Function-to-form mapping in corpora: Historical corpus pragmatics and the study of stance expressions

Daniela Landert

University of Zurich

Abstract

This paper presents a new approach to the study of pragmatic functions in corpora. In contrast to forms, functions cannot be retrieved automatically, which makes function-to-form approaches notoriously difficult. Exploiting the fact that pragmatic functions are not evenly distributed across corpora, but instead tend to co-occur, this study shows that it is possible to retrieve text passages from corpora that are particularly suitable and relevant for a detailed qualitative analysis. This method is demonstrated with the example of stance expressions. A set of 20 lexical items that are used to express epistemic and evidential stance were tagged in four Early Modern English corpora. A sample of 300-word passages with a high density of the tagged lexical items was then analysed manually to illustrate the kinds of observations that can be made based on such data. The findings show that this method can lead to new insights into stance marking in Early Modern English. For instance, it can be used to identify previously unstudied stance markers, (con)textual factors that deserve further investigation, and problems for the interpretation of quantitative findings. As such, the method complements existing approaches to the study of stance in corpora.

Keywords: Corpus pragmatics; historical pragmatics; epistemicity; evidentiality; Early Modern English; pragmatic functions

1. Introduction

Over the last two decades, an increasing number of corpora have become available to linguists, and more are in the process of being developed. These corpora cover an expanding range of language data and they keep growing in size. Such resources have provided linguists with exciting new research opportunities. Approaches that are mainly quantitative in nature benefit especially from this development. Larger amounts of data can lead to more reliable results, and data from different periods, varieties, text types and genres allow for new comparisons. Where scalable methods for automated linguistic analysis are available, they can be applied to larger corpora without leading to an increase in the amount of work that is proportional to the size of

the corpus. As long as no manual preparation and analysis of the data is required, the only limitations to the size of corpora that can be handled are posed by the technical infrastructure, such as computer memories and processors.

For the field of pragmatics scalability is the exception rather than the norm, though. Pragmatic meaning depends on context and, thus, most pragmatic studies require a considerable degree of manual analysis and interpretation. The detailed analysis of individual examples in their specific micro and macro context is essential to complement the quantitative results from large-scale evaluations of specific formal patterns (Rühlemann & Aijmer 2014, 3). This is perhaps even more the case in historical pragmatics, where information about the wider socio-cultural context is largely derived from textual evidence, and where variation of spelling and punctuation poses additional problems for automated analyses. Despite such problems, (historical) corpus pragmatics is a thriving field, as can be seen from the number of collected volumes and handbooks that have been published recently (e.g. Aijmer & Rühlemann 2014; Romero-Trillo 2008; Suhr & Taavitsainen 2012; Taavitsainen, Jucker & Tuominen 2014). Many studies deal with the challenges by investigating the pragmatic functions of a small number of fixed forms (form-to-function mapping), for instance in the study of discourse markers (for a recent overview, see Aijmer 2014). The forms can be retrieved automatically and their functions are then analysed manually. Typically, the automatic retrieval can be applied to new corpora without too many difficulties, but of course the manual analysis requires more time with increasing corpus size and a correspondingly increasing number of hits. Exploiting the newly available data tends to be even more challenging for studies that take as their starting point a pragmatic function, which is then investigated in its various forms of realisation (function-to-form mapping). Here the problem lies in the fact that functions cannot be identified automatically in corpora. As a consequence, function-to-form mapping requires more manual analysis and is often restricted to smaller corpora (see, for instance, Kohnen 2007, 2008).

From the point of view of (historical) pragmatics, the question arises how we can reconcile the need for detailed qualitative analysis with the increasing size of the available data. It is clear, for instance, that new historical corpora include many new, revealing examples of language use in earlier periods, but it is far from clear how such instances can be retrieved, especially if we are interested in realisations of pragmatic functions in forms that have not yet been described systematically. In addition, if detailed manual analysis of text passages is needed for the study of pragmatic functions, it would be desirable to have ways of automatically identifying text passages that are likely to be relevant for the analysis. Pragmatic functions tend to be distributed very unevenly across texts, which means that the study of a random selection of text passages is often a very inefficient approach. In

contrast, selecting texts from a specific context (e.g. genre) for which a high presence of a certain pragmatic function is expected results in insights that are restricted to this context; uses in contexts that do not correspond to our expectations are missed. At the moment, the existing methods provide only limited options for improving the qualitative analysis of corpus data and developing new methods is thus highly desirable.

In this paper I suggest one such approach. With the example of stance expressions, I show that it is possible to automatically retrieve text passages from corpora that are highly relevant for the analysis of stance. These passages include very explicit stance marking and they make it possible to see how conventionalised stance markers interact with other, non-conventionalised and context-dependent expressions of stance. They can also lead to new insights and starting points for further studies, for instance concerning the contexts in which stance is marked and the forms in which stance marking is realised. As a consequence, the analysis of such passages can complement more quantitatively-oriented studies of stance markers.

I will start by presenting some theoretical background on stance marking in Early Modern English and on function-to-form mapping approaches in pragmatics. In Section 3 I will briefly present the corpora that I used and then proceed to a more detailed discussion of my method. Section 4 will give an overview of the retrieved passages and in Section 5 I will present a few of these passages and give some examples of the kinds of observations that can be made on the basis of such data.

2. Theoretical background

2.1 Epistemic and evidential stance in Early Modern English

Markers of epistemic and evidential stance are expressions that reflect the speaker's or writer's attitude towards the certainty, reliability and source of information of their statements. There are various ways in which such stances can be expressed. Verbs like *think, believe,* and *seem* – often controlling *to*– or *that* complement clauses – are among the most commonly used linguistic resources for expressing stance in Early Modern English (see Example 1). Stance nouns and adjectives can likewise control complement clauses in constructions like *the fact that, it is certain that,* and *I am sure that*, and stance adverbs like *certainly* and *possibly* play an important role, too. In addition, stance can be expressed in ways that depend on interpretation in context, such as in Example 2 below in which the conditional clause *If my memory faile me not extreamly* expresses a very high degree of certainty. Insertions, like *at lest to my knowledge* in Example 3, are another way in which stance can be expressed.

