Grammatical stance marking across registers

Revisiting the formal-informal dichotomy

Tove Larsson
Uppsala University & Université catholique de Louvain

There are sets of grammatical stance markers that are morphologically and semantically related, but that differ with regard to their syntactic realization (e.g., *importantly*, *it is important that* and *the importance of*). Little attention has, however, been paid to how these pattern across registers. This study examines eleven such sets across five registers in apprentice and expert production to investigate which register(s) the apprentice writers' use is closest to and what that can tell us about their adherence to academic norms. The results show that there is a cline from the *a priori* more formal registers to the less formal registers for the stance markers investigated. When the apprentice writers' usage was mapped onto this cline, it became clear that their usage diverged slightly from that of the academic experts, thus indicating a lack of register awareness. Yet, very little evidence was found to support previous claims of the 'spoken-like' nature of learner writing.

Keywords: grammatical stance marking, registers, learner writing, expert production, level of formality

1. Introduction

Stance-marking devices help us position ourselves in relation to our claims, and as such, are of central importance in both speech and writing (e.g., Biber 2006b; Biber, Johansson, Leech, Conrad, & Finegan 1999). Following Gray and Biber (2012: 15), the term 'stance' (as realized through 'stance markers') is here defined as the "the linguistic mechanisms that convey a speaker or writer's personal attitudes and assessments". Because of the frequency and importance of stance marking in academic production (e.g., Biber 2006b: 87), it is important for apprentice student writers, regardless of native-speaker status (henceforth 'apprentice writ-

https://doi.org/10.1075/rs.18009.lar
Register Studies 1:2 (2019), pp. 243–268. issn 2542-9477 | e-issn 2542-9485

ers'), to attain a good command of these devices. However, many previous studies report that appropriate stance marking is something that many apprentice writers, in particular non-native writers, struggle with (e.g., Hasselgård 2015; Petch-Tyson 1998). Functional categories such as hedges (e.g., *seems*, *it is possible that*), boosters (e.g., *essential*, *importantly*) and attitude markers (e.g., *interestingly*, *noteworthy*) seem to be especially error-prone (see, e.g., Hinkel 2005; Larsson 2017a, 2017b).

To complicate matters further, there are sets of stance markers that are semantically and morphologically related, but that differ with regard to their syntactic realization, such as *possibly*, *it is possible that* and *the possibility of*. This group of stance markers encompasses members that have received ample individual attention, such as stance adverbs (e.g., *possibly*), stance complement clause constructions (e.g., *it is possible that*), and a stance noun followed by a prepositional phrase (e.g., *the possibility of*); the distribution of these individual categories has also been found to vary across registers (Biber et al. 1999: 979). However, little is known about the interrelation of the members of such morphologically and semantically related lexical sets and to what extent these vary across registers. Furthermore, the pragmatic and syntagmatic relations that each of the members of such sets participate in might be unknown to apprentice writers who may view these as fully synonymous variants from which to choose freely. In fact, previous research on sets such as IMPORT* (*importantly*, *important* and *importance*) has shown that learners tend not to be able to use such markers in a target-like manner compared to expert academic writers (Larsson 2017a).

Related to such difficulties is the question of appropriate level of formality. Many apprentice writers (particularly learners) are often reported to be overly informal in their writing (e.g., Altenberg & Tapper 1998). As some of the members of these sets of stance markers can be viewed as being more informal than others, in the sense that they are more strongly associated with speech than writing (see, e.g., Larsson 2017a for a discussion of *probably*), such sets present a good opportunity for further investigation of claims of informality. However, formality is still often (perhaps in particular in teaching contexts) perceived as being a binary concept.

In an attempt to provide a more nuanced and, at the same time, a more detailed picture of (in)formality in apprentice writing, the present study uses a method developed in Larsson and Kaatari (2019) where (in)formality is viewed as a continuum rather than a dichotomy (cf., e.g., Ädel 2008; Smith 1986) and where registers are placed along this continuum, from more formal to less formal, based on their respective situational characteristics. The distribution of any given linguistic features can then be investigated in registers in order to associate these features as more informal or more formal. Based on this, these features are sub-

sequently used to characterize the level of formality of non-native-speaker (NNS) and native-speaker (NS) apprentice writers. The different registers thus act as points of reference on the informal-formal cline, enabling a more fine-grained discussion of (in)formality. This approach also has the added advantage that we can study the extent to which NNS and NS apprentice writers adhere to register-specific differences.

Based on Biber et al. (1999:16) and Kaatari (2017:43), an *a priori* ordering of the registers used in the present study in descending order of formality based on their different situational characteristics looks as follows: academic prose, popular science, news, fiction and conversation. The proposed mapping is shown in Figure 1 (adapted from Biber et al. 1999:16; Kaatari 2017:43; cf. also Larsson & Kaatari 2019).

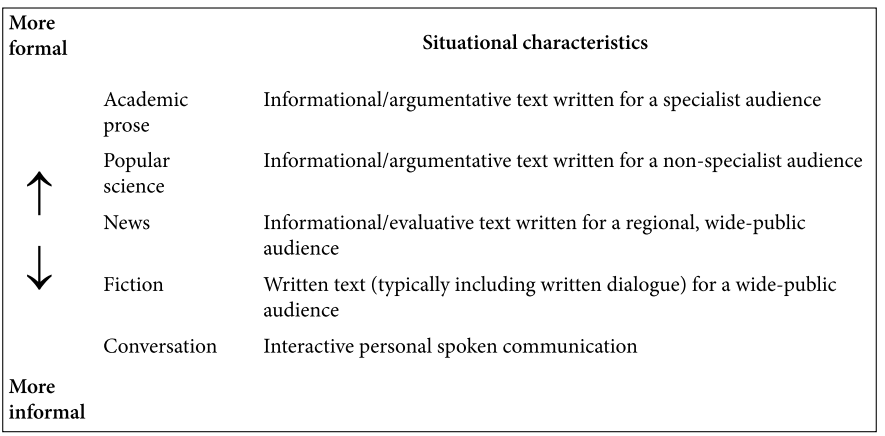


Figure 1. The registers mapped onto the informal-formal cline

Against this background, the present study aims to investigate the distribution of sets of morphologically related stance markers across registers to further explore previous claims of informality, focusing on the interplay between lexis and grammar. Specifically, using Biber et al.'s (1999:969–970) framework of grammatical stance marking (see Section 3.2), the study focuses on four sets of near-synonyms that are used for hedging claims – POSSIB*, PROBAB*, CONCEIVAB* and LIKEL* – as well as four sets that are used as boosters – IMPORTAN*, CRUCIAL*, ESSENTIAL* and IMPERATIVE* – and three sets that are used as attitude markers – INTEREST*, NOTEWORTH* and CURIOUS* (see Section 3.2 for an overview and a description of the selection process). The term 'hedges' is used here to denote constructions that express "possibility rather than certainty" and indicate "a lack of complete commitment to the truth of a proposition or […] a desire not to express that commitment categorically" (Hyland 1996:251). The term 'boosters' is

used here to denote linguistic features that strengthen the force of an utterance (see, e.g., Hewings & Hewings 2002: 373), whereas 'attitude markers' express "the writer's affective attitude towards what is stated in the clausal subject" (Larsson 2017b: 61). Considering that function has been shown to be an important factor to explain the distribution for some members of these sets (Larsson 2017b), controlling for function (hedges vs. boosters vs. attitude markers) enables more detailed analyses of the lexico-grammatical distribution of these sets than would otherwise be possible.

