RiCL Research in Corpus Linguistics

Evaluating stance annotation of *Twitter* data

Vasiliki Simaki – Eleni Seitanidi – Carita Paradis
Lund University / Sweden

Abstract – Taking stance towards any topic, event or idea is a common phenomenon on *Twitter* and social media in general. *Twitter* users express their opinions about different matters and assess other people's opinions in various discursive ways. The identification and analysis of the linguistic ways that people use to take different stances leads to a better understanding of the language and user behaviour on *Twitter*. Stance is a multidimensional concept involving a broad range of related notions such as modality, evaluation and sentiment. In this study, we annotate data from *Twitter* using six notional stance categories —contrariety, hypotheticality, necessity, prediction, source of knowledge and uncertainty— following a comprehensive annotation protocol including inter-coder reliability measurements. The relatively low agreement between annotators highlighted the challenges that the task entailed, which made us question the inter-annotator agreement score as a reliable measurement of annotation quality of notional categories. The nature of the data, the difficulty of the stance annotation task and the type of stance categories are discussed, and potential solutions are suggested.

Keywords – stance-taking; social media discourse; corpus annotation; inter-coder reliability

1. INTRODUCTION[1]

The development of social media platforms has given rise to a new type of discourse serving different purposes. The platforms are used by different actors to express opinions and assess other people's opinions but also to construct and establish their online identity over time. Despite their similarities, each social media platform has a different character and slightly different policies. These conditions have repercussions on how the platform users communicate. Especially in the case of *Twitter*, users are restricted to a specific tweet size and specific interaction functions —reply, like, retweet and share— which naturally affect the nature of the discourse. Stance-taking in tweets is pervasive. Expressions of stance are used to promote, reinforce or mitigate the communicative goals of the users such as, for instance, to search for information or make information more visible (Zappavigna 2012: 50ff.). Conversational practices through the various *Twitter* features, such as retweet, reply and mentions, have emerged (Boyd *et al.* 2010), and as

[1] The authors would like to thank two anonymous reviewers for their comments. This research was supported by the *Kamprad Family Foundation* (Reference No: 20180178).

Research in Corpus Linguistics 11/1: 53–80 (2023). Published online 2022. ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.11.01.03

Honey and Herring (2009) point out, opinions, sentiments and stances are present in such interactions. The study of the discursive ways that *Twitter* users employ to communicate their stances offers important insights about users' behaviour and language use.

Stance and stance-taking are concepts strongly related to modality and sentiment that have been widely studied in different research fields and for various purposes. Kaltenböck *et al.* (2020: 1) define stance as

> the way in which speakers express points of view, attitudes, feelings and evaluations, and position themselves in relation to some proposition (i.e. subjectivity) and to other speech participants (i.e. intersubjectivity) and their particular stances.

A similar definition is provided for stance-taking in Simaki *et al.* (2020: 217) 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.

Based on this definition, a stance framework with ten notional categories such as certainty, contrariety and necessity, among others, was introduced in Simaki *et al.* (2020). A general framework consisting of stance concepts that go beyond pro/con statements has the potential of important advances in stance studies in corpus pragmatics, computational linguistics, content analysis and other relevant disciplines. However, the annotation of texts using this stance framework is challenging since complexity, subjectivity and the background of the annotator can affect the annotation results and, consequently, the reliability of the dataset.

In this study, we test the validity of the abovementioned stance framework in order to show how suitable our categories are in a stance analysis task. Our stance framework was initially tested in data from blogs, and for this task we continued working with data from *Twitter*, as these data types fall within the social media discourse genre in the broad sense but are different in a range of ways from blog texts. Our purpose is to identify stance and attribute a stance label to the selected data, but we acknowledge the fact that this might not be possible for every tweet included in the data set. For this reason, in addition to the six stance categories that we used, namely, contrariety, hypotheticality, necessity, prediction, source of knowledge and uncertainty, we introduced a seventh category, entitled 'no label', which included those tweets that could not be attributed to any of the other stance labels.[2] The *Twitter* data in the study were annotated by two experts (annotators A and B), and the inter-annotator agreement was calculated. The relatively low level of agreement between the annotators led us to a broader discussion of discourse annotation, inter-annotator agreement measurements and what is considered to be an acceptable agreement level that ensures the reliability of the annotated data. The particular aims of this study are:

1. to evaluate the stance framework on annotated *Twitter* data of a wide thematic range;
2. to identify patterns and/or possible problems of the annotation scheme;
3. to propose solutions to improve the annotation results in the future;
4. to describe and analyse the complexity and the different components of tweets annotated as ‘no label’ for the refinement and improvement of the stance framework and annotation protocol.

2. BACKGROUND WORK

As a result of the expansion of social media platforms, social media discourse has become the focus of research from various perspectives in linguistics and other disciplines. The analysis of this discourse type can be a challenging task, because of ethical, formatting and language issues that may arise (Hernández 2014), as well issues related to the authors’ identity and communicative purposes (Yus 2011, 2016). *Twitter* data is special in many ways regarding the relations among users and the features that are available. *Twitter* users establish social relationships based on the notion of ‘following’ (the user has followers and follows other accounts), and this affects the tweets that are shown in their timeline, which has an impact on their network. When it comes to the platform’s features, the @ symbol is used for addressivity/communicative purposes among users and the # symbol as a feature of searchable tweets/conversations. These and other *Twitter* features have been extensively studied, especially hashtags (sequences starting with the # symbol) and their function that enables users to search for specific content and make comments searchable for others (Zappavigna 2015; Zhu 2016). A new type of publicness has emerged from *Twitter* with users presenting information of personal relevance (Schmidt 2014). *Twitter* is also used to create communities and networks sharing common

[2] See Section 3 for a detailed description and examples of all categories and labels.

experiences and/or similar values, and in such environments stance-taking is pervasive (Zappavigna and Martin 2018).

Broadly speaking, stance-taking is the way people use language to position themselves, express their opinions and assess their own and other people’s messages (Du Bois 2007). It has been studied in various contexts, and a whole range of aspects are involved, these including modality (Facchinetti *et al.* 2003; Marín-Arrese *et al.* 2014), evaluation (Hidalgo-Downing 2012; Fuoli 2018), evidentiality (Ekberg and Paradis 2009), subjectivity/intersubjectivity (Verhagen 2005; Marín-Arrese 2017) and sentiment (Taboada 2016). The analysis of speaker stance is a vibrant area in language sciences, with many studies aiming to understand better its role in human communication (Hunston and Thompson 2000; Berman *et al.* 2002) and its association to social roles, identities, interpersonal and social relationships (Jaffe 2009), while others focus on stance phenomena in specific types of discourse (Hyland 2005; Biber 2006; Perrin 2012), including social media discourse (Jacknick and Avni 2017), specific stance-taking expressions (Paradis 2003) and discourse markers that are strongly related to stance (Traugott 2020). Apart from the qualitative approaches, corpus-based methodologies offer important insights into the identification of stance and stance expressions in discourse. Such methods and tools offer the possibility to investigate stance-taking in large amounts of data and perform statistical tasks and analyses to identify patterns in the data across time and discourse types (Alonso Ameida 2015). Stance has also been studied from a computational perspective (Ghosh *et al.* 2019; Küçük and Can 2020) with many researchers addressing stance as a binary phenomenon of the speaker’s pro/con positioning in relation to a topic, an idea or an event (AlDayel and Magdy 2021). Stance annotation, in particular, has also been studied extensively with researchers aiming at creating as comprehensive annotation systems and tools as possible, which allows to use the annotation for automatic stance detection and classification (Kucher *et al.* 2016). Such tasks are performed in data extracted from ideological forum debates (Hasan and Ng 2014), news articles (Ferreira and Vlachos 2016), academic text data (Faulkner 2014) or other social media sources (Mohammad *et al.* 2016; Pamungkas *et al.* 2019).