1) *I think* shee liveth in Guildhall street .

   (PCEEC, Browne 010, 1678)

2) *If my memory faile me not extreamly,* it is taken out of two places over each eye;

   (EMEMT, Philosophical Transactions, Volume 3, Number 36, 1668)

3) It is no wonder indeed if our Barrister should be unapprized of Brereton's Case, it not

   being (*at least to my knowledge*) in Print;

   (Lampeter, Law, "Remarks on the trial of John-Peter Zenger", 1738)

There have been three main approaches to the study of stance expressions in Early Modern English. By far the largest body of research is devoted to the detailed study of individual stance markers and their development over time. The stance marker *I think* has attracted a great deal of attention (e.g. Aijmer 1997; Palander-Collin 1999), and other studies have dealt with groups of verbs, such as perception verbs (Whitt 2010, 2011), verbs of appearance (Gisborne & Holmes 2007), verbs of knowing (Hiltunen & Tyrkkö 2011), and verbs that share certain semantic properties (Bromhead 2009). In addition, some studies have focused on the development of specific structural patterns that are related to stance (e.g. Brinton 2008; López-Couso & Méndez-Naya 2014, 2015). These studies provide a very rich picture of the use and diachronic development of stance markers, including their grammaticalisation and pragmaticalisation. However, since they focus on a small number of markers, they can provide only limited insight into how stance was expressed overall.

This issue is addressed by the second group of studies, which aim to measure the frequency with which stance is expressed in different periods and different contexts (e.g. Biber 2004; Gray, Biber & Hiltunen 2011). These studies rely on the semi-automatic retrieval of conventionalised lexico-grammatical patterns that are frequently used to express stance, for instance stance verbs, nouns and adjectives followed by *that*-complement clauses. The restriction to specific grammatical patterns is necessary for distinguishing between stance and non-stance uses of lexical items. For instance, the verb *think* is a common stance marker in constructions like *I think that this is true*. In contrast, the verb does not express epistemic stance in constructions like *I'll think about your suggestion*. As a consequence, only instances are retrieved that are followed by the overt complementisers *that* and *to*, which introduce complement clauses – a pattern in which *think* consistently expresses stance. However, this means that patterns that cannot easily be retrieved are excluded, for instance stance verbs occurring with complement clauses without overt complementiser (see Example 1 above).

4

Likewise, more context-dependent stance expressions, like those given in Examples 2 and 3 above, cannot be included.

The third approach to studying stance in earlier periods of English can be found in studies that focus on the comprehensive analysis of stance markers in one particular genre. Examples are studies on medical writing (e.g. Alonso-Almeida & Mele-Marrero 2014; Taavitsainen 2000, 2001), witness depositions (Grund 2012, 2013) and, for Late Modern English, correspondence (Fitzmaurice 2003). Most of these studies focus on a limited amount of data which is analysed through detailed qualitative study of the texts. This limitation in the amount of data allows researchers to discuss all forms of stance marking, including non-conventionalised patterns and context-dependent expressions, and to produce rich interpretations of these markers. The results draw attention to the fact that context plays a very important role for when and how stance is expressed and what functions stance markers fulfil.

2.2 Function-to-form mapping

The search for linguistic forms that express a stance meaning is an instance of function-to-form mapping. One area in which function-to-form mapping has been applied is speech act analysis. Depending on the speech act, there are various ways how (semi-)automatic methods can be used to retrieve speech acts from corpora. Jucker (2013) describes three main approaches. First, for some speech acts, the search for illocutionary force indicating devices (IFIDS) can be used to retrieve relevant instances. For instance, the search for *sorry* retrieves apologies and the search for *please* retrieves requests. It is clear that both precision and recall are quite limited for this method. The search for *please* produces many hits that do not occur in requests (limited precision) and it misses all those requests that do not contain *please* (limited recall). Still, the method enables researchers to collect examples of speech acts, which can be used for further analysis. For the study of stance expressions, this method is not applicable, since there are no illocutionary force indicating devices that can be searched for.

The second method consists in the identification of fixed lexico-grammatical patterns that are typically used in a given speech act. For instance, Manes and Wolfson (1981) find that the most frequent way of paying compliments in American English takes the form of "NP *is/looks* (*really*) ADJ", accounting for slightly more than half of all the compliments they collected. Once such a pattern is identified, it can be used to retrieve realisations of the pattern from corpora – provided the corpora contain the necessary levels of annotation, in this case tagging for part of speech and possibly some level of syntactic annotation (see Jucker et al. 2008). This method is used in the studies by Biber

(2004) and Gray, Biber and Hiltunen (2011), who trace the distribution of stance markers by searching for common lexico-grammatical patterns.

The third approach to function-to-form mapping of speech acts is based on the identification of meta-communicative expressions. Jucker and Taavitsainen (2014) show that searching for the term *compliment* can retrieve passages that provide interesting insights into compliments of earlier periods. The term *compliment* is often used in situations in which people talk about compliments, for instance when negotiating whether or not an utterance should be considered a compliment. This means that many passages in which the term *compliment* is used contain (potential) compliments, and they can also reveal attitudes towards compliments. This method cannot be applied to the study of stance expressions, since there are no meta-communicative expressions that people use to refer to epistemic and evidential stance. What can be taken over from this approach, though, is the idea that it is possible to identify text passages in corpora that can be used for further qualitative analysis of a pragmatic phenomenon.

2.3 High-density passages, or the clustering of stance markers

In order to retrieve relevant passages for the detailed manual analysis of stance expressions it is necessary to identify formal characteristics that are typical of passages that can lead to rich insights into stance marking in Early Modern English. I started by reading through a sampler corpus in which I included texts from different sections of various Early Modern English corpora. When reading through the texts, it soon became clear that stance markers are not evenly distributed. There are many long text passages that do not contain any explicit stance markers, while there are other passages that are extremely rich in stance markers, such as Passages 4 to 6 below. These combine various conventionalised lexico-grammatical patterns like *I think, I am / it is certain, I am sure* and also include more context-dependent expressions (e.g. *it is not in the power of any own to overthrow [the hypothesis]*). Stance is expressed in a very explicit way in these passages, and the stance marking appears somewhat redundant from a present-day perspective.

