDICTION – UNCERTAINTY. In 86 sentences, the annotators attributed both PREDICTION and UNCERTAINTY as annotation labels, and in 23 sentences, they attributed both HYPOTHETICALITY and PREDICTION.

With respect to the combinations that have different feature values when combined with each other, we observe that CONTRARIETY is a highly discriminative category compared to the other stances. In all five possible combinations, CONTRARIETY is significantly different in terms of five out of 12 features. In most combinations, we find three common characteristics that are important: the conjunctions, the hapax legomena and the ratio of different words. The last two features have the lowest mean values in this category among all other five stances. Generally speaking, we see that CONTRARIETY uses more repeated forms. If we take into account the word and sentence length that appears to be among the significant features in some of the combinations, the lexical forms used in sentences annotated as CONTRARIETY are shorter, and they are found in longer sentences.

NECESSITY is the second most discriminative stance category, which, like CONTRARIETY, differs from the other stance in all five combinations. In these combinations, significant differences are observed in four to 12 features, and two of them are clues of differentiation in all NECESSITY comparisons to other stances: word length and sentence length in terms of words. The mean value of the first feature is the highest (0.061) among all other stance categories, and the mean value of the latter one is the lowest (0.43) one in the NECESSITY case. The high frequency of verbs in the NECESSITY sentences is an indication that there is a tendency of using subordinate clauses, which can be supported also by the high frequency of pronouns. When writers express NECESSITY, they use longer lexical forms but shorter sentences that are syntactically more complex. The punctuation feature also appears to be important in four out of five of the NECESSITY com-

binations, as in this category it shows the highest value (0.025) among all other stances. Another observation is that SOURCE OF KNOWLEDGE is the second longest stance category after CONTRARIETY in terms of both word and sentence length.

In Figure 1, we provide an overview of the stanced sentences in BBC. We summarise the important clues of differentiation among the six stance categories in Figure 1:

Figure 1: The most important characteristics of the six types of categories in BBC.

6 *Conclusions*

This exploratory study aimed to uncover whether stance in argumentative sentences from blog sources can be identified on purely formal grounds, which would benefit automatic identification of stance-taking in discourse. Two research questions were at the heart of this work. They concern (i) whether the

six stance categories are different in terms of formal clues, and (ii) what the salient features of the stanced discourse in a subset from BBC are. We described linguistic characteristics at character, lexical and sentence level in a stance-annotated data set, in order to derive elements of similarity and differentiation among six core stance categories in sentences from BBC. We showed that our data contain sentences with a limited number of special characters, interjections, digits, or less commonly used punctuation. The writers make use of a varied vocabulary when they express stance. We also found the differences across the six stance types and concluded that CONTRARIETY and NECESSITY are the most discriminative stance categories. The former makes use of longer sentences and shorter words, more conjunctions, and more repeated forms than all other stances. The latter makes use of longer lexical forms and shorter sentences.

These findings can be further analysed and expanded to other disciplines and purposes. From a Text Mining perspective, classification experiments based on the feature set analysed in the present study could be performed, as well as other NLP tasks in text analysis of stanced discourse. In addition, these formal features will provide useful information about stance types in the task of identification of new stance markers. The present study provides the initial insights to increase our understanding of the role of formal marking that can be useful for the retrieval of data on the basis of functional – notional categories, since manual annotation is time-consuming and costly. Our findings suggest radically new paths and new knowledge about the investigation of writers' attitudes in discourse generally or under specific communicative circumstances. The features used in this study were tested and analysed in Simaki et al. (2017b) as part of a broader feature set that also included the present six stance categories. The classification results of that study support our findings and conclusions here, and show that CONTRARIETY and NECESSITY are the most discriminative categories in the corpus. This indicates that the proposed framework is a robust point of departure for further studies in speaker stance identification. As a future step, the features used here will have to be analysed in more depth in order to determine their full potential as informative clues for stance identification for computational purposes (feature selection), and to sift out new data for new experimentation. This stance framework and the linguistic patterns derived here should also be evaluated in different types of texts from social media or other sources. Our categories and the linguistic clues that are related to each stance type can be of great value for applications for product and service reviews. Also, a very interesting future aspect of our study would be to uncover the sociodemographic information of speakers as a function of stance-taking, or *vice versa* stance-taking as a function of speaker sociodemographics.