The study uses subsets from two learner corpora, two NS student corpora and one reference corpus (see Section 3.1). The following research questions are investigated:

- What differences and similarities in the distribution of morphologically related stance markers can be noted across registers?
- Which of the registers is the apprentice writers' academic prose closest to linguistically, and what can this tell us about (in)formal uses of such stance markers in apprentice writing?

2. Background

Stance is a widely studied concept, with work by Biber and colleagues (e.g., Biber 1995, 2006a, 2006b; Biber et al. 1999; Biber & Zhang 2018; Gray & Biber 2012) and Hyland (e.g., Hyland 1996, 2005) arguably being the most pivotal in the field. Stance covers the study of devices that express epistemic or attitudinal meaning (Biber et al. 2018: 198). Traditionally, stance has been investigated using quantitative corpus-based methods looking at lexico-grammatical features, whereas related theoretical constructs such as evaluation (e.g., Hunston & Thompson 2000), appraisal (e.g., Martin & White 2005) and attitude (e.g., Halliday 1994) have been studied more qualitatively by primarily looking at individual words and expressions in context (cf. Biber & Zhang 2018). While some recent studies have sought to bridge this paradigm gap (e.g., Biber & Zhang 2018; Biber et al. 2018), showing for example that studies of stance and evaluation can offer different and complementary views on evaluative language, the present study is situated within the framework of stance, as described by Biber and colleagues.

Stance has been investigated using both specific and inclusive study designs. For example, whereas some studies focus on specific subtypes of stance marking, such as extraposition (e.g., *it is interesting to note*; Larsson & Kaatari 2019) and reporting clauses followed by a *that*-clause (e.g., *Jones argues that*; Charles 2006), other studies take a more inclusive approach. An example of the latter is reported

in Biber et al. (1999), where grammatical stance marking was investigated. This type denotes "a stance relative to some other proposition", as in *I just **hope** that she is here now* (Biber et al. 1999: 969), and thus differs from affective words (e.g., *wonderful*, *happy*) (Biber et al. 1999: 968). The framework includes categories such as stance adverbials (e.g., *unfortunately*), stance noun + prepositional phrase (e.g., ***the possibility** of disagreement is higher now*) and stance complement clauses (e.g., *he is **happy** that she finally arrived*) (Biber et al. 1999: 969–970). Grammatical stance marking, which is the framework used in the present study (see Section 3.2), was also used in Larsson (2017a) to investigate sets of stance markers, similar to those included in the present study, such as IMPORTAN* (e.g., *the **importance** of*, ***importantly*** and *it is **important** to*) in published research articles and learner data. The study examined to what extent factors such as level of expertise in academic writing, first-language (L1) transfer and lexis can be seen to influence the distribution of these realizations. The results showed that all three of these factors had an effect. For example, there was clear inter-lexical variability between the base forms (e.g., IMPORTAN*, as above); moreover, tahe learners tended to struggle with individual stance markers, such as ***interestingly*** and *it is **interesting** to*.

A factor that was not investigated in relation to stance in Larsson (2017a), but that has received considerable attention elsewhere in the literature is 'register'. Following Biber et al. (1999: 15), register distinctions are here defined "in non-linguistic terms, with respect to situational characteristics such as mode, interactiveness, domain, communicative purpose, and topic" (see Lee 2001 on the terms 'register', 'genre' and 'text type' and Biber & Conrad 2009: 2 for a discussion of register in relation to 'genre' and 'style'). Biber et al. (1999: 979ff), who looked at grammatical stance marking in four registers – conversation, fiction, news and academic prose – noted certain differences across the registers. For example, with regard to individual grammatical categories, adverbial stance markers (with 'single adverbs' being the most common subcategory) were most frequent in conversation, although, as was noted, this category was also "relatively common" in academic prose (Biber et al. 1999: 979); these findings were echoed by Biber (2006b: 103), where stance adverbs were reported to be "generally much more common in the spoken registers than in the written registers". Register has also been discussed in relation to level of formality (e.g., Biber 1995), where written data have been described as being more likely to exhibit features typical of formal language than spoken data. Other studies have looked at specific features often referred to as 'informal', such as contracted forms (e.g., *it's*, *what's*) (Olohan 2003) and omission of the complementizer *that* (e.g., Kaatari 2017) and noted that these are common in less formal registers such as fiction and speech.

Furthermore, learners have sometimes been reported to be more informal in their writing than NS students, in that the former group tends to use features that

are typical of spoken language in their writing in English. For example, Petch-Tyson (1998:116) found that the learner groups investigated (L1 French, Swedish, Finnish and Dutch) exhibited a higher degree of interpersonal involvement than the NS students, in the sense that almost all features of reader/writer visibility, including stance marking, were overused by the learners. Similar observations in L1 Swedish data led Herriman and Boström Aronsson (2009:118) to conclude that L1 Swedish learners' argumentative writing "is similar to NS' spoken language". However, in Larsson and Kaatari's (2019) study looking at extraposition, it was noted that some of the preferences described in the literature could be largely attributed to text type, meaning, in this context, that the (L1 Swedish) learners' academic prose was found to be more similar to the expert writers' academic prose than to any other register, whereas their argumentative writing was found to differ considerably from the experts' academic prose. As text type (academic prose vs. argumentative texts) was found to be an important indicator that could have a negative impact on corpus comparability (see also Ädel 2006, 2008; Callies 2013), the present study has controlled for this factor.

Based on previous research, we thus know that the kinds of stance markers investigated are likely to exhibit register-specific distributional differences. However, little is known about how individual stance markers vary along lexical and syntactic dimensions and what this distribution looks like across registers. As the sets of stance markers studied here are morphologically related, they provide an opportunity to obtain a more complete picture of these dimensions and their interaction, along with information about how frequent each of the members of these sets are in relation to the other members. Further, since previous research has shown somewhat contradictory results with regard to whether learners are 'informal' (or 'spoken-like') in their academic writing, an investigation of these stance markers in relation to their register distribution will contribute to a more complete picture of (in)formality in apprentice writing than has been done in previous studies looking at only one construction (e.g., Larsson & Kaatari 2019).

3. Data and method

In this section, the material used is presented in Section 3.1, and the method used is described in Section 3.2.

3.1 The corpus data

The study uses subsets from two learner corpora, ALEC and VESPA, two NS student corpora, BAWE and MICUSP, and one reference corpus, BNC-15. ALEC (the

Advanced Learner English Corpus; Larsson 2014) is a 1.3-million-word corpus of NNS academic writing from Swedish university students of English linguistics and English literature. VESPA (the *Varieties of English for Specific Purposes Database*; Paquot, Hasselgård, & Oksefjell Ebeling 2013) is a multi-million-word corpus of academic writing in disciplines such as linguistics and business communication. In the present study, a subset from the L1 Swedish components of ALEC and VESPA (ALEC-SE and VESPA-SE) are included for investigation. This subset includes untimed linguistics and literature theses written by students who are in their third year of university studies on average.