4) Nay, undoubtedly I might very well beleeve you without swearing; for surely, there was then excessive Sports, Pastimes, & Revels, that it would make a bodies haire stand an end to heare and see it. (CED, Miscellaneous, "Women will have their will", 1648)

5) And I am so absolutely confirm'd in the truth of my Hypothesis, that I am sure it is not in the power of any one to overthrow it. Nay the Hypothesis on which I build, I am sure is true. (EMEMT, "Dialogue between Alkali and Acid", 1698)

6) Tis most Certaine, that our Emperour would have bin to mee rather a Jaylor then a husband, and tis as true that (though for my owne sake I think I should not make an ill wife to any body) I can not bee a good one to any, but one. I know not with what constancy you could heare the sentence of your Death, but I am certaine there is nothing I could not heare with more, and if your interest in mee bee dearer to you then your life, it must necessarily follow that tis dearer to mee then any thing in the worlde besydes, therfore you may bee sure I will preserve it with all my care. (PCEEC Osborne 028, 1653)

The main formal characteristic of these passages is that they contain clusters of stance markers or, put differently, that they have a high density of stance markers. Some of these markers contain lexical items that are closely associated with epistemic stance, such as *think, believe, sure, true, truth, truly* and *surely*, even though many instances are not used in the lexico-grammatical patterns that express stance consistently and that can be identified automatically (see Section 2.1). Since most of these lexical items can also be used with non-stance meanings, searching for all instances of the items does not provide reliable results on where and how frequently stance is expressed. However, searching for text passages in which many of these lexical items co-occur is very likely to produce passages in which stance is marked explicitly. The search does not need to be restricted to specific lexico-grammatical patterns, for even if some lexical items are used without stance meaning, it is quite unlikely that a number of lexical items often used to express stance occur together with non-stance meanings. And even if this is the case, the passage can simply be dismissed for further analysis, given that the focus lies on qualitative rather than quantitative interpretation.

3. Identifying passages with a high-density of stance markers

The analysis is based on four Early Modern English corpora, the Corpus of English Dialogues 1560–1760 (CED), the Early Modern English Medical Texts Corpus (EMEMT), the Lampeter Corpus of Early Modern English Tracts (LC) and the Parsed Corpus of Early English Correspondence (PCEEC).[1] Collectively, the four corpora include over 6 million words and they cover a range of different text types and genres (see Table 1). All four corpora were transformed into an XML representation that was specifically designed for this project.

---

[1] For the CED, the XML version was used. The corpus version without spelling normalisation was used for the EMEMT. For the PCEEC, the plain text version without syntactic annotation was used.

Table 1. Overview of corpora. Information based on Claridge (2003 [1999]); Kytö and Walker (2006); Taavitsainen and Pahta (2010); Taylor et al. (2006)

	CED	EMEMT	LC	PCEEC
period	1560-1760	1500-1700	1640-1740	1410-1681
intervals	40 years	–	10 years	70 years
word count	1.2 mio	2.0 mio	1.2 mio	2.2 mio
released	2006	2010	1999	2006
based on	speech-relatedness	text domain: medical writing	publication form: pamphlet	publication form: letter
text domain / text type / genre	drama comedy, didactic works, prose fiction, witness depositions, trial proceedings	medical writing	economy, law, politics, religion, science, miscellaneous	correspondence

First, the corpora were tagged for a set of 20 lexical items that are commonly used to express epistemic and evidential stance (see Table 2). The lexical items include verbs, adjectives and adverbs. They were selected based on the results of a pilot study in which I identified lexical items that are frequent within their part of speech category and which are reliable indicators of stance. To establish the frequency within the part of speech category, the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) was used. A set of 66 lexical items was then tagged in the four corpora, including all word forms and spelling variants. For each lexical item, a stratified sample of 40 instances was manually analysed to establish whether or not it was used to express stance. I adopted a broad definition of stance for this step, including also reported and attributed stance. The 20 items that were selected for the final tagging were all used to express stance in at least two thirds of all instances; for most of them, more than 80% of all instances were used with a stance-related meaning. Overall 42,988 instances of these lexical items were tagged in the four corpora.

Table 2. Lexical items tagged as stance markers

verbs	*believe*, doubt[1], know, perceive, seem*, suppose*, think**

adjectives	*confident, evident*, (un)likely*, manifest, (im)possible*, (im)probable*, (un)sure*
adverbs	*certainly*, perhaps*, plainly, surely*, truly*, verily**

* more than 80% of all instances have stance-related meaning

[1] includes instances of *doubt* as a noun; more than 80% of all instances of *doubt* (noun and verb) have stance-related meaning

In a second step, a Python script was run on the tagged corpora to identify text passages with a high number of tagged items. The script counts the number of tagged items in text windows of a fixed size. For the purpose of the present study, the window size was set to 300 words. The windows are overlapping in order to avoid missing high-density passages that occur at a window boundary. Corpora annotations are not included in the word count and windows cannot span across different files within a corpus. The script sorts the passages according to the number of tagged markers included in each and returns those *N* passages with the highest numbers.