In Simaki *et al.* (2020), the point of departure is a notional definition of speaker stance rather than a lexical one. According to this definition, discussed in Section 1 above, the concept of stance is defined as a psychological state involving speakers’ beliefs and attitudes, stance-taking as human performance in communication and expressions of stance as the constructions used for stance-taking in discourse. As a result, and based on the literature in the field, an original stance framework consisting of ten notional stance categories was proposed.[3] These categories were manually identified and attributed to utterances extracted from blogs thematically related to the 2016 UK referendum. The final output of this procedure resulted in the *Brexit Blog Corpus* (BBC).[4] Simaki *et al.* (2020) showed that stance-taking is common practice in discussions of controversial political matters such as the Brexit. The distribution of the categories showed that contrariety was the most frequent category in the corpus, while the category of volition was the least frequent one. The presence of more than one instance of stance-taking in the same utterance was also shown to be a frequent phenomenon. The calculation of the inter-coder reliability showed good agreement scores for the categories of contrariety, hypotheticality, necessity and uncertainty.

In subsequent studies, the BBC was computationally (Simaki *et al.* 2017a) and statistically (Simaki *et al.* 2018a) evaluated in order to test the framework's efficacy and to provide new insights about linguistic patterns for the identification of stance in discourse in future work. In Simaki *et al.* (2019), the aim was to identify specific constructions that are related to the six most frequent stances in the BBC categories. A quantitative analysis of the annotated corpus data and a meta-annotation procedure to identify lexical forms (stance markers) that are stance-specific for each category were performed. The results of the two techniques were then compared, and a list of constructions of stance-related discourse as particularly salient expressions of each stance type was proposed.[5] Part of this list is used in the present study, as will be shown in Section 4.

3. Methodology

In this study, our hypothesis was that the stance framework mentioned above is suitable for the analysis of stance in discourse and its use can be generalised to social media discourse types other than blogs and a wide variety of topics. For that purpose, we used texts retrieved from *Twitter* which were extracted on the basis of specific criteria from a social media corpus (see Section 4). We selected *Twitter* as the source of data for our

[3] See the full framework with a brief description and examples for each category in Appendix 1.
[4] https://snd.gu.se/en/catalogue/study/snd1037
[5] These constructions are presented in Appendix 2.

study since tweets vary from blog texts, the most important difference being the character limitation of the tweet in contrast to blog texts, which can be as long as the blog author wants. In addition, *Twitter* is frequently the source of data in which researchers from different disciplines dive into to explore people's ideas, beliefs and opinions about various topics, and we have prior experience with the particularities and challenges of such data type.

We used the six most frequent stance categories distinguished in the stance framework (Simaki *et al.* 2020), namely, contrariety, hypotheticality, necessity, prediction, source of knowledge and uncertainty. The category of contrariety includes instances where the authors express a compromising/contrastive opinion (e.g., *Hate the end result, but #thegame always delivers. always. best rivalry in sports*).[6] Hypotheticality is attested in utterances where authors express a possible consequence of a condition, mostly formulated with conditional clauses (e.g., *If you use this Kim Kardashian hashtag thing it's an instant unfollow*). Necessity includes cases in which authors express requests, recommendations, instructions or obligations (e.g., *I really need to start utilizing a day minder*). Prediction is attested when authors make a guess or a conjecture about a future event (e.g., *@lazycat99 I knew someone would catch that reference. Well done!*). Source of knowledge occurs when authors express the origin of what they say (e.g., *One mustn't be much concerned with living, but with living well... Socrates to Crito, in Plato, 'Crito', 48b*). Finally, the category of uncertainty concerns authors' doubt regarding the likelihood of what they say (e.g., *Stand up special starts in 20 mins. I think 7 on West Coast. 10 on East. I actually have no fucking idea*). As already stated (see Section 1), in the present study, we introduced another category, 'no label', which deals with tweets that did not fit in any of the abovementioned categories. This includes neutral statements (e.g., *rt @ankhmarketing: ms. lauryn hill [@mslaurynhill] performing live may 12th [@thewarfield!!] @goldenvoicesf*), questions (e.g., *@amwalkush @britenyc but we are still going to decorate gourds, right?*), ambiguous and/or illegible tweets (e.g., *Me. Stretch. Hollywood. rt @will_blackmon: I wear a 3 piece suit in a cab son. Who needs a limo!*) and tweets expressing sentiment (e.g., *So grateful for u all and ur kind comments. May this brighten ur day. Love u! #standbyyou*) or more than one stance (e.g., *@jjenas8 You might be right but you're wrong*). Two annotators annotated the data (see Section 5.1), and the inter-annotator agreement was calculated.

[6] Unless otherwise stated all instances have been retrieved from dataset used in this study.

4. CORPUS DESCRIPTION

We use data from the *Twitter* part of the social media corpus used in Simaki *et al.* (2017b). Simaki *et al.*'s (2017b) corpus was compiled with data from the official *Facebook* and *Twitter* profiles of public figures such as actors, authors or athletes. It was manually annotated with the authors' sociodemographic information such as their gender, age, profession and any other additional information available as, for instance, their educational background. In contrast to the BBC, this corpus consists of texts on various topics, such as personal branding, social and political matters, nature, etc. The corpus was compiled from September to December 2015 at the same time the BBC was build. It includes texts from 838 different authors (535 male and 303 female authors) and its overall size is of 13.4 million words distributed in 721,033 entries. The data were further processed and normalised and, as a result, features typical of *Twitter* discourse (e.g., multimodality, the use of upper/lower case letters, hashtags or emojis/emoticons, among others) were excluded and, therefore, are also disregarded in the present study. However, some features such as hashtags (#), mentions (@) and links have been included in the data.