The study also uses an NS-student corpus, composed of subsets from MICUSP (the *Michigan Corpus of Upper-level Student Papers*) and BAWE (the *British Academic Written English*). In order to ensure comparability to the greatest extent possible, these corpora were carefully sampled to be as similar as possible to the NNS corpus with regard to text type, year of study, discipline and the number of words that each student has contributed. Nonetheless, certain unavoidable differences remain between the NNS and the NS corpora, mainly pertaining to the length of the texts, with the NS texts being shorter than the NNS texts on average.

The final corpus to be included is BNC-15 (Kaatari 2017), which is a methodically sampled 3-million-word subset of the BNC (the *British National Corpus*; Burnard 2007). BNC 15 allows for register comparisons in comparable subsets. It includes five registers: academic prose, popular science, news, fiction and conversation. An overview of the subcorpora included for investigation can be found in Table 1.

Table 1. An overview of the subcorpora included

	Subcorpus	Word count	Number of texts	Mean text length	L1
Apprentice corpora	ALEC-SE	905,572	103	8,792	Swedish
	VESPA-SE	155,469	22	7,067	Swedish
	MICUSP	349,242	119	2,935	English
	BAWE	150,593	46	3,274	English
BNC-15	Academic Prose	600,117	60	10,002	English
	Popular Science	600,122	60	10,002	English
	News	600,326	60	10,005	English
	Fiction	600,334	60	10,006	English
	Conversation	600,049	60	10,001	English
Total		**4,561,824**	**590**		

Due to the nature of the corpora deemed most suitable for the focus of the study, it falls outside the scope of the study to investigate possible disciplinary differences. Nonetheless, an effort was made to increase the comparability of the corpora by using reference corpora and learner corpora that comprise data from a mix of disciplines, rather than from one single discipline.

3.2 Method

To select suitable sets of stance markers for the present study, a combination of a bottom-up and a top-down approach was applied. Furthermore, since the definition of the widely used terms 'stance' and 'stance marking' varies across studies, a decision was made to use Biber's (2006b: 92–93) extensive list of grammatical stance markers as a benchmark to make this study more easily replicable.

To enable careful comparisons of the distribution of the stance markers across the registers and corpora, it was considered important to look at sets that were as similar as possible with regard to their discourse function. The study therefore started out from three high-frequency adjectives that have been found to be particularly frequently used by L1 Swedish learners, namely *possible*, *important* and *interesting* (Larsson 2016; Larsson & Kaatari 2019). In addition to potentially being problematic for learners, these three adjectives and their morphologically related equivalents belong to three different functional categories (hedges, boosters and attitude markers, respectively), thereby making it possible for the study to control for function. A decision to focus on L1 Swedish learners was made, as it not only facilitates comparisons to the results of Larsson and Kaatari's (2019) study, but also enables investigation of texts written by users of English as a foreign language with advanced-level proficiency in English (cf. the *Common European Framework of References for Languages;* Council of Europe 2001). While a large-scale study of several different L1s would most certainly add to our knowledge of non-native uses of stance markers, this kind of investigation falls outside the scope of the present study and will thus have to be left for future studies.

The synonym function (based on *WordNet*) in the multi-million-word *Corpus of Contemporary American English* (COCA) (Davies 2008) was subsequently used to obtain a list of near-synonyms for these adjectives; this list was then cross-referred with Biber's (2006b) list of stance markers to identify those markers for which there was overlap between the lists. Four such adjectives were found that can be used for hedging (*possible*, *likely*, *probable* and *conceivable*), four adjectives that can function as boosters (*important*, *essential*, *crucial* and *imperative*) and three adjectives that can function as attitude markers (*interesting*, *noteworthy* and *curious*). The corresponding nominal and adverbial forms were subsequently

added, to enable investigation of the full set of syntactic variants of interest to this study, as described below.

The study considers three of Biber et al.'s (1999: 969–970) categories of grammatical stance markers: (i) stance adverbials (e.g., *conceivably*), (ii) stance noun + prepositional phrase (*NP+PP*; e.g., ***the possibility** of agreement is denied*) and (iii) stance complement clauses (*comp*; e.g., *I'm **happy** that…*); the third category includes extraposed structures (e.g., *it is **noteworthy** that…*). As not all corpora used are part-of-speech tagged, lexical searches were carried out to extract the tokens. One advantage of using lexical searches (rather than search strings) is that they improve the recall (i.e., how many relevant/valid tokens are identified).

It was, nonetheless, necessary to take certain measures in order to increase the precision (and thereby, the manageability) of the searches and results. Tokens belonging to the stance complement clause construction were identified through a search for the corresponding adjective followed by *that* (for the hedges) and by *that* or *to* (for the boosters and attitude markers). The reason for this discrepancy between the functional categories is the polysemous character of some of the adjectives in the hedges category; for example, whereas *it is possible that* serves a hedging function, *it is possible to find* expresses ability (see, e.g., Groom 2005: 259 for a more detailed discussion of such differences). Following Biber (2006a: 100), participial clauses and *wh*-clauses have been excluded. The subcategory that has been demonstrated to comprise the vast majority of tokens in the category of stance adverbials (Biber et al. 1999: 982), *single adverbs*, was selected for analysis; the members of this category were searched for individually. Tokens belonging to the stance noun + PP construction were identified by a search for the corresponding noun followed by *of*. An overview of the sets of grammatical stance markers and the search patterns used can be found in Table 2.

It was considered important for the analysis that the members of these sets be sufficiently similar not only morphologically, but also semantically; the adverbial form of ESSENTIAL* (i.e., *essentially*) was therefore excluded, as this form was considered to be too far removed semantically to merit inclusion in the set. Moreover, since not all of the members of these sets allow for negative prefixes (e.g., **uncrucial*, **uncurious*), this angle was not pursued further.

Any invalid tokens caught by the search patterns were excluded manually; examples of such excluded tokens include tokens where *to* is prepositional (1) rather than an infinitival marker (2) or where *likely* is adjectival (3) rather than an adverb (4). Instances where the stance markers were used as linguistics examples in the texts, as in (5), were also removed.

Table 2. The sets of grammatical stance markers included in the study

Function	Base form	Stance complement clause constructions	Stance adverbs	Stance noun + PP
Hedges	POSSIB*	possible that	possibly	possibility of
	LIKEL*	likely that	likely	likelihood of
	PROBAB*	probable that	probably	probability of
	CONCEIVAB*	conceivable that	conceivably	conceivability of
Boosters	IMPORTAN*	important to/that	importantly	importance of
	ESSENTIAL*	essential to/that	*NA*	essentialness of
	CRUCIAL*	crucial to/that	crucially	crucialness of
	IMPERATIVE*	imperative to/that	imperatively	imperativeness of
Attitude markers	INTEREST*	interesting to/that	interestingly	interest of
	NOTEWORTH*	noteworthy to/that	noteworthily	noteworthiness of
	CURIOUS*	curious to/that	curiously	curiosity of

(1) […] but the latter is **essential to** effective social care. (BNC_ALN_academic)

(2) In these circumstances **it is essential to ensure** that the foster mother has finished. (BNC_ EV6_academic)

(3) […] the Jewish American subject was therefore a **likely** convention. (MICUSP_555.G2.06.1)