4. **Overview of frequency of stance markers across windows**

Before I analyse the high-density passages in detail, I will provide a very general overview of the distribution of the tagged lexical items and the high-density passages across the four corpora. Figure 1 displays the frequency of core stance markers in 300-word windows, normalised per 1,000 windows. For instance, the high-density analysis of the Lampeter corpus produced 7,395 300-word windows, 1,795 of which do not contain any core stance markers. For the normalised frequency, this means that 243 out of 1,000 windows do not contain any stance markers. The frequency of windows with one stance marker is slightly higher, at 253 per 1,000 windows. Three of the four corpora have their peak at one stance marker per 300-word window. The only exception is the EMEMT, which contains almost twice as many windows with zero markers than with one marker. Overall, the CED and the PCEEC contain more windows with a high-density of stance markers than the Lampeter corpus and the EMEMT. This may be related to the higher degree of linguistic immediacy in dialogues and letters compared to (other forms of) mass media and scientific writing.
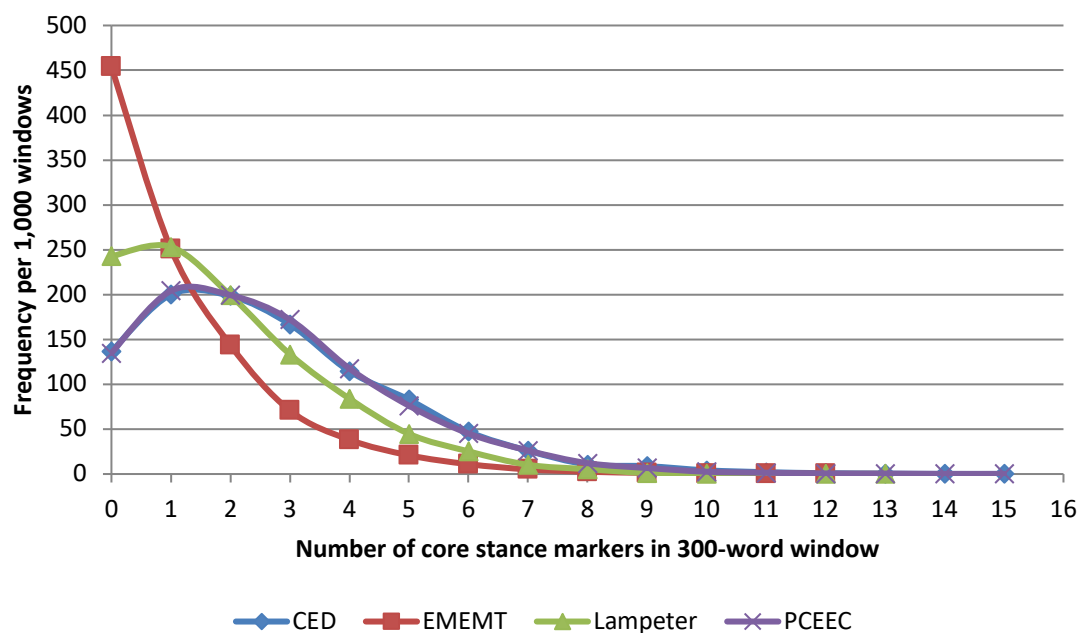
Figure 1. Overview of the density of core stance markers in 300-word windows across corpora

These results show that most of the passages from the four corpora contain a very low density of core stance markers, often no more than one or two tagged items per 300 words. However, all four corpora contain passages with ten or more tagged items within 300 words.

Due to the way in which the corpora were tagged, it cannot be assumed that each tagged item is a stance marker. However, a closer look at the passages with the highest density of stance markers shows that the large majority of these indeed includes very explicit marking of authorial stance. In a sample of 60 passages – the 15 passages from each corpus with the highest density of stance markers – only four passages do not contain explicit marking of authorial stance. One passage is from the trial proceedings section of the CED and it contains an exchange between the judge and a witness about whether or not another witness knew the defendant. Due to the topic of the exchange, the passage contains 13 instances of the verb *know* with third person and past reference (e.g. "He did not *know* him"), which do not express stance. One passage from the Lampeter corpus contains three instances of the formulation "things which a Christian ought to *know* and *believe* to his Soul's Health", together with some other instances of *know* and *believe* which are instructive rather than stance related. Another passage from the Lampeter corpus recounts past events and contains several expressions of attributed stance in the past tense (e.g. "as he *supposed*"; "he was not *sure* he had ever read it"). The remaining 57 passages all contain explicit marking of the stance of the speaker or writer, and most of them include stance expressions apart from the lexical items that were tagged.

10

5. Results and observations

I will now present and discuss some of the high-density passages that were retrieved from the corpora. Rather than presenting a comprehensive analysis, the discussion will draw attention to the different kinds of observations that can be made on the basis of such passages. For space reasons, I will only quote short extracts from the retrieved passages. Two examples of complete passages can be found in the appendix.

The first question concerns the contexts in which high-density passages can be observed. As mentioned above, passages with ten or more tagged markers were found in each of the four corpora. This shows that explicit and repeated expressions of stance are not restricted to a single context. Also within the corpora high-density passages can be found across different corpus sections. A look at the passages with the highest density of stance markers from each corpus can illustrate this point. From the CED 12 passages with 12 or more tagged markers were retrieved. Two of the passages are overlapping. The 11 non-overlapping passages come from three of the five sections of the corpus, namely Drama Comedy, Trial Proceedings, and Didactic Works. For the EMEMT, 11 non-overlapping passages with 11 or more tagged markers were retrieved from five of the seven sections, namely Appendix, Surgical and Anatomical Treatises, General Treatises and Textbooks, Specialised Treatises, and the Philosophical Transactions. From the Lampeter, 10 non-overlapping passages with 9 or more tagged markers were retrieved from all sections except Politics (Economics, Law, Science, Religion, and Miscellaneous). And the 10 non-overlapping passages with 12 or more tagged markers that were retrieved from the PCEEC come from 8 different letter collections and 4 periods and cover various types of relationships between author and recipient of the letter. Thus, it appears that there are only few restrictions concerning the text domain and text type in which very explicit and multiple marking of stance can be found.

Further observations result from the analysis of the stance expressions in sixty non-overlapping high-density passages (fifteen from each corpus). For instance, some of the passages contain stance markers that have not been included in previous studies. In 7, *collect that* has a meaning very similar to *conclude that*, a meaning which is also attested in the *Oxford English Dictionary* (collect *v.* 5.). In contrast to *collect*, *conclude* has been included in previous studies of stance marking as a factive verb expressing certainty (Biber 2004, 134; Gray, Biber & Hiltunen 2011, 230). In 8, *I credit* is used with the meaning of *I believe*, another common stance marker. Again, this meaning of *credit* is attested in the *Oxford English Dictionary* (credit *v.* II. 4. a.).