Acknowledgments. This research is part of the StaViCTA project,[12] supported by the Swedish Research Council (framework grant the Digitized Society Past, Present, and Future, No. 2012-5659).

Notes

1. The Brexit Blog Corpus (BBC) is publicly available through the Swedish National Data Service (SND): https://snd.gu.se/en/catalogue/study/snd1037
2. Measures the frequency of the ˜ , @, /, $, %, ˆ , &, *, -, =, +, >, < symbols.
3. Measures the frequency of the (,), [,]. —, ,, ;, ?, ., !, :, ’, “, ” symbols.
4. Measures the frequency of different forms within a sentence.
5. Measures the frequency of forms appearing once in the sentence.
6. Measures the frequency of forms appearing twice in the sentence.
7. The features of this category are based on the Penn Treebank tagset (that NLTK’s default tagger uses), and they are grouped into the following grammatical categories: nouns (NN, NNS, NNP, NNPS), pronouns (PRP, WP, WP$, PRP$), adjectives (JJ, JJS, JJR), verbs (VBG, VBD,VBN, VBP, VBZ, VB), adverbs (RB, WRB, RBS, RBR), prepositions (TO, IN), conjunctions (CC), interjections (UH), determiners (DT, WDT), and particles (RP).
8. Gavagai API: https://developer.gavagai.se
9. NLTK: http://www.nltk.org/
10. https://www.gov.uk/guidance/style-guide
11. http://www.plainenglish.co.uk/files/howto.pdf
12. StaViCTA project: http://cs.lnu.se/stavicta/

References

Adar, Eytan, Li Zhang, Lada A. Adamic and Rajan M. Lukose. 2004. Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem* 13 (1): 16989–16995.

Agarwal, Nitin and Huan Liu. 2008. Blogosphere: Research issues, tools, and applications. *ACM SIGKDD Explorations Newsletter* 10 (1): 18–31.

Anand, Pranav, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani and Michael 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*, *WASSA ’11*, 1–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bassiouney, Reem. 2012. Politicizing identity: Code choice and stance-taking during the Egyptian revolution. *Discourse & Society* 23 (2): 107–126.

Benveniste, Émile. 1971. Subjectivity in language. In M. E. Meek (ed.). *Problems in general linguistics*, 223–230. Coral Gables, FL: University of Miami Press.

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

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

Cabrejas-Peñuelas, Ana B. and Mercedes Díez-Prados. 2014. Positive self-evaluation versus negative other-evaluation in the political genre of pre-election debates. *Discourse & Society* 25 (2): 159–185.

Cataldi, Cataldi, Mario, Luigi Di Caro and Claudio Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining* 4, 1–10. Washington, DC, USA: Association for Computing Machinery.

Chaemsaithong, Krisda. 2012. Performing self on the witness stand: Stance and relational work in expert witness testimony. *Discourse & Society* 23 (5): 465–486.

Chiluwa, Innocent and Presley Ifukor. 2015. 'War against our Children': Stance and evaluation in #BringBackOurGirls campaign discourse on Twitter and Facebook. *Discourse & Society* 26 (3): 267–296.

Conrad, Susan and Douglas Biber. 2000. Adverbial marking of stance in speech and writing. In G. Thompson (ed.). *Evaluation in text: Authorial stance and the construction of discourse*, 56–73. Oxford: Oxford University Press.

Downing, Angela. 2001. "Surely you knew!": Surely as a marker of evidentiality and stance. *Functions of Language* 8 (2): 251–282.

Du Bois, John. 2007. The stance triangle. In R. Englebretson (ed.). *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 139–182. Amsterdam: John Benjamins.

Ekberg, Lena and Carita Paradis. 2009. Editorial: Evidentiality in language and cognition. *Functions of Language* 16 (1): 5–7.

Englebretson, Robert. 2007. Stancetaking in discourse: An introduction. In R. Englebretson (ed.). *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 1–25. Amsterdam: John Benjamins.