(4) […] they have most **likely** studied it since the age of seven. (ALEC_4.003)

(5) Predicate adjectives, e.g., sure, certain, **probable, likely, conceivable**, doubtful. (BAWE_6038a)

The statistical environment *R* (R Core Team 2018) was used for data management and to test the differences found for statistical significance. A multinomial log-linear model was fitted to investigate to what extent any independent variable (in this case, *base form*; see Section 4.1) affects a dependent, categorical variable with more than two levels (in this case, *construction*; see Section 4.1). Put in another way, the model investigates to what extent there is a correlation between the base forms investigated (e.g., INTEREST*) and the constructions (e.g., stance adverbs). Moreover, multiple correspondence analyses (MCA) were used to detect patterns in the data. MCAs make use of a multivariate space reduction technique for exploratory investigations of categorical data. Since the present study seeks to summarize and identify underlying structures in the data and since all the factors

investigated here (base form, construction and L1) are categorical, such analyses fit the objectives of the present study well. MCAs enable plotting of variables on a two-dimensional plane, where the distance indicates degree of similarity between variables; the closer the distance, the stronger the correlation. To make this possible, the frequencies of the variables under investigation are converted into matrices of distance (between rows and columns, respectively), which are subsequently plotted on the plane (see, e.g., Glynn 2014; Baayen 2008:128ff). Put more simply, this method provides an overview of how all the data pattern, thereby enabling researchers to draw conclusions about, for example, which registers are most similar and which base forms and constructions are associated with which register(s).

4. Results and discussion

In this section, the frequency distribution across registers in the BNC-15 will be discussed in Section 4.1 and subsequently compared to those of the apprentice writers in Section 4.2. Section 4.3 provides a concluding summary and discussion.

4.1 Distribution across registers

In total, there were 1,992 valid tokens in the BNC-15 data. The distribution of the morphologically related stance markers under scrutiny can vary on two different axes: a lexical one and a syntactic one; however, based on previous research (e.g., Larsson 2017a), we can also expect there to be an interaction between the two. In this subsection, the distribution across base form (i.e., the lexical distribution) will be presented first, after which the distribution across construction (i.e., the syntactic distribution) will be displayed; the interactions will subsequently be explored.

The proportion of each base form per register is shown in Figure 2. The first thing to note is that there are clear differences between the base forms across the registers, suggesting that lexis is an important factor to take into consideration when wishing to draw conclusions about stance markers of this kind, in line with the findings in Larsson (2017a).

Examples of the five most frequent base forms, PROBAB*, POSSIB*, IMPORTAN*, LIKEL* and INTEREST* can be found in (6)–(10) below.

(6) […] I'm **probably** not going to get a hundred pounds at the office […].
(BNC_KCB_conversation)

(7) […] the man might just **possibly** have meant this in a humorous sort of way […].
(BNC_AR3_fiction)

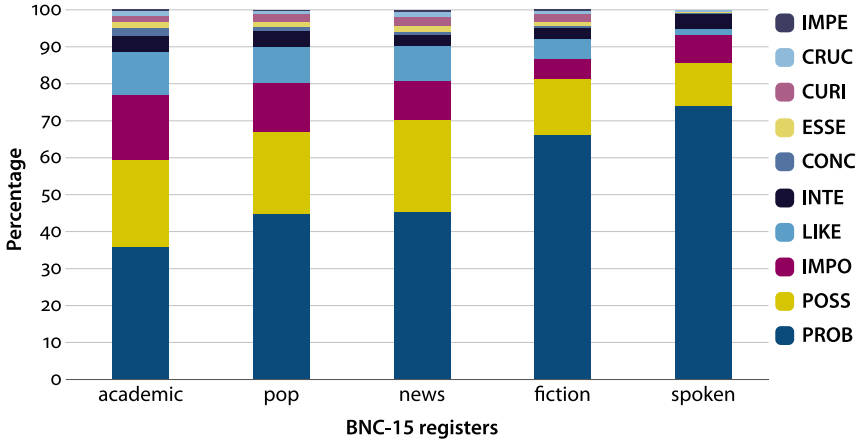


Figure 2. Proportions of each base forms across the BNC-15 registers

(8) **The importance of** these sources of income has implications for the income in old age […]. (BNC_CKP_academic)

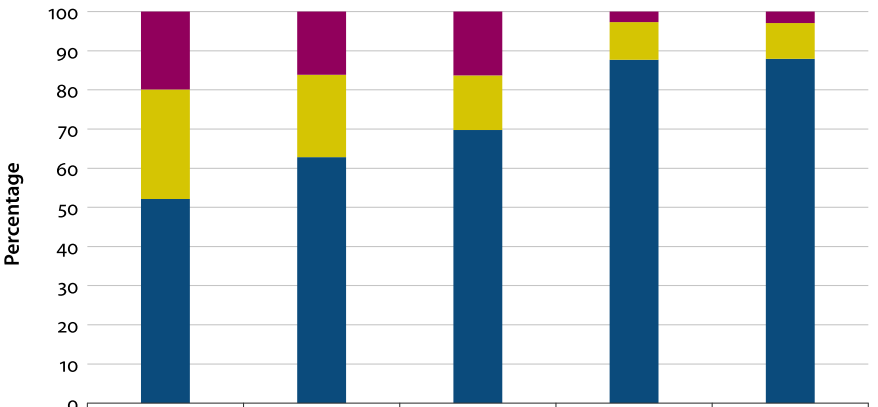
(9) **It is likely that** this group was betrayed by an informer. (BNC_A5R_news)

(10) **Interestingly**, there are three species of waxbills that share identical mouth markings […]. (BNC_CJ3_popular science)

Something else worth noting is that the base form PROBAB* (*probable*, *probably*, *probability*) represents the lion's share of the tokens, accounting for approximately half of the tokens (1,036/1,992 = 52 percent). In fact, PROBAB* is the most frequent base form in all five registers. In terms of the proportional distribution, however, this base form is most strongly associated with conversation; it makes up no less than 74 percent (326/440) of the tokens in this register, compared to only 36 percent (203/568) in academic prose.

Among the remaining base forms, by contrast, there is a clear trend visible for the raw frequencies in the data, namely that of a gradual decline in frequency corresponding largely to the *a priori* order of registers in decreasing level of formality, with academic prose displaying the largest variability and the spoken data the least variability. This gradual cline is even more clearly visible when we turn to the normalized frequencies of the constructions (i.e., the syntactic distribution) across registers, as shown in Figure 3.

The proportion of stance adverbs decreases with the *a priori* expected level of formality, whereas the proportion of stance complement clause constructions (henceforth: *comp*) and noun phrase + prepositional phrase constructions (henceforth: *NP+PP*) increases. The high frequency of stance adverbs in particular in the spoken data is in line with Biber et al.'s (1999: 979) findings, where the full list of



	academic	pop	news	fiction	spoken
NP+PP	113	70	41	8	13
comp	159	91	35	29	40
adverb	296	272	175	263	387

Figure 3. Proportions of each construction across the BNC-15 registers

stance markers was studied in four of the five registers included in the present study, namely academic prose, news, fiction and conversation. These results thus suggest that the relative frequency of adverbs (i.e., the proportion of adverbs per register) may be used as an indicator of formality.