In Simaki *et al.* (2019), a list of stance markers for each stance category was compiled containing both stance-related forms, such as *but*, *if*, *must*, and forms that do not unambiguously evoke a specific type of stance but are stance-related in the sense that they occur frequently in long sequences that express stance, such as *then* (e.g., *If you're not willing to risk it all then you do not want it bad enough*) and *would* (e.g., *It would be cute if they didn't draw on me*) that are frequent forms in hypothetical sentences. The markers are based on a two-fold analysis of the BBC data: first, the extraction of the statistically significant lexical items per category and, second, the identification of the stance-related lexical chunks by one of the annotators, who —five months before this task— had conducted the initial annotation task. The findings from both analyses were combined and the results are shown in Appendix 2. For the present study, we refined that list by excluding forms that would create noise in the data selection process such as *I*, *be*, *is*, *have* and *it*. To avoid a high number of neutral or irrelevant tweets, we selected texts from the *Twitter* set of the corpus in which at least one stance marker from the refined list was present. This list contains 20 markers, and 1,000 tweets were extracted. This was possible for many of the markers that are used frequently in tweets but, in some cases, the search retrieved fewer tweets. In Table 1, we present the list of the stance markers that were searched for in the data, the corresponding stance categories for these markers and the number of tweets extracted per marker.

Stance category	Stance marker	Number of tweets
Contrariety	*But*	1,000
	Than	1,000
	While	436
Hypotheticality	*Could*	1,000
	If	1,000
	Would	1,000
	Then	819
Necessity	*Need*	1,000
	Must	396
	Needs	259
	Should	1,000
Prediction	*Will*	1,000
Source of knowledge	*As*	1,000
	Said	582
	Show	1,000
	That	1,000
Uncertainty	*Think*	1,000
	Might	350
	Maybe	337
	Probably	168
Uncertainty/ prediction	*May*	521
	Total	**15,868**

Table 1: Stance markers used for the extraction of the data listed according to the stance categories they pertain to, and the number of the tweets extracted

As illustrated in Table 1, the total size of the dataset is 15,868 texts (274,697 words). The markers *may*, *maybe*, *might*, *must*, *needs*, *probably*, *said*, *then* and *while* were limited in number (fewer than 1,000) and, thus, all tweets in which they were present were extracted. The relevance of these stance markers to the annotation results and our research findings will be discussed in Section 6.

5. Corpus annotation and results

In this section, we describe the annotation procedure and the annotation results from the pilot and the final annotation rounds.

5.1. Corpus annotation procedure

The annotation of the data was carried out by two annotators with a background in linguistics. More specifically, annotator A holds a PhD in linguistics and computational linguistics, whereas annotator B holds a Master in English applied linguistics. A comprehensive annotation protocol was introduced, comprising six steps, as presented in Table 2.

1	Presentation of the main concept, the stance categories and familiarisation with previous studies.
2	New label for current task: ‘no label’ category in which neutral statements, questions, ambiguous and/or illegible tweets and tweets with sentiment or more than one stance are stored.
3	Discussion between Annotator A (research expert) and Annotator B (research assistant) about the task.
4	Pilot annotation of 664 tweets by annotator A and Annotator B.
5	Discussion between Annotators A and B about conflicting assessments/ambiguous cases.
6	Annotation of 6,659 tweets by Annotator A and 15,868 by Annotator B.

Table 2: The annotation protocol followed in the study

As shown in Table 2, firstly, annotator A explained the main concept, the stance framework and the categories. Instructions about the annotation process were also provided, so that both annotators would base their decisions on the overall meaning of each text and would not merely rely on the potential presence of a specific stance marker. Annotator B studied previous work to become familiar with the task. Secondly, a new category was added: the ‘no label’ category. Annotators attributed this label to neutral statements that do not express any stance, ambiguous or illegible tweets, tweets with more than one stance and tweets expressing sentiment but not stance. In addition, we did not exclude the tweets in which a question mark was present, as in many cases it is not used to form a question (e.g., *@imsoforserious calling someone ugly, stupid or a cunt is hardly criticism. but thanks for trying to teach me something super obvious (?)*). The idea to add this category stems from Simaki *et al.* (2018b), in which many texts were stance-free, neutral, expressing sentiment or irrelevant. Further analysis of this category will provide feedback about the discourse of *Twitter* and improve the stance annotation process. In the third step of the annotation process, annotators A and B discussed the task while the fourth step was a pilot annotation round of 664 tweets by both annotators. In step five, after the pilot round, the two annotators discussed the challenges of the task, problematised conflicting or ambiguous tweets, and problems were resolved. Finally, in step 6, the final annotation round was conducted.

5.2. Pilot round annotation results

The reliability of the annotated set of 664 tweets in the pilot round was tested by calculating the level of agreement between the annotations by annotators A and B. We used the coefficient kappa (Cohen 1960) to calculate the inter-annotator agreement score with a confidence level 95 per cent. The results are shown in Table 3 which provides the distribution of the annotated tweets in each stance category, and the inter-annotator agreement score. The highest level of agreement between the two annotators was achieved for the source of knowledge category (0.77), followed by the necessity (0.59) and contrariety (0.58) categories. The overall inter-annotator agreement for this set of tweets is 0.54, which can be characterised as moderate according to Landis and Koch (1977).

		Annotator A								
	Categories	**Contrariety**	**Hypotheticality**	**Necessity**	**Prediction**	**Source of knowledge**	**Uncertainty**	**No label**	**Total**	**Kappa**
Annotator B	**Contrariety**	71	5	1	6	8	7	21	119	**0.58**
	Hypotheticality	4	49	1	3	2	2	7	68	0.53
	Necessity	7	38	91	9	2	12	15	174	**0.59**
	Prediction	3	0	0	31	1	15	4	54	0.48
	Source of knowledge	7	2	5	6	86	0	8	114	**0.77**
	Uncertainty	4	2	1	8	1	63	34	113	0.51
	No label	4	4	1	1	0	1	11	22	0.13
	Total:	100	100	100	64	100	100	100	**664**	**0.54**

Table 3: Annotation results of the pilot round and kappa scores

5.3. Final round annotation results

After steps four and five, the final annotation round was carried out: 6,659 tweets were annotated by annotator A and 15,868 tweets by annotator B. Table 4 shows the overall results of the annotation. As can be noticed, the ‘no label’ category is the largest category according to the annotations of both annotators. This category is more than twice as large when compared to the rest of the annotated categories, which shows the extent to which our stance annotation criteria did not apply. For annotator A, the most frequent categories are uncertainty, contrariety and necessity. For annotator B, necessity is the most frequent type of stance, which is followed by contrariety and hypotheticality. For both annotators, prediction is the least frequent category.

Stance categories	**Annotator A**	**Annotator B**
Contrariety	687	1,632
Hypotheticality	386	1,207
Necessity	628	2,235
Prediction	64	278
Source of knowledge	134	809
Uncertainty	730	1,034
No label	4,030	8,673
Total	**6,659**	**15,868**

Table 4: Final annotation round results

We, then, calculated the interrater reliability of the annotated tweets, which is shown in Table 5.