7) And this I *Collect* from these Considerations. 1. That there are innumerable Animalcula discovered in Semine Masculo omnium Animalium. (EMEMT, Garden, George, Philosophical transactions, volume 16, number 192, 1688)

8) And why should I *credite* more these, then others, who assygned other prynciples? And howe will you knowe how disguestion is made in the stomack. Truly, I for my part, dare no more *credite* the one, then the other. (EMEMT, Mexía, Pedro: Pleasaunt dialogue, 1580)

Both *credit* and *collect* are not very frequent, but they are used repeatedly with similar stance meanings in all four corpora. This suggests that analysing passages with a high density of stance markers can help identify previously unstudied stance markers.

Another observation that can be made concerns the integration of the tagged lexical items into larger stance expressions. Many of the markers do not express stance individually, but in combination with other markers. In order to understand the stance that is expressed, the markers need to be interpreted in the context of the sentence or passage in which they occur. For example, Extract 9 below includes two instances of *seem*, and one each of *truly, think* and *impossible*. In addition, there are a number of lexical items that were not included in the tagging, but which are also frequently used as stance markers, e.g. *may*, *consider*, *true* and *absurd.* If the passage is read in its context, it becomes clear that the stance the author expresses here could be paraphrased as *it is possible* or *it is probable*, but this meaning cannot be derived directly from any of the lexical stance markers.

9) And he who considers the nature of Vision, that it does not give us the true magnitude, but the proportion of Things, and that what *seems* to our naked Eye but a Point, may *truly* be made up of as many Parts as *seem* to us to be in the whole visible World will not *think* this an absurd or *impossible* thing. (EMEMT, Garden, George, Philosophical transactions, volume 16, number 192, 1688)

This extract illustrates several problems for studies that approach stance from a mainly quantitative perspective. First many of the lexical markers do not occur in fixed lexico-grammatical patterns that can easily be retrieved. For example, the first instance of *seems* is not followed by a complement clause with overt complementiser, and the same is true for *think*. In addition, *impossible* is used as an attribute to *thing*, rather than in the construction *it is impossible that*. This means that all these

instances cannot be retrieved automatically by approaches that rely on the overt marking of the complementiser.

Another issue that can be illustrated with Extract 9 concerns the interpretation of frequency information. Various problems of quantifying stance are known from research on present-day English, especially from research focusing on evaluative stance (for an overview, see Hunston 2011, Chapter 4). For instance, it has been shown that evaluative terms can have positive, negative, or neutral meaning depending on context, and that evaluative meaning can be difficult to pin down to specific lexical items (e.g. Hunston 2007). Similar points can be observed here. If each lexical item is counted independently, this short extract would add five to ten hits (depending on which items are included) to the overall count. However, one could argue that stance is only expressed once, albeit in a very elaborate way. The problem is further complicated by the fact that the sentence overall introduces a stance that applies to a proposition that is expressed outside of the sentence. The reference of *this* can be traced to the end of the previous sentence, given in 10.

10) [...] it seems most probable that the Stamina of all the Plants and Animals that have been,
    or ever shall be in the World, have been formed ab Origine Mundi by the Almighty
    Creator within the first of each respective kind. (EMEMT, Garden, George,
    Philosophical transactions, volume 16, number 192, 1688)

That stance markers refer to propositions outside of the sentence in which they occur is a characteristic that can often be observed in my data. Sometimes this is made very explicit, as in the use of *collect* in Extract 7 above, which introduces a list of six conclusions which spans 383 words. Therefore, without taking into account the scope of stance markers, frequency information alone cannot adequately represent to what degree stance is expressed in a corpus. The interpretation of frequency information can also be misleading in another way. As I argued above, the stance meaning that is expressed in Extract 9 cannot be calculated directly from the meaning of the individual stance markers. While the lexical meaning of most items expresses either weak support (*may, consider, seem, think*) or disagreement (*impossible, absurd*), the stance that is expressed overall is one of support. This means that frequency information about the occurrence of lexical stance markers from different semantic groups does not readily reflect how frequently the corresponding stance meanings are expressed in a corpus.

The analysis of the passages also leads to a number of observations that could be used as a starting point for further investigations. For instance, rhetorical questions are used in the 15 highest-density

passages from three of the four corpora, namely in the EMEMT (see Extract 8 above), in the scientific section of the Lampeter corpus (Extract 11) and in trial proceedings from the CED (Extract 12).

11) In Moral Philosophy (which is not established on demonstrations) and Physique I am sure it will put us all to confusion; for who knowes the efficacy of things Naturall? how Purges and Vomits &c. do worke? Nay, who knowes whether second causes have any effect, but that God doth operate "ad praesentiam causarum secundarum"? as the admirable Durandus held. And may not the Divel do so? How shall this be reduced to Faith? I am sure, and have seen things ascribed to the Divel, that it is now confessed have another Originall. (Lampeter, Miscellaneous, "The Miraculous Conformist", 1666)

12) M. B.: You doe not thinke then that the deuill doeth teach her?

The good wife R.: How should I thinke that the Deuill doeth teach her? Did you euer heare that the deuill did teach any good thing?

M. B.: Doe you know that was a good thing?

The good wife R.: Was it not a good thing to driue the euill spirit out of my creame?

(CED, Handbook, "Concerning Witches", 1593)

The rhetorical questions in these passages can be interpreted as an additional way of expressing or boosting stance. Quirk et al. (1985, 825–826) describe rhetorical questions as having the "force of [...] strong assertion[s]" and they use the stance adverbial *surely* to paraphrase their meaning. This means that instead of expressing commitment to a proposition through constructions like *surely, I believe* or *I am sure*, language users can present the proposition in the form of a rhetorical question and thereby express their commitment to its truthfulness. Hiltunen (1996, 28–29) describes the use of this device by defendants in the Salem witchcraft trials. He finds that the use of rhetorical questions is restricted to defendants who plead not guilty, and it is particularly common in one of the longest and most powerful defences in his data (1996, 28–29). Further studies would be needed to see how wide-spread this use of rhetorical questions is in Early Modern English texts and how rhetorical questions combine with other forms of stance marking.