As discussed above, there are clear differences across registers with regard to both the base forms (lexis) and the constructions (syntax). These factors are, however, intertwined, as the base forms can be expected to have varying constructional preferences (cf. Larsson 2017a). To further investigate such interactions, a multinomial log-linear model was fitted onto the data; the confidence intervals of the model output can be found in the Appendix. The model was fitted only for the five base forms with the highest frequencies: IMORTANT*, INTEREST*, POSSIB*, PROBAB* and LIKEL*; as was shown in Figure 2, the remaining base forms exhibited very low frequencies. The results are summarized in an effect plot from the effects package (Fox & Hong 2009) in R; the plot can be found in Figure 4. The effect plot displays the predicted probabilities for how likely a given base form is to be realized through one of the three constructions (*NP+PP*, *comp* or *adverb*); the probabilities add up to 1.0 vertically. The shaded bands mark the confidence intervals.

As can be seen from the figure, there is clear lexico-grammatical variation, thereby offering further evidence to support the claim that it is important to take the lexical dimension into consideration. For example, the predicted probability of the base form PROBAB* being realized as a stance adverb is almost 1, whereas

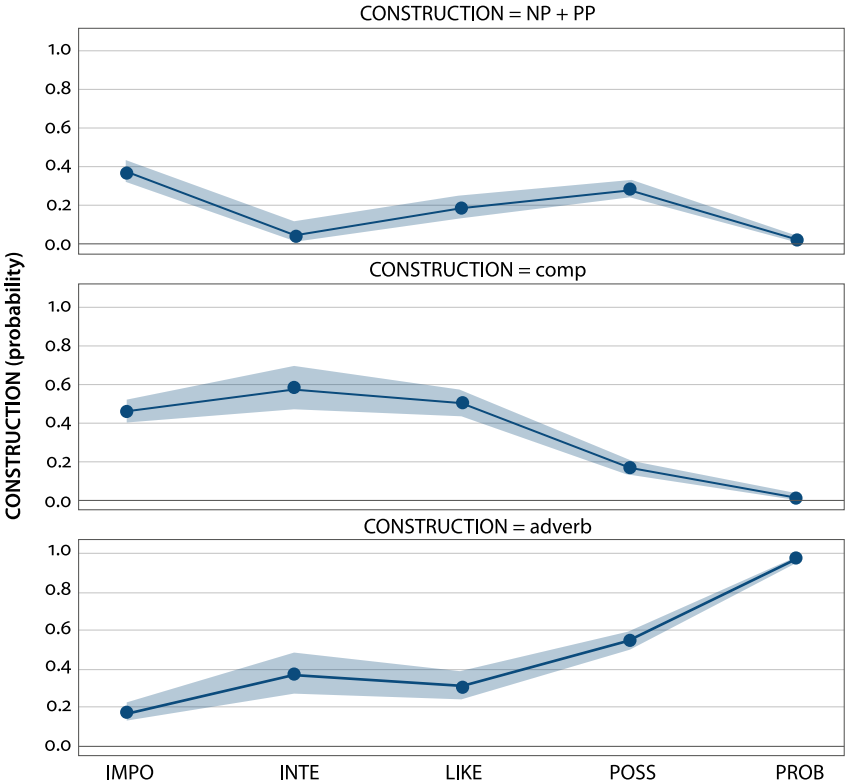


Figure 4. Distribution across the base forms for each construction for the BNC-15 registers

base forms such as IMPORT* and LIKEL* do not show the same preference for one single form. The base form INTEREST* is predicted to be almost equally likely to be realized as a stance adverb as through the stance complement construction.

A closer look at the data shows that these lexico-grammatical preferences also appear to be register-specific to a large degree. For example, while the news data include a comparatively large proportion of POSSIB* as realized through the *NP+PP* construction (11), the academic prose data display a comparatively large proportion in particular of POSSIB* as realized through the *comp* construction, as exemplified in (12) below.

(11) **The possibility of** Transkei becoming a base of operations for the ANC does not, suddenly, seem far-fetched. (BNC_A28_news)

(12) **It is possible that** sea water may exert a chemical action […]. (BNC_GVo_academic)

In order to summarize all the underlying dimensions statistically across all the registers, an MCA was fitted onto the data (Figure 5), using the R package *FactoMineR* (Le, Josse, & Husson 2008). As can be recalled from Section 3.2, this kind of correspondence analysis is used to display underlying structures in the data by mapping these onto a two-dimensional space, where spatial proximity indicates similarity between the variables.

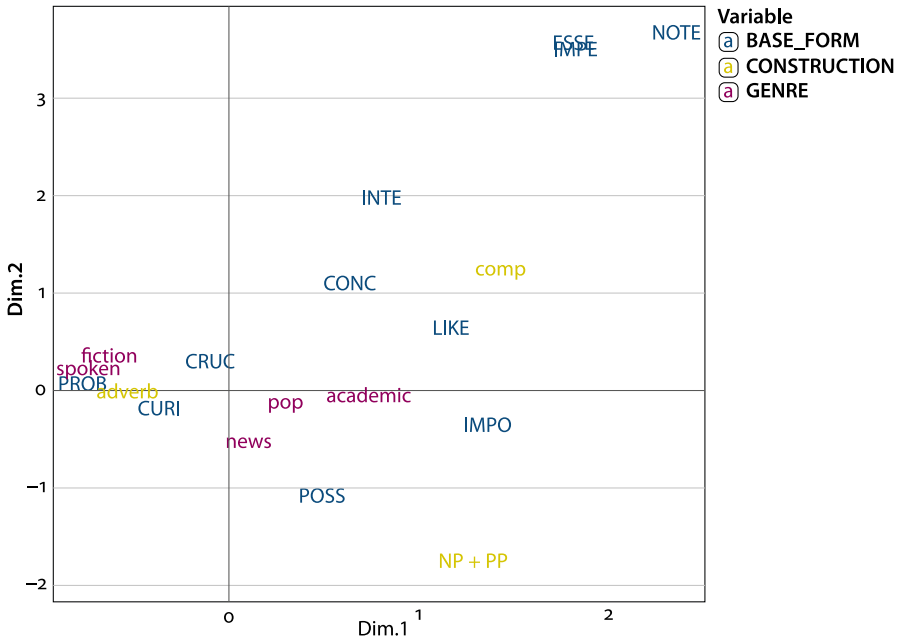


Figure 5. MCA of the base forms and constructions across the BNC-15 registers

As can be seen, with all the underlying (raw and relative) frequencies summarized through an MCA, the registers are plotted in descending order of formality along the x-axis (Dimension 1). The two *a priori* most formal registers, academic prose and popular science, display similar behavior; this is also the case for the two most informal registers, fiction and conversation, which is evident from the fact that they are plotted closely to one another. It can also be noted that both the adverb construction and the base form PROBAB* (and, to a lesser extent, CURIOUS*) are strongly associated with fiction and the spoken data, as these are plotted very close to these registers.

With all the comparisons taken into consideration, it can thus be concluded that the distribution of stance markers investigated corresponds largely to the *a priori* mapping of the registers onto the informal-formal continuum, meaning that the distribution is systematic and thus seems to be useful for characterizing the

(in)formality of the apprentice writing. I will now turn to the comparison between expert and apprentice writing; here, the results will be presented separately for the three functional categories: hedges, boosters and attitude markers.