		Annotator A								
	Categories	**Contrariety**	**Hypotheticality**	**Necessity**	**Prediction**	**Source of knowledge**	**Uncertainty**	**No label**	**Total**	**kappa**
Annotator B	**Contrariety**	439	3	9	5	2	12	332	802	**0.54**
	Hypotheticality	5	164	2	3	0	6	224	404	0.38
	Necessity	5	66	383	2	0	15	397	868	0.45
	Prediction	0	0	3	17	1	12	69	102	0.19
	Source of knowledge	15	11	20	4	91	33	202	376	0.34
	Uncertainty	4	13	15	12	0	429	187	660	**0.57**
	No label	219	129	196	21	40	223	2,619	3,447	0.32
	Total:	687	386	628	64	134	730	4,030	**6,659**	**0.42**

Table 5: Final annotation results and kappa scores

As can be noticed, the kappa score for the total set of annotated data is 0.42, which is a much lower score than the score in the pilot round. In this set, we achieved the highest inter-annotator agreement score for the uncertainty category (0.57), as well as the second highest score for the contrariety category (0.54). Among the labelled tweets, these two categories are the most frequent ones. Interestingly, source of knowledge, which was the stance category with the highest inter-annotator agreement score in the pilot round, shows a lower kappa score in this round (0.34). The most important finding is the frequency of the ‘no label’ category in this annotation round (55% of the corpus). However, the low agreement score (0.32) on the tweets grouped in this category suggests that the annotators faced difficulties in applying the annotation instructions in the same way. This difficulty can be due to the relatively high level of the subjectivity of the task since the categories are notional rather than being identified through lexical items. In comparison with the very low kappa score in the pilot round (0.13), in this round, the kappa indicates a better agreement score (0.32).

The low overall kappa score in the final annotation round led to the implementation of alternative measures which have more advantages regarding the type of data that they support and the handling of the missing data. We calculated the Krippendorff's Alpha coefficient (K alpha; Krippendorff 2011), which is another standard and relevant metric for the calculation of the interrater reliability as a reference metric. In addition, we calculated the Gwet's AC_1 coefficient (Gwet 2002), a more recent metric which has been suggested as a more robust solution to evaluate annotations of discourse data, in which skewed data and variability in the distribution of categories are quite common phenomena (Hoek and Scholman 2017). Table 6 shows, the inter-annotator agreement scores, which are calculated by using three different metrics. The results show that the kappa and K alpha scores have similar values (0.42 and 0.41, respectively), while Gwet's AC_1 shows a higher score (0.58). These results gave rise to methodological considerations regarding the annotation of discourse data and the annotated data reliability and quality. This will be discussed in Section 7.

Metric	**Score**
Kappa	0.42
K alpha	0.41
Gwet's AC_1	0.58

Table 6: Results of the different metrics

6. Analysis of the annotation results

When it comes to the analysis of the annotated data, we start with the frequency of the six stance categories. As shown in Table 5, the most frequent stances for annotator A were uncertainty, contrariety, necessity and hypotheticality. For annotator B, necessity was the most frequent stance, followed by contrariety, hypotheticality and uncertainty. Examples (1)–(4) illustrate the most frequent categories in the annotated data.

(1) (@carlykimmel I think we get 5 years of this. It ends in 7th grade, just like my grandmother promised me. (Uncertainty)

(2) @chelseaolson3 @andygrammer I did find this but haven't used it yet. (Contrariety)

(3) I must apologise straight away for leaving the question mark off the end of my previous tweet. (Necessity)

(4) Damn it! If I go one week without seeing game of thrones I have to start from the beginning again. (Hypotheticality)

These tweets are examples in which both annotators agreed on their label attribution. In these categories, the best inter-annotator agreement score was achieved (see Table 5). The high number of tweets annotated as ‘no label’ is of great interest as well. It turns out that this was the largest category of the dataset with more than half of the data annotated as ‘no label’. The annotators faced several challenges during the annotation of the data that led them to attribute this label to different reasons: the presence of symbols and/or special characters that made tweets difficult to comprehend, tweets consisting only of a hyperlink, incomprehensible abbreviations, slang language, the absence of enough context and the absence of any stance category in many cases. Additionally, the selected texts cover a wide thematic range, where stance-taking may not always be among the main communicative purpose of the tweeter. This contrasts with the BBC, where the discussion about a controversial political matter invites people to express their stance in a bolder manner. Therefore, identifying stance in *Twitter* data creates more noise in our corpus, with content that cannot be grouped under the predefined stance categories.

A closer look at the ‘no label’ data confirms the diversity of this category and various patterns may be observed. According to our guidelines, tweets that express neutral and stance-free statements should be in this category. This type of tweets is frequent, and two examples of neutral and stance-free tweets are shown in (5)–(6).

(5) The economy added 280,000 jobs in May marking 63 consecutive months of private-sector job growth.

(6) FYI I just sat down to google “how to use pomade” and somehow tweeted that.

In these examples, tweeters either make a neutral statement or describe aspects of their lives without taking any stance. More specifically, the authors share neutral information probably derived from news sources (cf. 5) or describe their everyday experiences (cf. 6). In many cases, the tweets are narrated in a confessional/emotional tone to forge connection with their followers, bond with them and/or increase their network. Tweets in which gratitude, love and wishes are expressed by public figures to their followers were quite frequent in our data. Some examples are provided in (7)–(9).

(7) That was fun times #moa!!!! love you mean it Minneapolis!!! @mallofamerica

(8) That plane saga made my night. Happy thanksgiving to you & yours! rt @briankoppelman happy tg. have u been following @theyearofelan tonight?

(9) Happy Valentines Day! May you love, be loved and make love, all in excess!

These examples illustrate tweets expressing sentiments, such as gratitude, love, appreciation and enthusiasm that public figures express to their fans. This type of tweets usually creates interaction and followers respond with likes, retweets and replies. The initial tweet becomes more visible, while the author builds stronger ties with their followers and gets more followers. The follower, in turn, gets the chance to establish a 'real' connection with the public figure they admire. As a result, the public figure has a larger and more loyal audience to which self-branding and promoting strategies can be efficient, as shown in (10)–(12).

(10) We are in Orlando, fl @waltdisneyworld for an amazing event. social media moms celebration. I must have spoken well last year. I'm back again

(11) Looking for some new music for the weekend? check my #liveinthefuture top 10 chart at beatport *HYPERLINK*

(12) Seriously the best cafe in California is @cafegratitudevb [...] if you haven't already tried it you need?? *HYPERLINK*

Examples serving the purpose mentioned above can also be attested and are annotated with a stance label, but most tweets related to promoting and self-branding content or providing advice about health and lifestyle choices were grouped in the 'no label' category, even in cases where indications of stance could be detected, as in (13) and (14).

(13) Free tickets, a free round-trip flight and free swag? What else could a steelers fan ask for?! Enter here!