Facchinetti, Roberta, Frank Palmer and Manfred Krug (eds.). 2003. *Modality in contemporary English* (Topics in English Linguistics 44). Berlin: Walter de Gruyter.

Faulkner, Adam. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. *Science* 376 (12): 86.

Ferreira, William and Andreas 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,* 1163–1168. Sheffield, UK.

Fuoli, Matteo. 2012. Assessing social responsibility: A quantitative analysis of Appraisal in BP's and IKEA's social reports. *Discourse & Communication* 6 (1): 55–81.

Glynn, Dylan and Mette Sjölin. 2015. Subjectivity and epistemicity: Corpus, discourse, and literary approaches to stance. In D. Glynn and M. Sjölin (eds.). *Corpus, discourse, and literary approaches to stance* (*Lund Studies in English* 117), 360–410. Lund: Lund University.

Granger, Sylviane. 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, Bethany and Douglas Biber. 2014. Stance markers. In K. Aijmer and C. Rühlemann (eds.). *Corpus pragmatics: A handbook*, 219–248. Cambridge: Cambridge University Press.

Gu, Xiang. 2015. Evidentiality, subjectivity and ideology in the Japanese history textbook. *Discourse & Society* 26 (1): 29–51.

Hasan, Kazi Saidul and Vincent Ng. 2013a. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceeding of IJCNLP 2013: The 6th International Joint Conference on Natural Language Processing,* 1348–1356. Nagoya, Japan.

Hasan, Kazi Saidul and Vincent Ng. 2013b. Frame semantics for stance classification. In *Proceedings of CoNLL 2013: The Seventeenth Conference on Computational Natural Language Learning,* 124–132. Sofia, Bulgaria.

Hasan, Kazi Saidul and Vincent Ng. 2013c. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 2: Short Papers), 816–821. Sofia, Bulgaria.

Hasan, Kazi Saidul and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 751–762. Doha, Qatar.

Hunston, Susan and Geoffrey Thompson (eds.). 2000. *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7 (2): 173–192.

Jiang, Feng Kevin. 2017. Stance and voice in academic writing. *International Journal of Corpus Linguistics* 22 (1): 85–106.

Kanté, Issa. 2010. Mood and modality in finite noun complement clauses: A French-English contrastive study. *International Journal of Corpus Linguistics* 15 (2): 267–290.

Kärkkäinen, Elise. 2003. *Epistemic stance in English conversation: A description of its interactional functions, with a focus on* I think (Pragmatics & Beyond New Series 115). Amsterdam: John Benjamins.

Kessler, Brett, Geoffrey Numberg and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 32–38. Association for Computational Linguistics.

Kucher, Kostiantyn, Andreas Kerren, Carita Paradis and Magnus 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, Kostiantyn, Teri Schamp-Bjerede, Andreas Kerren, Carita Paradis and Magnus Sahlgren. 2016b. Visual analysis of online social media to open up the investigation of stance phenomena. *Information Visualization* 15 (2): 93–116.

Kucher, Kostiantyn, Carita Paradis, Magnus Sahlgren and Andreas Kerren. 2017. Active learning and visual analytics for stance classification with ALVA. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7 (3): 1–31.

Martin, James R. and Peter R. White. 2003. *The language of evaluation*. London: Palgrave Macmillan.

Mathioudakis, Michael and Nick Koudas. 2010. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, 1155–1158. Association for Computing Machinery.

Mohammad, Saif M., Parinaz Sobhani and Svetlana Kiritchenko. 2016. Stance and sentiment in tweets. *arXiv preprint arXiv:1605.01655*.

Mukherjee, Arjun and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 207–217. Association for Computational Linguistics.

Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media,* 439–448. Cambridge, Massachusetts, USA.

Pak, Alexander and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of The Seventh International Conference on Language Resources and Evaluation (LREC)* (Vol. 10), 1320–1326. Valletta, Malta.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2 (1–2): 1–135.

Paradis, Carita. 2003. Between epistemic modality and degree: The case of *really.* In R. Facchinetti, F. Palmer and M. Krug (eds.). *Modality in contemporary English* (Topics in English Linguistics 44), 191–222. Berlin: DeGruyter.