Another related observation is that dialogue and direct address can be found in passages from all four corpora. This is not surprising for the CED and the PCEEC, since they consist of dialogues and letters and direct address can thus be expected to be the norm. However, dialogues were also found in several of the passages from the EMEMT and the Lampeter corpus. In medical writing, dialogues are often used for didactic purposes (Taavitsainen 2009, 2010, 54), and similar uses can be found in

14

pamphlets (Claridge 2001, 57). In addition, when pamphlets are published in response to previous pamphlets, the text as a whole becomes part of a dialogue between the authors (Claridge 2005). An example of a dialogic text from the EMEMT is given in Extract 8 above, which comes from a text that is structured as a dialogue between four interlocutors. Three out of the fifteen high-density passages from the EMEMT come from three different texts that have a dialogic structure, and one additional passage contains an imagined dialogue between the author and the reader. Seven out of the fifteen passages from the Lampeter corpus contain direct address forms. One of the passages comes from a text that is structured as a dialogue, similar to the examples from the EMEMT. The other passages come from pamphlets that are published in the form of letters (two), as replies to authors of previously published pamphlets (two), or as a reply to political representatives (one). In one case, the addressee is only vaguely defined as "all Persons that have any thing to do with ELECTIONS" (Lampeter, Law, "An argument of a learned judge", 1704). Gray, Biber and Hiltunen (2011, 244) mention a possible correspondence between dialogic texts and stance marking in their study of Early Modern English medical texts, but further research is needed to find out how systematic this relation is.

6. Conclusion

In this paper I presented a method of using corpora for the detailed qualitative analysis of epistemic and evidential stance. The method retrieves passages that contain a large number of lexical stance markers in close vicinity. These passages tend to contain explicit and elaborate stance expressions that can lead to new insights into how stance is marked in Early Modern English. They can help identify previously unstudied stance markers, they can point to problems for the interpretation of quantitative findings and they can provide indications of (con)textual characteristics that deserve further attention, such as rhetorical questions and direct address. Perhaps the most important finding is the fact that passages with a high-density of stance markers and very explicit stance expressions can be found across various contexts in Early Modern English. This is particularly noteworthy in light of previous findings that indicate that stance expressions occur with lower frequency in Early Modern English than in Present-day English, which has been interpreted as "a general shift in cultural norms: speakers and writers are simply more willing to express stance in recent periods than in earlier historical periods" (Biber 2004, 129-130). My findings show that there are situations in which speakers and writers of Early Modern English decided to mark their stance in very explicit ways and the findings also indicate that their use of stance expressions is at times difficult to capture in numbers,

for instance due to the need for contextual interpretation and the potentially large scope of stance markers. Thus, the method I presented in this paper can be a valuable complementation to existing methods that focus on individual forms, on quantitative evaluations of fixed lexico-grammatical patterns, and on the comprehensive study of smaller text samples from well-defined domains.

I used this method to study stance expressions, but similar approaches could be used to study other pragmatic phenomena, as long as it is possible to tag formal characteristics (e.g. certain lexical items) that tend to co-occur frequently with specific pragmatic functions. As far as stance is concerned, it would be interesting to use different sets of lexical items as a starting point for identifying high-density passages. In this study, I used a set of 20 lexical items that frequently express epistemic and evidential stance, but similar studies could be carried out based on markers of evaluative stance. The identification of clusters of lexical items might not work equally well for the study of other pragmatic functions. However, most pragmatic phenomena are not evenly distributed across corpora. Instead, they tend to be restricted to some text passages, and they often occur in repeatedly in close vicinity. Therefore, the core idea of the method – the retrieval of relevant passages for qualitative analysis – may hold potential for the study of pragmatic features more generally.

An important characteristic of the method I presented is that it is scalable; as long as corpora are available as text files that can be saved locally on a computer, the Python scripts can be used to retrieve passages with a large number of stance markers. Some adjustments are needed to address differences in corpus design and mark-up, but these are not very time-consuming and they are not dependent on corpus size. This means that new and larger corpora can be used to retrieve passages that are suitable for qualitative analysis. A larger amount of corpus data is likely to yield more passages with a higher density of stance markers and thus may lead to even richer insights without requiring an increase in the amount of work that is proportional to corpus size. In this way, new and larger corpora can be fruitfully used not only for quantitative but also for qualitative analysis.

Acknowledgements

Appendix: Examples of retrieved passages

Additional text at the beginning and end of the retrieved 300 word passages is added in square brackets.

Passage 1

11 hits in 300 words

EMEMT, Section 6, Garden, George, Philosophical transactions, volume 16, number 192, 1688

[and from hence we may probably conjecture that so curiously an organized Creature as an Animal, is not the sudden product of a Fluid or Colliquamentum, but does much rather proceed from an Animalcle of the same kind, and has all its] little Members folded up according to their several Joynts and Plicatures, which are afterwards enlarged and distended, as we see in Plants. Now though this Consideration alone may *seem* not to bear much weight, yet being joyn'd to the two former they do mutually strengthen each other. And indeed all the Laws of Motion which are as yet discovered, can give but a very lame account of the forming of a Plant or Animal. We see how wretchedly Des Cartes came off when he began to apply them to this Subject; they are form'd by Laws yet *unknown* to Mankind, and it *seems* most *probable* that the Stamina of all the Plants and Animals that have been, or ever shall be in the World, have been formed ab Origine Mundi by the Almighty Creator within the first of each respective kind. And he who considers the nature of Vision, that it does not give us the true magnitude, but the proportion of Things, and that what *seems* to our naked Eye but a Point, may *truly* be made up of as many Parts as *seem* to us to be in the whole visible World will not *think* this an absurd or *impossible* thing.

But the second thing which later discoveries have made *probable* is, that these Animalcles are originally in Semine Marium & non in fúminis. And this I Collect from these Considerations. 1. That there are innumerable Animalcula discovered in Semine Masculo omnium Animalium. Mr. Leewenhoeck has made this so *evident* by so many Observations, that I do not in the least question the truth of the thing. The reason of their multitude, and some of the difficulties which arise thereupon, he has cleared to very good purpose, so that I shall not repeat them. 2. The observing the [Rudiments of the Fœtus in Eggs, which have been fecundated by the Male, and the seeing no such thing in those which are not fecundated, as appears from Malpighius his Observations, make it very probable that these Rudiments proceed originally from the Male, and not from the Female.]