(14) Be strong & courageous. do not be terrified or discouraged, for the lord your god will be with you wherever you go. -josh 1:9 be #blessed !!

In (13), the public figure urges their followers to join a competition to win free tickets, while in (14) lifestyle/religious advice is given. In both examples, the authors recommend their followers about specific choices, and the necessity label could be used in both tweets, but due to their overall meaning, the utterances, do not only express necessity. More specifically, in (13), the text includes two questions, with the second question also expressing uncertainty, and it ends up with an exhortation to the followers to join the competition. In (14), adding to the recommendation expressed in the imperative, we can also identify prediction (*…the lord will be with you…*) and source of knowledge (*-josh 1:9*). The co-occurrence of different types of stances in the same text was already observed in the analysis of the BBC (Simaki *et al.* 2020) and is also a frequent phenomenon in the present data. In (15)–(17), this co-occurring pattern may be observed.

(15) **Need** to sleep **but** my stupid brain won't shut offffff. Hummmblfukkdstfjff *HYPERLINK*. (Necessity and contrariety)

(16) #morningjah "you change **if** you change from babylon to rasta, **but** you can't change from rasta to anything". (cont) *HYPERLINK*. (Hypotheticality and contrariety)

(17) **Don't know if** you guys saw @hitrecordjoe be one of the first to do it **but** he did @nickiminaj like nobody's biz. go!x *HYPERLINK*. (Uncertainty, hypotheticality and contrariety)

Other patterns of tweets characterised as 'no label' are texts with stance-taking but, since the text (or part of it) is a question, they have been excluded. Some examples are provided in (18)–(20).

(18) I'm sorry but did you see my last post? My fans care about others in a manner I can't even begin to explain. proud. Let's change the world!

(19) Heartbreaker but the entire group is still alive. why not us? #ibelieve

(20) If obama's asia trip wasn't suspicious, why are all of his meetings taking place while Americans are asleep? *HYPERLINK*

Finally, new constructions related to stance were attested in this category. We identified various patterns of commonly used expressions that can be associated to our stance categories or form new ones. For instance, an interesting pattern is the 'not sure' construction that can be aligned with the uncertainty category, but it frequently co-occurs with other stances or is part of a question. While it certainly evokes a sense of uncertainty to the whole text, this pattern made the annotators doubt as regards the strength of the 'not sure' construction in dominating the overall meaning of the text, especially when other stances could also be identified. As a result, such cases were annotated as 'no label'. Some examples of this pattern are provided in (21)–(23).

(21) @jh0ps maybe...maybe not....probably maybe tho...but, maybe not also...dunno...could have...not sure...:):) (say hi next time!)

(22) @andavis1 college Wasn't right for me. Not sure what you mean about venture capital but a Boston based firm, spark, invested in jelly.

(23) I do. Not sure if I'm allowed to tell my prediction. I will check with NBC. Back later. RT @lolabeauty33 Any favorite acts from yesterday???

In (21), the tweet is illegible and not clear, while the contrariety marker *but* is present. In (22)–(23) other stances and sentiments may also be identified. Constructions, such as the 'not sure' construction, are strongly related to stance and should be studied in depth as they can enrich not only our stance markers list, but also our stance categories. For instance, and in contrast to the 'not sure' construction, the certainty category, for which we identified examples of the 'I'm sure' construction in the data, could be added. Examples are provided in (24)–(25).

(24) @1nataliemaines: I'm sure this haircut will be coming back around any day now. I think you should have it now.

(25) nicert @stephpalmer15: @mooremaya are you planning on coming to #passion2013? I'm sure @lecrae would share the stage!

Nevertheless, we also need to address the strength of the certainty that the 'I'm sure' construction carries in connection to the occurrence of other stances: prediction and necessity in (24) or the presence of the question in (25). In the present study, we have not addressed the issue of the presence of expressions of different stances in the same text, as we focus on the annotation process, the observation and the analysis of the results.

Overall, the annotation results mostly confirm the validity of our stance framework: the six stance categories tested here are attested in the *Twitter* data and, especially for the cases of uncertainty and contrariety, we observed a moderate but acceptable level of inter-annotator agreement. The overall agreement score (0.42) highlights the challenges of the stance annotation of the *Twitter* data, but also the potential of such a task. Sentiments are also present in the data, and constructions related to sentiment can be identified for a more in-depth linguistic analysis of the data.

In addition to the analysis of the annotation results, we studied whether the stance label which was attributed by the annotators to each text corresponded to the stance that, according to Table 1, the marker that was present in this text indicated. The goal of this task was first to test whether the presence of each of the stance markers provides a robust indication for the overall meaning (stance-taking or not) of the text and, second, to confirm whether each selected marker was perceived as related to a specific stance category by the annotators. Our hypothesis has been that the frequency of the stance types in the data does not only reflect the way that *Twitter* authors position themselves in their text, but it is also linked to the selection of the data, which is based on the list of predefined stance markers described in Section 4. This list includes a range of stance markers from relatively clear ones (*but* and *if*) to items that do not refer to a specific stance category (*that*, *show* and *think*). We compared the annotation labels that the annotators attributed to the stance marker according to which text was selected to be part of the dataset. We also investigated whether the stance category related to each of the 21 stance markers coincided with the annotation label that the annotators decided to attribute to the tweet: for instance, are texts in which *as* and *but* (markers for source of knowledge and contrariety respectively) annotated as source of knowledge and contrariety? We then associated the annotations to the stance markers and the results are shown in Figure 1 below.

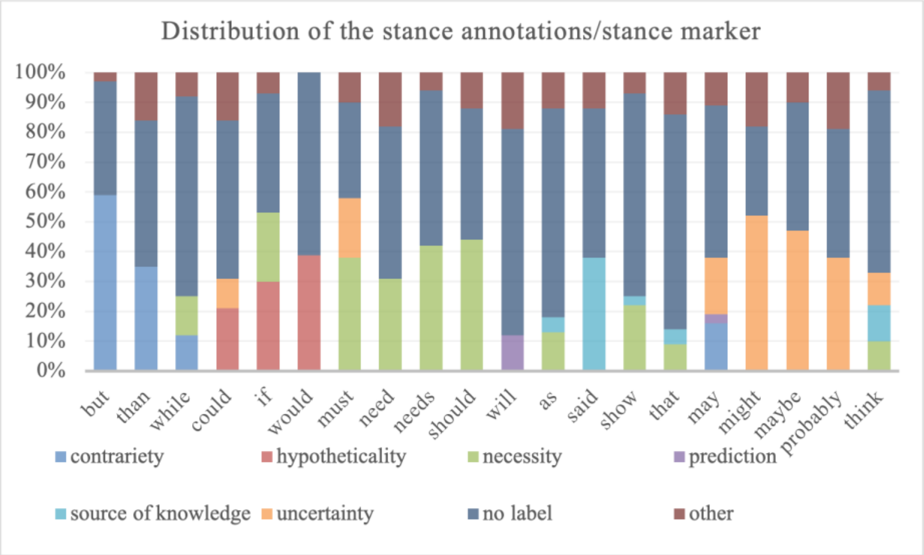

Figure 1: The percentage distribution of annotations per texts in which the same stance marker is present