Park, Jaram, Young Min Baek and Meeyoung Cha. 2014. Cross-cultural comparison of nonverbal cues in emoticons on twitter: Evidence from big data analysis. *Journal of Communication* 64 (2): 333–354.

Paterson, Laura L., Laura Coffey-Glover and David Peplow. 2016. Negotiating stance within discourses of class: Reactions to Benefits Street. *Discourse & Society* 27 (2): 195–214.

Peersman, Claudia, Walter Daelemans and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the Third International Workshop on Search and Mining User-Generated Contents*, 37–44. Association for Computational Linguistics.

Persing, Isaac and Vincent Ng, V. 2016. Modeling stance in student essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2174–2184. Association for Computational Linguistics.

Põldvere, Nele, Matteo Fuoli and Carita Paradis. 2016. A study of dialogic expansion and contraction in spoken discourse using corpus and experimental techniques. *Corpora* 11 (2): 191–225.

Precht, Kristen. 2003. Stance moods in spoken English: Evidentiality and aspect in British and American conversation. *Text* (Special issue: *Negotiating Heteroglossia: Social Perspectives on Evaluation*) 23 (2): 239–257.

Rajadesingan, Ashwin and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 153–160. Berlin: Springer International Publishing.

Read, Jonathon and John Carroll. 2012. Annotating expressions of appraisal in English. *Language Resources and Evaluation* 46 (3): 421–447.

Saurí, Roser and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation* 43 (3): 227–268.

Scheffé, Henry. 1999 [1959]. *The analysis of variance*. New York City: John Wiley & Sons.

Schwartz, Andrew, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin Seligman and Lyle Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOSONE* 8 (9): e73791.

Simaki, Vasiliki. 2015. *Sociolinguistic research on web textual data* (Doctoral dissertation, in Greek). University of Patras, Greece. Retrieved from: http://hdl.handle.net/10889/9422

Simaki, Vasiliki, Christina Aravantinou, Iosif Mporas and Vasileios Megalooikonomou. 2015a. Using sociolinguistic inspired features for gender classification of web authors. In *International Conference on Text, Speech, and Dialogue (TSD)* (Lecture Notes in Computer Science, vol. 9302), 587–594. Berlin: Springer International Publishing.

Simaki, Vasiliki, Christina Aravantinou, Iosif Mporas and Vasileios Megalooikonomou. 2015b. Automatic estimation of web bloggers' age using regression models. In *International Conference on Speech and Computer (SPECOM)*, 113–120. Berlin: Springer International Publishing.

Simaki, Vasiliki, Christina Aravantinou, Iosif Mporas, Marianna Kondyli and Vasileios Megalooikonomou. 2017a. Sociolinguistic features for author gender identification: From qualitative evidence to quantitative analysis. *Journal of Quantitative Linguistics* 24 (1): 65–84.

Simaki Vasiliki, Carita Paradis and Andreas Kerren. 2017b. Stance classification in texts from blogs on the 2016 British Referendum. In A. Karpov, R. Potapova and I. Mporas (eds.). *Speech and computer. SPECOM 2017* (Lecture Notes in Computer Science, vol. 10458), 700–709. Berlin: Springer International Publishing.

Simaki, Vasiliki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher and Andreas Kerren. 2017c. Annotating speaker stance in discourse: The Brexit Blog Corpus. *Corpus Linguistics and Linguistic Theory.* DOI:10.1515/cllt-2016-0060

Somasundaran, Swapna and Janyce 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, Dhanya, Lise Getoor and Marilyn 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, Maryland, USA.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60 (3): 538–556.

Stamatatos, Efstathios, Nikos Fakotakis and George Kokkinakis. 2000. Automatic authorship attribution. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, 158–164. Association for Computational Linguistics.

Stamatatos, Efstathios, Nikos Fakotakis and George Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities* 35 (2): 193–214.

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

Tracy, Karen. 2011. What's in a name? Stance markers in oral argument about marriage laws. *Discourse & Communication* 5 (1): 65–88.

Tukey, John W. 1949. Comparing individual means in the analysis of variance. *Biometrics* 5 (2): 99–114.