4.2 Distribution across the expert and apprentice writing

In the present section, a frequency overview will be provided in Section 4.2.1, followed by a more detailed comparison between the apprentice writing and the registers from BNC-15 across the functional categories in Sections 4.2.2–4.2.4.

4.2.1 *A frequency overview*

In addition to the 1,992 valid tokens found in the BNC-15 data, there were 1,073 valid tokens in the learner apprentice writing and 516 valid tokens in the NS apprentice writing, thus adding up to 3,581 tokens in total. The dispersion across the texts (normalized per 10,000 words) is summarized in boxplots using the R package *ggplot2* (Wickham 2009) in Figure 6. The boxes show the interquartile range of the data. The median and mean are marked by a vertical black line and an x, respectively; the notches display the confidence intervals around the means. The dots represent outliers.

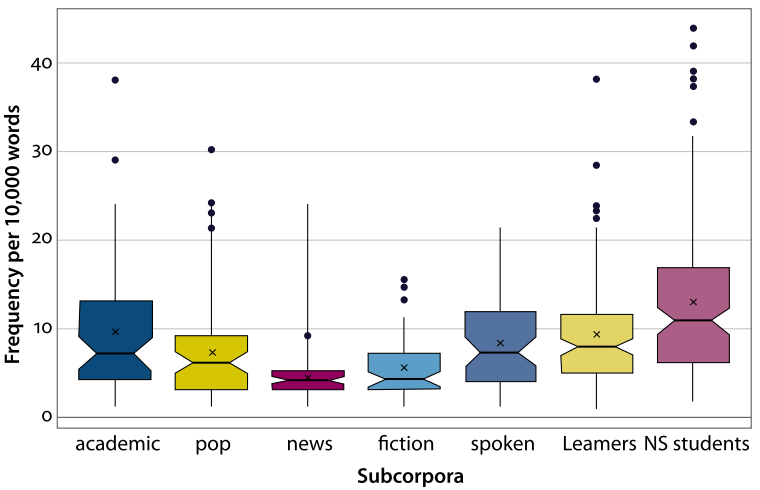


Figure 6. Boxplot of the per-text frequencies (per 10,000 words) of all valid tokens across the subcorpora

As can be seen, while some differences can be noted between the subcorpora, the range between the medians is not particularly large; the median is somewhere between 4 and 7 occurrences per 10,000 words in all the subcorpora but the NS student subcorpus, where the median is just over 10 per 10,000 words.

4.2.2 *Hedges*

In this subsection, the hedges (PROBAB*, LIKEL*, POSSIB* and CONCEIVAB*) are placed under further scrutiny. Of the three functional categories investigated, the hedges category is the largest, comprising 2,415 (67 percent) of the total number of valid tokens. To obtain an overview of how the apprentice writers' usage patterns in relation to the registers, an MCA was fitted onto the data; the output can be seen in Figure 7.

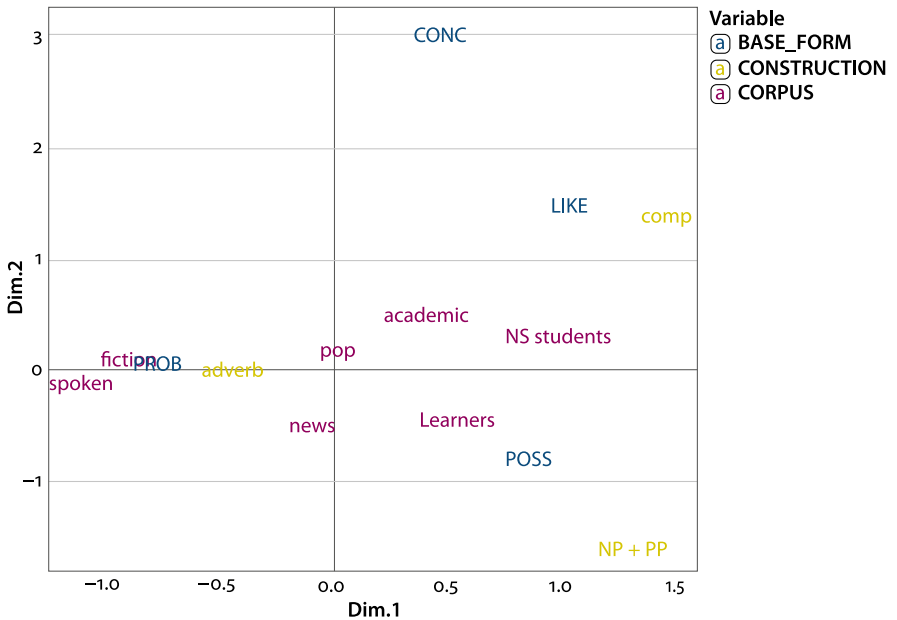


Figure 7. MCA of the base forms and constructions across the subcorpora for the hedges

As can be seen, while neither the NS students' nor the learners' use of hedges is identical to that of the academic experts, the NS students' use is slightly closer overall. However, this is mainly the case on dimension 2 (displayed on the y-axis), where the NS students' use is found to be in-between the experts' academic writing and the experts' popular science writing, whereas the learners' use is most similar to that of the experts' news texts. On dimension 1 (displayed on the x-axis), by contrast, the learners' use can be found in-between that of the academic experts and that of the NS students. It would thus seem that the apprentice writers' use bears resemblance to three of the expert registers: academic prose, news and popular science.

A closer look at the data shows that the main differences between the academic experts and the apprentice writers pertain to overall frequencies, lexical

preferences and – to a lesser degree – constructional preferences. With regard to the overall frequencies, the apprentice writers' numbers are more similar to those of the popular science texts than to those of the academic texts (academic: 693 instances per million words (pmw); popular science: 558 pmw; NS students: 538 pmw; learners: 513 pmw). Here, we can note in passing that these results thus converge with the results of previous studies where learners with different L1s have been found to make less frequent use of hedges than NS students (e.g., Hinkel 2005; Larsson 2017b); however, one should bear in mind that the proficiency levels and the set of hedges investigated differ between the studies.

With regard to lexical preferences, the apprentice writers generally make less frequent use of the base form PROBAB* than the academic experts. Instead, the learners make comparatively more frequent use of the base form POSSIB*, and the NS students make more frequent use of LIKEL*. Nonetheless, in terms of how these base forms are used in the discourse, the apprentice writers' usage is very similar to that of the academic expert writers in that the base forms included in this functional category are most often used to comment on their own or other researchers' findings or claims. Some examples of each of these base forms in their most common instantiations can be found in (13)–(15).

(13) […] it **probably** takes much more cultural energy to teach bellicosity […]. (BNC_HTP_academic)

(14) These results could **possibly** indicate a perception similar to that found in previous studies […]. (ALEC_3.009)

(15) His reading was **likely** not unique in the connection it draws […]. (MICUSP_201.1)

Nonetheless, as the comparatively higher frequencies of POSSIB* and LIKEL* found in the apprentice writing is unlike any of the BNC-15 registers, this could serve as a reminder of the fact that there still are features that are unique to apprentice writing, and that are thus not possible to map onto the register continuum of the reference corpus.