Figure 1 shows how, in terms of frequency, the six stance categories and the ‘no label’ category are attributed to the data. For each marker subset, we assumed that the distribution of the labels would reflect the correlation of the annotations to the corresponding stance marker which, in turn, is associated with a given stance category. For instance, in the case of *but* (that may be a marker of contrariety), almost 60 per cent of the extracted data, in which *but* was present, was annotated as contrariety. The results here show that, to an extent, the well-established stance markers (*but*, *if* type) were annotated with the stance category to which they are strongly related (*but* with contrariety and *if* with hypotheticality), despite the high percentages of the ‘no label’ category in all subsets. Words that are less evident as markers of a specific stance, such as *as*, *that*, *think* and *will* are more rarely annotated with the anticipated stance label. For instance, *as* was identified as a marker for source of knowledge in Simaki *et al.* (2019), but, in the present study, texts containing *as* are mostly annotated (70%) as ‘no label’, and also as necessity (13%). Only 5 per cent of these texts were annotated as source of knowledge. This shows that the same marker is used to express different stance. Another example of multifunctionality and maybe text type/genre sensitivity is *show*, which in the BBC was used in source of knowledge constructions as a verb (e.g., *The data from the study show that…*). In the present data, it mostly refers to artistic performances and, despite the large number of ‘no label’ cases, it is frequently attested in necessity texts, where public figures encourage/urge their followers (e.g., *Dear everyone in Perth, u must see this show. It won best cabaret last year & it opens tomorrow for one week only!*). This phenomenon is due to the different types of content, topics and text types of both corpora, so we can assume that the same form that is identified as stance marker in one dataset does not work in the same way in a different dataset. Overall, the results in Figure 1 provide interesting insights about the validity of these stance markers when tested in a different dataset.

7. Stance annotation methodological considerations

Most of the discussion since the first annotation round of the BBC in Simaki *et al.* (2020) has been about the challenging nature of the stance annotation task. Difficulties were inevitable due to most of the framework’s categories and the nature of the BBC (limited size and duplicates due to stance co-occurrences). Likewise, all efforts of quantitative and computational tasks within the given setting have possibly resulted in overfitting stance-related markers as only the BBC was tested. Thus, it becomes even more challenging to evaluate the findings from those studies in a different setting, and the results provided in Table 5 confirm such a challenge.

Similarly, less encouraging results can still be discussed, and important insights with a potential to address them in future studies can be achieved. Methodological questions can be raised and method-related issues problematised as to the effectiveness of our annotation protocol in applying a notional stance scheme. In this study, the weak inter-annotator agreement made authors reflect on the annotation results, and more specifically on the metric used (Cohen's kappa) and how suitable this metric is for the task. For this purpose, after an extensive bibliographic search, we have concluded that there is no consensus on which metric is the most appropriate one for calculation of inter-coder reliability, despite all efforts to develop reliable metrics and tools. There are several recommendations for Cohen's kappa, which is among the most frequently used metric. A common issue that arises when calculating the reliability of annotated data in a scheme with more than two labels is that infrequent categories emerge from the annotation process, which leads to an uneven distribution of categories that produces unbalanced datasets, and subsequently leads to a lower reliability score (McHugh 2012). This is a common phenomenon in discourse annotation studies where similar distributions of categories between different types of discourse are not always feasible (Hoek and Scholman 2017).

The frequency of a specific label is due to the frequency of the type of relation it refers to (e.g., condition, reason, opposition, etc.). Discourse is also characterised by an uneven distribution of connective constructions that mark the various relations and link the different parts of a sentence. Some of these connectors are very frequent while others are less frequent, and the distribution of relation types that specific connectives mark may also vary. As Hoek and Scholman (2017: 1) state,

> annotators tend to agree more when annotating explicit coherence relations, which are signalled by a connector or cue phrase (*because, for this reason*) than when annotating implicit coherence relations, which contain no or less linguistic markers on which annotators can base their decision.

This is important, as such markers not only are explicitly mentioned, but they are also less prone to ambiguity, so they cannot easily be interpreted in an ambiguous way. In the present study, this has been confirmed in the three most frequent categories in the annotated data (contrariety, necessity and uncertainty). Markers that signal the corresponding stance type, such as *but, need, must* and *might,* occur frequently. The annotators could identify and label those markers that were explicitly associated to these stances and, as a result, the highest degree of agreement scores was attested. In some of the other categories, we can argue that due to the lower prevalence of stance-related items, there is insufficient information in the data, not only for the annotators to make decisions in a structured and homogeneous way based on discriminate factors, but also for us to assess the annotators' ability. As a consequence, kappa may underestimate the true agreement (Hripcsak and Heitjan 2002). An interesting case is the hypotheticality category which, despite the highly discriminative item *if,* is about half as frequent as the contrariety or the uncertainty categories and shows a lower level of agreement (0.35).

These issues can influence a reliability measurement such as a kappa score, which seems to be very sensitive to typical characteristics of discourse data, such as the ones mentioned above. In those cases, the kappa paradox is attested (Feinstein and Cicchetti 1990); in other words, the values are sometimes relatively low, despite the high percentage of observed agreement. We considered the agreement percentage as an alternative measurement that is easy to calculate and interpret but, as Lombard *et al.* (2002) argue, it fails to account for agreement that occurs by chance. Instead, as shown in Table 6 (see Section 5.3), we used two other metrics to calculate the interrater reliability: 1) the K alpha (Krippendorff 2011), which is also a standard metric, and 2) the relatively new AC_1 measure (Gwet 2002), which is used to solve some of the problems in the Cohen's kappa. This metric estimates the agreements between annotators as they are not partly due to chance (expected agreement) and it is less affected by the prevalence of categories and the marginal probability than the Cohen's kappa. AC_1 shows a higher score (see Table 6), which is encouraging for future research and suggests that it can be an important alternative measure when it comes to the calculation of the interrater reliability of discourse data.