Passage 2

13 hits in 300 words

PCEEC, Arundel_010, Anne Howard (née Dacre) to Thomas Howard III, 1609

My good Sonne

(...) Your opinion, sweete harte, for sending to my Lady Lumley, I will follow by writing very shortly.

But I besech y[u] consider well consarning my entry of Hayling,

for I *doute* the deferring of it may rather bring harme than otherwise.

This is my reason.

My Lady Lumley *knoweth* very well that it was both *thought* by us, as hath been said to hirself, that my L. hir late husband did great wronge in keeping of itt, and therefor that I *think* it, my right cannot *seme* strang to hir; besids she sayed, as I *think* y[u] told me, that hir Lord had made some formar estate befor he made it to y[or] Lo.; which to avoide I *think* it fittest to holde my owne and y[or] true right on foote, rather than by forbering to geve advantage, if in the Lady Lumley's life, or after, that gifte shall appere.

For being beholdinge to my L. Lumley so much as I was in time of my hevy fortune, I had cause to forbeare him; yett y[u] *know*, swett harte, he payed me still the olde rentt,

and my *doute* is that if y[u] speake to Sir Thomas Savage of it, or if it be done, happely she will desire to forbear it, which will not be good, I *think*, for y[u] nor me.

But, if it were done, y[u] might say to him that y[u] see y[u] must be beholding to hir and me bothe for Haling; and then y[u] may forbeare any question aboute the farme till y[u] see time.

I writt what I *think*; but, good Sonne, if my brother and y[u] *think* it better to be deferred, use y[or] discretion.

But I *think* you will happely, if y[u] talke wth Sr Thomas [Savage and in general desir his furtherance for what is fitt for y[u] in Sussex, to have it before an other ...]

Data

CED. A Corpus of English Dialogues 1560-1760. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).

EMEMT. Early Modern English Medical Texts. 2010. Compiled by Irma Taavitsainen (University of Helsinki), Päivi Pahta (University of Tampere), Martti Mäkinen (Svenska handelshögskolan), Turo Hiltunen, Ville Marttila, Maura Ratia, Carla Suhr, Jukka Tyrkkö (University of Helsinki).

PCEEC. Parsed Corpus of Early English Correspondence, parsed version. 2006. Annotated by Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

PPCEME. Penn-Helsinki Parsed Corpus of Early Modern English. 2004. Compiled by Anthony Kroch, Beatrice Santorini, and Ariel Diertani.

The Lampeter Corpus of Early Modern English Tracts. 1999. Compiled by Josef Schmied, Claudia Claridge, and Rainer Siemund. (In: ICAME Collection of English Language Corpora (CD-ROM),

Second Edition, eds. Knut Hofland, Anne Lindebjerg, Jørn Thunestvedt, The HIT Centre, University of Bergen, Norway.)

References

Aijmer, Karin. 1997. *I think*: An English modal particle. In Toril Swan & Olaf Jansen Westvik (eds.), *Modality in Germanic languages: Historical and comparative perspectives*, 1–47. Berlin: Mouton de Gruyter.

Aijmer, Karin. 2014. Pragmatic markers. In Karin Aijmer & Christoph Rühlemann (eds), *Corpus pragmatics: A handbook*, 195–218. Cambridge: Cambridge University Press.

Aijmer, Karin & Christoph Rühlemann (eds.). 2014. *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.

Alonso-Almeida, Francisco & Margarita Mele-Marrero. 2014. Stancetaking in seventeenth-century prefaces on obstetrics. *Journal of Historical Pragmatics* 15(1). 1–35.

Biber, Douglas. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics* 5(1). 107–36.

Brinton, Laurel. 2008. *The comment clause in English*: *Syntactic origins and pragmatic development* (Studies in English Language). Cambridge: Cambridge University Press.

Bromhead, Helen. 2009. *The reign of truth and faith: Epistemic expressions in 16th and 17th century English* (Topics in English Linguistics 62). Berlin: Mouton de Gruyter.

Claridge, Claudia. 2001. Structuring text: Discourse deixis in Early Modern English. *Journal of English Linguistics* 29(1). 55–71.

Claridge, Claudia. 2003 [1999]. *'Life is ruled and governed by opinion': The Lampeter Corpus of Early Modern English Tracts. Manual of information*. 2nd edn.

Claridge, Claudia. 2005. Questions in Early Modern English pamphlets. *Journal of Historical Pragmatics* 6(1). 133–68.

Fitzmaurice, Susan. 2003. The grammar of stance in early eighteenth-century English epistolary language. In Pepi Leistyna & Charles F. Meyer (eds.), *Corpus analysis: Language structure and language use*, 107–31. Amsterdam & New York: Rodopi.

Gisborne, Nikolas & Jasper Holmes. 2007. A history of English evidential verbs of appearance. *English Language and Linguistics* 1(1). 1–29.

Gray, Bethany, Douglas Biber & Turo Hiltunen. 2011. The expression of stance in early (1665-1712) publications of the philosophical transactions and other contemporary medical prose: Innovations

in a pioneering discourse. In Irma Taavitsainen & Päivi Pahta (eds.), *Medical writing in Early Modern English* (Studies in English Language), 221–57. Cambridge: Cambridge University Press.

Grund, Peter J. 2012. The nature of knowledge: Evidence and evidentiality in the witness depositions from the Salem Witch Trials. *American Speech* 87(1). 7–38.

Grund, Peter J. 2013. *I saw ye child burning in ye fire*: Evidentiality in Early Modern English witness depositions. In Andreas H. Jucker, Daniela Landert, Annina Seiler & Nicole Studer-Joho (eds.), *Meaning in the history of English: Words and texts in context* (Studies in Language Companion Series 148), 319–41. Amsterdam & Philadelphia: John Benjamins.

Hiltunen, Risto. 1996. *Tell me, be you a witch?*: Questions in the Salem Witchcraft Trials of 1692. *International Journal for the Semiotics of Law* 9(1). 17–37.