When it comes to constructional preferences, the proportions of the three constructions – *adverbs*, *comp* and *NP+PP* – in the apprentice data is somewhere in-between those of the experts' academic writing and the experts' popular science writing. The main difference between the apprentice writers and the academic experts is the slightly more frequent use of the stance complement clause construction in the expert data than in the apprentice data.

All in all, assuming that the academic experts' use can be taken to represent formal writing, the apprentice writers' use is relatively similar to that of the academic experts. However, their use also bears a resemblance to popular science

and news texts, which might indicate somewhat lacking register awareness and, by extension, somewhat insufficient level of formality when it comes to the hedges studied.

4.2.3 *Boosters*

The boosters category comprises 800 valid tokens and is made up of the following base forms: IMPORTANT*, ESSENTIAL*, IMPERATIVE* and CRUCIAL*. However, it should be noted that the most frequent base form, IMPORTANT*, makes up no less than 92 percent (722/800) of the tokens; while interesting, this lexical bias will naturally have an impact on the generalizability of the results to the full category of boosters, which should be kept in mind. Nonetheless, there are clear differences found across the subcorpora, as is shown in Figure 8.

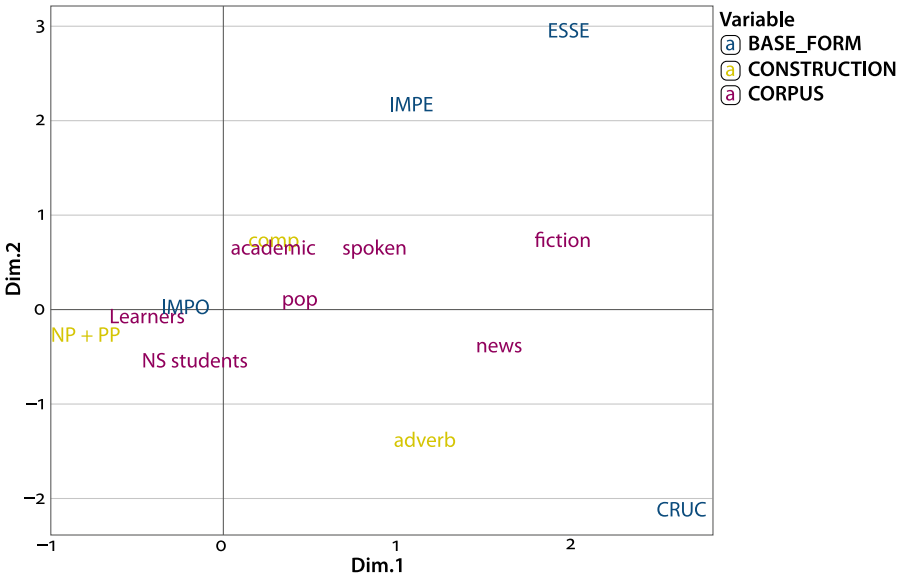


Figure 8. MCA of the base forms and constructions across the subcorpora for the boosters

The first thing to note is that there are clear similarities between the two apprentice writer groups with regard to all the underlying structures in the data, as these are plotted together in the bottom-left corner of the graph, separate from the BNC-15 registers. These results thus seem to be in line with those of Römer (2009) and Larsson (2018) in which a subcategory of these stance markers was investigated and only very minor differences across NS status were noted.

Of the BNC-15 registers, the apprentice writers' usage is closest to that of the academic and popular science texts. The main explanation for this is that the

apprentice writers, like the academic and popular science expert writers, make frequent use of IMPORTANT* in its stance complement clause construction; some examples of this lexico-grammatical preference can be found in (16)–(18).

(16) **It is also important that** the thin sections counted are representative of the sequence studied […]. (BNC_H9S_academic)

(17) […] **it is important that** there is no confusion […]. (BAWE_6009c)

(18) **It is however important to** understand the complexity of each culture […]. (VESPA-SE_0120)

However, there are also clear differences between the apprentice writers' use and that found in the BNC-15 data (regardless of register), which could explain the fact that the two apprentice groups are grouped closely together in their own corner of the MCA graph. One such difference is the fact that the apprentice writers made considerably more frequent use of the boosters category as a whole (325 instances pmw in the learner data and 348 instances pmw in the NS student writing); the closest register with regard to frequencies is academic prose, which includes 197 instances pmw. Another difference is the comparatively high reliance on the base form IMPORTANT* in relation to the other base forms; this base form makes up 94 percent of the tokens in each of the apprentice groups, which can be compared to 84 percent in the experts' academic prose. Thus, while the learners' usage shares features with some of the BNC-15 registers, their usage is, in fact, most similar to that of the NS students (and vice versa).

4.2.4 *Attitude markers*

There were 366 valid tokens in the attitude markers category, thereby making it the smallest of the three functional categories. This category includes the base forms INTEREST*, CURIOS* and NOTEWORTH*, where the most frequent base form, INTEREST*, makes up a large proportion of the tokens (317/366 = 87 percent). An MCA of all the variables is presented in Figure 9.

As can be seen, unlike the other functional categories, here the apprentice writers' usage is most similar to that of the spoken data. A closer look at the results shows that the main explanation for these similarities seems to be that the apprentice writing, like the spoken data, includes a high proportion of INTEREST* in relation to the other base forms (learner data: 93 percent; NS student data: 93 percent; BNC-15 spoken data: 100 percent), which would explain why this base form is plotted in close proximity to these subcorpora. Furthermore, the stance complement clause construction of INTEREST*, as in Example (19), is frequent in particular in the learner data.

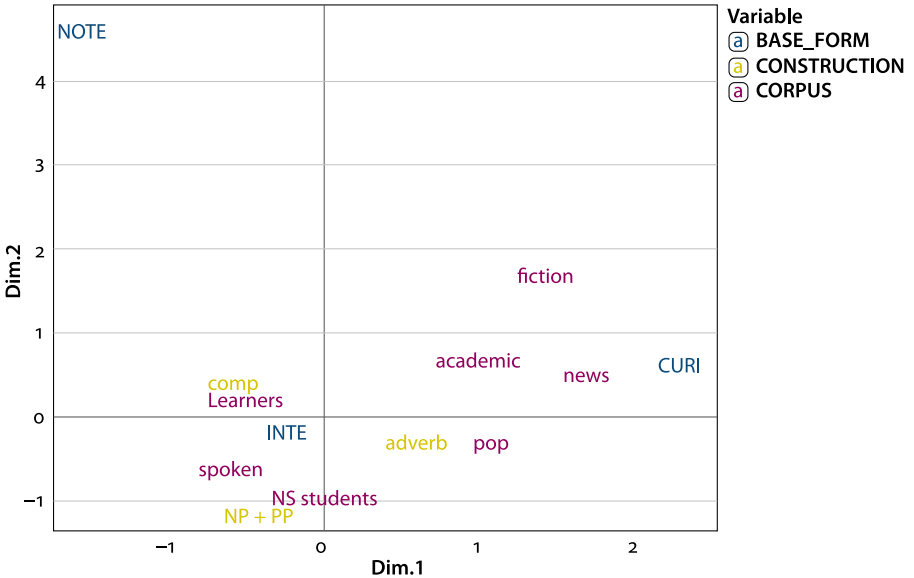


Figure 9. MCA of the base forms and constructions across the subcorpora for the attitude markers

(19) […] **it is thus interesting that** the two last points fit rather well with the encounter […]. (ALEC_3.046)

It should also be mentioned that the apprentice writers, and in particular the learners, made considerably more frequent use of the attitude markers category than the experts; there were 173 instances pmw in the learner data and 146 pmw in the NS student data, but only 57 pmw in the expert academic data, which was the register with the most occurrences.