The calculations of the inter-annotator agreement raised another important question about the interpretation of the results: What can be considered as an acceptable level of reliability? Does 0.58 here indicate that our set of annotated data is a reliable value for replication and usability purposes? The answer to these questions is problematic and, as Neuendorf (2017: 168) summarises, there are no established standards and "coefficients that account for chance (e.g., Cohen's kappa) of .80 or greater would be acceptable to all, .60 or greater would be acceptable in most situations and, below that, there exists disagreement”. Landis and Koch (1977) suggest a scale for the interpretation of kappa scores that was originally designed for the medical field: 0.41–0.60 values signal a moderate agreement, 0.61–0.80 substantial agreement and 0.81–1 perfect agreement. Poesio (2004) suggests 0.80 as a threshold that ensures an annotation of reasonable quality. McHugh (2012) states that kappa is not very well supported for factors such as rater independence, which lowers the estimate of agreement excessively. In addition, the fact that this metric cannot be directly interpreted leads researchers to accept lower kappa values. When it comes to the publication of annotated corpora, Artstein and Poesio (2008) argue that setting a specific agreement threshold should not be a prerequisite as long as a detailed report on data collection methodology, statistical significance of agreement and agreement table are included in the data description. Our opinion, which is based on the experience with different discourse annotation tasks, is in line with Artstein’s and Poesio’s (2008). We agree that interrater reliability is an important indication of the quality of annotated data and that it is important to use such measurements, but it is always worthwhile taking the analysis a step beyond the interpretation of the agreement score and, in doing so, draw more insightful conclusions.

8. CONCLUSION

In this study, our goal has been to evaluate 1) Simaki *et al.*’s (2020) stance framework and 2) the suitability of our categories in a different social media text type. We selected *Twitter* texts in which at least one stance marker from a predefined list was present. The data were annotated by two annotators and the inter-annotator agreement score was calculated. The findings show that taking stance differs across different social media platforms and that the stance categories, which appear to be salient in blogs, are less salient in *Twitter*. Our prior experience with stance annotation showed that it is not possible to identify stance in every text since people use *Twitter* (and social media in general) for different communicative purposes. Thus, many tweets can be stance-free in the sense of expressing sentiments or asking for information, while other issues, such as ambiguity or just illegible content, may still be present. For this reason, we created an additional category to cater for all tweets that did not conform to any of the six stance categories. According to the annotation results, this ‘no label’ category emerged as the largest one, which made us question the suitability of the stance categories for *Twitter* data. However, a closer look to the data made us realise that it provides an excellent benchmark to further explore and develop this framework. A refined annotation protocol, a more cautious filtering of the data and adjustments in the existing framework are likely to lead to more efficient annotation and more reliable data. An alternative approach can also be considered, namely, to address the disagreements in the annotations and resolve the conflicting cases. Overall, our study confirms that the stance categories analysed here can be identified and attributed in *Twitter* data, despite the challenges of the nature of the task. In a follow-up study, sentiments, functions such as self-branding and new stances or other phenomena can be incorporated to the framework. In addition, the emerging patterns in the 'no label' category can be further analysed, and new categories can be considered to enrich the existing ones. In this study, both annotators devoted a large amount of time to a laborious cognitive task. Especially relevant has been the annotators' fatigue due to the manual task and its possible effects on the annotations, dealing with the very complex concept that stance is. As for the quantitative analysis, the statistical outcome has not been as rewarding as we had hoped. Should these results determine the reliability of the task, or is there room to derive important insights about stance-taking on *Twitter* data? The relatively low interrater agreement highlighted challenges related to the nature of the task, the categories of the framework and the text type, but it also pointed to methodological issues discussed in relation to our results and to the literature. We believe that the annotation protocol and our annotations will be a good basis for future studies since there are no duplicates in our set (all tweets have only one label), and this will be helpful in the replicability of the study. Finally, our annotated data can also be used in computational tasks such as stance detection and classification tasks.

References

AlDayel, Abeer and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58/4: 102597. https://doi.org/10.1016/j.ipm.2021.102597

Alonso Ameida, Francisco. 2015. Introduction to stance language. *Research in Corpus Linguistics* 3: 1–5.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34/4: 555–596.

Berman, Ruth, Hrafnhildur Ragnarsdóttir and Sven Strömqvist. 2002. Discourse stance: Written and spoken language. *Written Language & Literacy* 5/2: 255–289.

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

Boyd, Danah, Scott Golder and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on *Twitter*. *Proceedings of the 43rd Hawaii International*

Conference on System Sciences. Washington: IEEE Computer Society, 1–10. https://ieeexplore.ieee.org/document/5428313

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20/1: 37–46.

Du Bois, John W. 2007. The stance triangle. In Robert Englebretson ed. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: John Benjamins, 139–182.

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

Facchinetti, Roberta, Frank Palmer and Manfred Krug. 2003. *Modality in Contemporary English*. 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. In William Eberle and Chutima Boonthum-Denecke eds. *Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference*. Florida: Association for the Advancement of Artificial Intelligence, 174–179.

Feinstein, Alvan R. and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43/6: 543–549.

Ferreira, William and Andreas Vlachos. 2016. Emergent: A novel data-set for stance classification. In Kevin Knight, Ani Nenkova and Owen Rambow eds. *Proceedings of the Association for Computational Linguistics: Human Language Technologies*, 1163–1168. https://aclanthology.org/N16-1138/

Fuoli, Matteo. 2018. A stepwise method for annotating APPRAISAL. *Functions of Language* 25/2: 229–258.

Ghosh, Shalmoli, Prajwal Singhania, Siddharth Singh, Koustav Rudra and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato and Nicola Ferro eds. *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. Cham: Springer, 75–87.

Gwet, Kilem. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment* 1: 1–5.

Hasan, Kazi Saidul and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In Alessandro Moschitti, Bo Pang and Walter Daelemans eds. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha: Association for Computational Linguistics, 751–762.

Hernández, Nuria. 2014. New media, new challenges: Exploring the frontiers of corpus linguistics in the linguistics curriculum. *Research in Corpus Linguistics* 1: 17–31.

Hidalgo-Downing, Laura. 2012. Grammar and evaluation. *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal1471

Hoek, Jet and Merel Scholman. 2017. Evaluating discourse annotation: Some recent insights and new approaches. In Harry Bunt ed. *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. Tilburg: Tilburg University, 1–13. https://aclanthology.org/W17-7401/

Honey, Courtenay and Susan C. Herring. 2009. Beyond microblogging: Conversation and collaboration via *Twitter*. *Proceedings of the 42nd Hawaii International Conference on System Sciences*. Waikoloa: IEEE Computer Society, 1–10. https://ieeexplore.ieee.org/document/4755499

Hripcsak, George and Daniel F. Heitjan. 2002. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics* 35/2: 99–110.

Hunston, Susan and Geoffrey Thompson. 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.

Jacknick, Christine M. and Sharon Avni. 2017. Shalom, bitches: Epistemic stance and identity work in an anonymous online forum. *Discourse, Context & Media* 15: 54–64.

Jaffe, Alexandra. 2009. *Stance: Sociolinguistic Perspectives*. Oxford: Oxford University Press.

Kaltenböck, Gunther, María José López-Couso and Belén Méndez-Naya. 2020. The dynamics of stance constructions. *Language Sciences* 82: 101330. https://doi.org/10.1016/j.langsci.2020.101330

Krippendorff, Klaus. 2011. *Computing Krippendorff's Alpha-reliability*. https://repository.upenn.edu/asc_papers/43. (24 November, 2022.)