Hiltunen, Turo & Jukka Tyrkkö. 2011. Verbs of knowing: Discursive practices in Early Modern vernacular medicine. In Irma Taavitsainen & Päivi Pahta (eds.), *Medical writing in Early Modern English* (Studies in English Language), 44–73. Cambridge: Cambridge University Press.

Hunston, Susan. 2007. Using a corpus to investigate stance quantitatively and qualitatively. In Robert Englebretson (ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (Pragmatics & Beyond New Series 164), 27–48. Amsterdam & Philadelphia: John Benjamins.

Hunston, Susan. 2011. *Corpus approaches to evaluation: Phraseology and evaluative language*. New York: Routledge.

Jucker, Andreas H. 2013. Corpus pragmatics. In Jan-Ola Östman & Jef Verschueren (eds.), *Handbook of pragmatics*, 1–17. Amsterdam & Philadelphia: John Benjamins.

Jucker, Andreas H., Gerold Schneider, Irma Taavitsainen & Barb Breustedt. 2008. Fishing for compliments: Precision and recall in corpus-linguistic compliment research. In Andreas H. Jucker & Irma Taavitsainen (eds.), *Speech acts in the history of English* (Pragmatics & Beyond New Series 176), 273–94. Amsterdam & Philadelphia: John Benjamins.

Jucker, Andreas H. & Irma Taavitsainen. 2014. Complimenting in the history of American English: A metacommunicative expression analysis. In Irma Taavitsainen, Andreas H. Jucker & Jukka Tuominen (eds.), *Diachronic corpus pragmatics* (Pragmatics & Beyond New Series 243), 257–76. Amsterdam & Philadelphia: John Benjamins.

Kohnen, Thomas. 2007. Text types and the methodology of diachronic speech act analysis. In Susan M. Fitzmaurice & Irma Taavitsainen (eds.) *Methods in historical pragmatics*, 139–66. Berlin: Mouton de Gruyter.

Kohnen, Thomas. 2008. Tracing directives through text and time: Towards a methodology of a corpus-based diachronic speech-act analysis. In Andreas H. Jucker & Irma Taavitsainen (eds.), *Speech acts in the history of English* (Pragmatics & Beyond New Series 176)*, 295–310. Amsterdam & Philadelphia: John Benjamins.

Kytö, Merja & Terry Walker. 2006. *Guide to a corpus of English dialogues 1560-1760* (Studia Anglistica Upsaliensia 130). Uppsala: Acta Universitatis Upsaliensis.

López Couso, María José & Belén Méndez-Naya. 2014. From clause to pragmatics marker: A study of the development of "like"-parentheticals in American English. *Journal of Historical Pragmatics* 15(1). 36–61.

López Couso, María José & Belén Méndez-Naya. 2015. Evidential/epistemic markers of the type verb + complementizer: Some parallels from English and Romance. In Andrew D. M. Smith, Graeme Trousdale & Richard Waltereit (eds.), *New directions in grammaticalization research* (Studies in Language Companion Series 166), 93–120. Amsterdam & Philadelphia: John Benjamins.

Manes, Joan & Nessa Wolfson. 1981. The compliment formula. In Florian Coulmas (ed.), *Conversational routine: Explorations in standardized communication situations and prepatterned speech*, 115–32. The Hague: Mouton.

Palander-Collin, Minna. 1999. *Grammaticalization and social embedding:* I think *and* methinks *in Middle and Early Modern English*. Helsinki: Société Néeophilologique.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Romero-Trillo, Jesús, (ed.). 2008. *Pragmatics and corpus linguistics: A mutualistic entente*. Berlin & New York: Mouton de Gruyter.

Rühlemann, Christoph & Karin Aijmer. 2014. Corpus pragmatics: Laying the foundations. In Karin Aijmer & Christoph Rühlemann (eds.), *Corpus pragmatics: A handbook,* 1–26. Cambridge: Cambridge University Press.

Suhr, Carla & Irma Taavitsainen (eds.). 2012. *Developing corpus methodology for historical pragmatics*. (Studies in Variation, Contacts and Change in English 11). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/11/

Taavitsainen, Irma. 2000. Metadiscursive practices and the evolution of Early English medical writing 1375-1550. In John M. Kirk (ed.), *Corpora galore: Analyses and techniques in describing English,* 191–207. Amsterdam: Rodopi.

Taavitsainen, Irma. 2001. Evidentiality and scientific thought-styles: English medical writing in Late Middle English and Early Modern English. In Maurizio Gotti & Marina Dossena (eds.), *Modality in specialized texts*, 21–52. Bern: Peter Lang.

Taavitsainen, Irma. 2009. Authority and instruction in two sixteenth-century medical dialogues. In Matti Peikola, Janne Skaffari & Sanna-Kaisa Tanskanen (eds.), *Instructional writing in English: Studies in honour of Risto Hiltunen* (Pragmatics & Beyond New Series 189), 105–24. Amsterdam & Philadelphia: John Benjamins.

Taavitsainen, Irma. 2010. Discourse and genre dynamics in Early Modern English medical writing. In Irma Taavitsainen & Päivi Pahta (eds.), *Early Modern English medical texts: Corpus description and studies*, 29–53. Amsterdam & Philadelphia: John Benjamins.

Taavitsainen, Irma & Päivi Pahta (eds.). 2010. *Early Modern English medical texts: Corpus description and studies*. Amsterdam & Philadelphia: John Benjamins.

Taavitsainen, Irma, Andreas H. Jucker & Jukka Tuominen (eds.). 2014. *Diachronic corpus pragmatics* (Pragmatics & Beyond New Series 243). Amsterdam & Philadelphia: John Benjamins.

Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk & Terttu Nevalainen. 2006. *The Parsed Corpus of Early English Correspondence*. http://www-users.york.ac.uk/~lang22/PCEEC-manual/index.htm. (last accessed 28 September 2016).

Whitt, Richard J. 2010. *Evidentiality and perception verbs in English and German*. Bern: Peter Lang.

Whitt, Richard J. 2011. (Inter)subjectivity and evidential perception verbs in English and German. *Journal of Pragmatics* 43(1). 347–60.