Van de Kauter, Marjan, Bart Desmet and Véronique Hoste. 2015. The good, the bad and the implicit: A comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation* 49 (3): 685–720.

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

Walker, Marilyn, Pranav Anand, Robert Abbott and Ricky 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, Marilyn, Pranav Anand, Robert Abbott, Jean E. Fox Tree, Craig Martell and Joseph King. 2012b. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems* 53 (4): 719–729.

Walker, Marilyn, Jean E. Fox Tree, Pranav Anand, Robert Abbott and Joseph King. 2012c. A corpus for research on deliberation and debate. In *Proceedings of The Eighth International Conference on Language Resources and Evaluation (LREC)*, 812–817. Istanbul, Turkey.

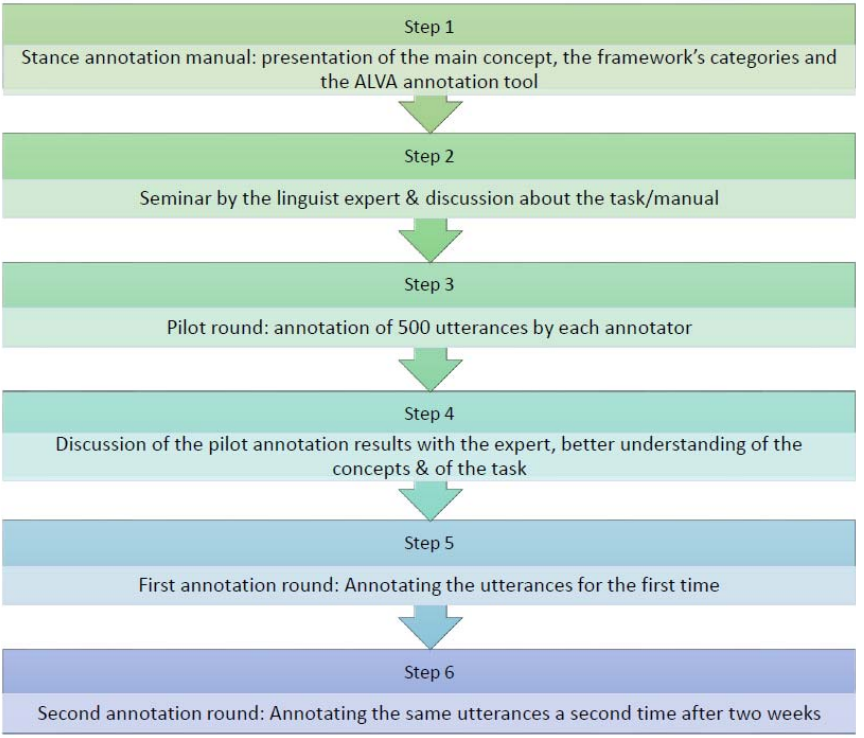
White, Peter R. 2003. Beyond modality and hedging: A dialogic view of the language of intersubjective stance. *Text* 23 (2): 259–284.

Wiebe, Janyce, Theresa Wilson and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39 (2): 165–210.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics* 30 (3): 277–308.

Zheng, Rong, Jiexun Li, Hisnchun Chen and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology* 57 (3): 378–393.

Appendix 1: The protocol followed in the annotation process of BBC. Two annotators, one who is a professional translator with a Licentiate degree in English Linguistics and the other one with a PhD in Computational Linguistics, carried out the annotations independently of one another. For the ALVA annotation tool, see Kucher et al. (2016a, 2017).



Appendix 2: The inter- and intra-annotator agreement sets in terms of the F- and Kappa scores.

Stance categories	Inter-annotator agreement set		Intra-annotator agreement set	
	Mean F-score	**Mean Kappa**	**Mean F-score**	**Mean Kappa**
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
UNCERTAINTY	0.62	0.58	**0.81**	**0.79**
CERTAINTY	0.21	0.20	0.58	0.58
AGREEMENT/DISAGREEMENT	0.45	0.42	0.67	0.65
TACT/RUDENESS	0.55	0.54	0.78	0.77
VOLITION	0.44	0.43	0.71	0.70