Kucher, Kostiantyn, Andreas Kerren, Carita Paradis and Magnus Sahlgren. 2016. Visual analysis of text annotations for stance classification with ALVA. In Tobias Isenberg and Filip Sadlo eds. *Proccedings of the Eurographics Conference on Vizualization*, 49–51. http://dx.doi.org/10.2312/eurp.20161139

Küçük, Dilek and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys* 53/1: 1–37.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33/1: 159–174.

Lombard, Matthew, Jennifer Snyder-Duch and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research* 28/4: 587–604.

Marín-Arrese, Juana I. 2017. Stancetaking and inter/subjectivity in journalistic discourse: The engagement system revisited. In Ruth Breeze and Inés Olza eds. *Evaluation in Media Discourse: European Perspectives*. Bern: Peter Lang, 21–48.

Marín-Arrese, Juana I., Marta Carretero, Jorge Arús Hita and Johan Van der Auwera eds. 2014. *English Modality: Core, Periphery and Evidentiality*. Berlin: Mouton de Gruyter.

McHugh, Mary L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22/3: 276–282.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov and Tortsten Zesch eds. *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego: Association for Computational Linguistics, 31–41. https://aclanthology.org/S16-1003/

Neuendorf, Kimberly. 2017. *The Content Analysis Guidebook*. Thousand Oaks: SAGE publications.

Pamungkas, Endang Wahyu, Valerio Basile and Viviana Patti. 2019. Stance classification for rumour analysis in *Twitter*: Exploiting affective information and conversation structure. *arXiv preprint arXiv: 1901.01911*. https://doi.org/10.48550/arXiv.1901.01911

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

Perrin, Daniel. 2012. Stancing: Strategies of entextualizing stance in newswriting. *Discourse, Context & Media* 1/2–3: 135–147.

Poesio, Massimo. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In Bonnie Webber and Donnna Byron eds. *Proceedings of the Workshop on Discourse Annotation*. Barcelona: Association for Computational Linguistics, 72–79.

Schmidt, Jan-Hinrik. 2014. Twitter and the rise of personal publics. In Katrin Weller, Alex Bruns, Jean Burgess, Merja Mahrt and Cornelius Puschmann eds. *Twitter and Society*. Bern: Peter Lang, 3–14.

Simaki, Vasiliki, Carita Paradis, Panagiotis Simakis and Andreas Kerren. 2017a. Stance classification in texts from blogs on the 2016 British referendum. In Alexey Karpov, Rodmonga Potapova and Losif Mporas eds. *Proceedings of the 19th Speech and Computer International Conference.* Charm: Springer, 700–709.

Simaki, Vasiliki, Carita Paradis and Andreas Kerren. 2017b. Identifying the authors' national variety of English in social media texts. In Ruslan Mitokov and Galia Angelova eds. *Proceedings of the Recent Advances in Natural Language Processing Conference*, 700–709. https://acl-bg.org/proceedings/2017/RANLP%202017/pdf/RANLP086.pdf

Simaki, Vasiliki, Carita Paradis and Andreas Kerren. 2018a. Evaluating stance-annotated sentences from the *Brexit Blog Corpus*: A quantitative linguistic analysis. *ICAME Journal* 42: 133–165.

Simaki, Vasiliki, Panagiotis Simakis, Carita Paradis and Andreas Kerren. 2018b. Detection of stance-related characteristics in social media text. In Nikos Fakotakis and Vasileios Megalooikonomou eds. *Proceeding of the 10th Hellenic Conference on Artificial Intelligence*. Patras: Association for Computing Machinery, 1–7. https://doi.org/10.1145/3200947.3201017

Simaki, Vasiliki, Carita Paradis and Andreas Kerren. 2019. A two-step procedure to identify lexical elements of stance constructions in discourse from political blogs. *Corpora* 14/3: 379–405.

Simaki, Vasiliki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher and Andreas Kerren. 2020. Annotating speaker stance in discourse: The *Brexit Blog Corpus*. *Corpus Linguistics and Linguistic Theory* 16/2: 215–248.

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

Traugott, Elizabeth Closs. 2020. Expressions of stance-to-text: Discourse management markers as stance markers. *Language Sciences* 82: 101329. https://doi.org/10.1016/j.langsci.2020.101329

Verhagen, Arie. 2005. *Constructions of Intersubjectivity: Discourse, Syntax, and Cognition*. Oxford: Oxford University Press.

Yus, Francisco. 2011. *Cyberpragmatics: Internet-mediated Communication in Context*. Amsterdan: John Benjamins.

Yus, Francisco. 2016. Discourse, contextualization and identity shaping: The case of social networking sites and virtual worlds. In María Luisa Carrió-Pastor ed. *Technology Implementation in Second Language Teaching and Translation Studies*. Singapore: Springer, 71–88.

Zappavigna, Michele. 2012. *Discourse of Twitter and Social Media: How we Use Language to Create Affiliation on the Web*. London: A&C Black.

Zappavigna, Michele. 2015. Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25/3: 274–291.

Zappavigna, Michele and James R. Martin. 2018. # Communing affiliation: Social tagging as a resource for aligning around values in social media. *Discourse, Context & Media* 22: 4–12.

Zhu, Hongqiang. 2016. Searchable talk as discourse practice on the internet: The case of "# bindersfullofwomen." *Discourse, Context & Media* 12: 87–98.

Corresponding author
Vasiliki Simaki
Lund University
Faculties of Humanities and Theology
Centre for Languages and Literature
Helgonabacken 12
Box 201, SE 221 00
Lund
Sweden
E-mail: vasiliki.simaki@englund.lu.se

received: April 2022
accepted: November 2022
published online: December 2022

Appendices

Appendix 1: The framework's text stance categories in alphabetical order, followed by a brief description and examples (Simaki *et al.* 2020).

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.*
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.*
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.*
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.*
Necessity	The speaker expresses a request, recommendation, instruction or an obligation.	*I must hand back all the books by tomorrow.* *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.*
Tact/rudeness	The speaker expresses pleasantries and unpleasantries.	*Please, do give my love to him.* *You lazy bastard. Get lost.*
Uncertainty	The speaker expresses doubt as to the likelihood or truth of what she or he is saying.	*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.	*I wish I could join you next summer.* *I prefer to stay in a cheap hotel.*

Appendix 2: Full list of stance markers for each stance category based on Simaki *et al.* (2019).

Contrariety	**Hypotheticality**	**Necessity**	**Prediction**	**Source of knowledge**	**Uncertainty**
And	*A*	*Have*	*Be*	*As*	*Could*
But	*Be*	*Let*	*May*	*Has*	*I*
Not	*Could*	*Must*	*Not*	*I*	*May*
Than	*If*	*Need*	*Is*	*Said*	*Maybe*
While	*In*	*Needs*	*It*	*Show*	*Might*
	Then	*Should*	*The*	*That*	*Probably*
	Will	*To*	*To*	*The*	*Think*
	Would	*We*	*Will*	*To*	