Thus, assuming that similarity to the experts' academic writing would constitute appropriate register awareness, both apprentice groups once again exhibited somewhat lacking register awareness, as operationalized in the present study. Nonetheless, what is noteworthy about the attitude markers category is that this is the only category where the apprentice writers' usage is closest to the *a priori* least formal register. This means that if we would have treated formality as a binary factor (informal vs. formal) and if 'spoken-like' usage is to be considered informal usage, then the boosters category would have been the only category where the apprentice writers' usage would have been counted as informal. By extension, any discrepancies between the apprentice writers and the academic experts for the other functional categories would have been overlooked. The fact that other differences between the apprentice and expert writing were found in the present study therefore suggests that a more nuanced view of formality is preferable.

4.3 Concluding discussion

In the first part of this study, the aim was to investigate the lexico-grammatical distribution of sets of morphologically related stance markers across registers; in the second part, NNS and NS apprentice data was added to explore which of the registers the apprentice writers' usage was closest to. The results show that the lexico-grammatical preferences vary across registers; for example, the base form PROBAB* (in particular in its stance adverb form: *probably*) is most strongly associated with spoken data, whereas LIKEL* and POSSIB* (in particular when realized through the stance complement clause construction; e.g., *it is likely that* and *it is possible that*) are strongly associated with the academic register. Thus, while the syntactic distribution across register is largely in line with previous research, the present study shows that the lexical dimension is also important, suggesting that the latter deserves more attention in studies of this kind than has previously been the case. Furthermore, a cline with regard to these lexico-grammatical preferences was found across the registers that corresponded to the *a priori* ordering of the registers from more informal to more formal. This suggests that the use of the stance markers varies systematically across the registers.

When the apprentice data was added to the analysis, it became clear that the learners and the NS students exhibited surprisingly similar behavior with regard to the stance markers investigated. This suggests that factors other than the oft-investigated factor *NS status* might be of more interest to the field, although it should of course be kept in mind that only one L1 was investigated in the present study. Furthermore, while the apprentice writers generally used the stance markers similarly to the academic experts, they also shared many preferences with the other registers, which could be seen as indicating somewhat lacking register awareness, and, by extension, a somewhat insufficient level of formality. For two of the three functional categories, hedges and boosters, the apprentice writers' use was found to be somewhere in-between the experts' academic prose, popular science and news, which is perhaps not surprising given that these registers can be expected to be important sources of written input for these students. Only for the least frequent functional category, attitude markers, was the apprentice writers' use most similar to that in the least formal register, conversation.

Two things follow from these results. First, oft-cited claims about the 'spoken-like' nature of apprentice language (and in particular learner language) were only found to be accurate for one of the categories investigated. Second, the traditional view of (in)formality as binary lacks nuance, which speaks to the usefulness of an approach such as the one taken in the present study. These findings will, however, need to be investigated and confirmed by more large-scale studies. Further avenues for future research include studies of other features of writing, in more

L1s, taking other factors into consideration. For example, since it falls outside the scope of the present study to investigate possible disciplinary differences, investigations of this factor in other data would serve as an important complement to the results of this study.

The results can also be used in the English for Academic Purposes (EAP) classroom when teaching students about how to successfully use evaluative language in their academic writing. It can be noted here that both student groups – NNS *and* NS – exhibited some preferences that could be described as non-target like in an academic context, such as their strong preference for the base form INTEREST*. This suggests that both student groups would benefit from targeted teaching of stance marking. All in all, it is hoped that the results of this study will contribute to nuancing discussions of (in)formality, thereby benefitting both EAP instruction and theories of stance marking.

Acknowledgements

I am grateful to the editorial team and the anonymous reviewers for their very helpful comments and suggestions. I would also like to express my gratitude to Gregory Garretson at Uppsala University for writing the Perl program used to extract the tokens.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 35–53). Amsterdam: Rodopi.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80–93). London: Longman.

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

British Academic Written English (BAWE). Corpus compiled at the Universities of Warwick, Reading and Oxford Brookes in 2004–2007. http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

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

Biber, D. (2006b). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., & Zhang, M. (2018). Lexis and grammar as complementary discourse systems for expressing stance and evaluation. In M. Gómez González & J. L. Mackenzie (Eds.), *The construction of discourse as verbal interaction* (pp. 201–226). Amsterdam: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Biber, D., & Zhang, M. (2018). Expressing evaluation without grammatical stance: Informational persuasion on the web. *Corpora*, 13(1), 97–123.

Burnard, L. (2007). *Reference guide for the British National Corpus (XML edition)*. <http://www.natcorp.ox.ac.uk/docs/URG/>

Callies, M. (2013). Agentivity as a determinant of lexico-grammatical variation in L2 academic writing. *International Journal of Corpus Linguistics*, 18(3), 357–390.

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes*, 25(3), 310–331.

Council of Europe. (2001). *Common European framework of reference for languages: Learning teaching assessment*. Cambridge: Cambridge University Press.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 520 Million Words, 1990-present. <http://corpus.byu.edu/coca/>

Fox, J., & Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, 32(1), 1–24. <http://www.jstatsoft.org/v32/i01/>

Glynn, D. (2014). Correspondence analysis: An exploratory technique for identifying usage patterns. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods in cognitive semantics: Quantitative studies in polysemy and synonymy* (pp. 443–485). Amsterdam: John Benjamins.

Gray, B., & Biber, D. (2012). Current conceptions of stance. In K. Hyland & C. S. Guinda (Eds.), *Stance and voice in written academic genres* (pp. 15–33). Houndmills: Palgrave Macmillan.

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, 4(3), 257–277.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). London: Edward Arnold.

Hasselgård, H. (2015). Lexicogrammatical features of adverbs in advanced learner English. *International Journal of Applied Linguistics*, 166(1), 163–189.

Herriman, J., & Boström Aronsson, M. (2009). Themes in Swedish advanced learners' writing in English. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 101–120). Amsterdam: John Benjamins.

Hewings, M., & Hewings, A. (2002). 'It is interesting to note that…': A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes*, 21(4), 367–383.

Hinkel, E. (2005). Hedging, inflating and persuading in L2 academic writing. *Applied Language Learning*, 15, 29–53.

Hunston, S., & Thompson, G. (Eds.). (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Hyland, K. (1996). Talking to the academy: Forms of hedging in science research articles. *Written Communication*, 13(2), 251–281.

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

Kaatari, H. (2017). Adjectives complemented by *that*- and *to*-clauses: Exploring semantico-syntactic relationships and genre variation (Unpublished doctoral dissertation), Uppsala University, Uppsala, Sweden.

Larsson, T. (2014). Introducing the Advanced Learner English Corpus (ALEC): A new learner corpus. Poster presented at the *2014 LOT Winter School*, VU University Amsterdam, Amsterdam, The Netherlands, 20 January, 2014.

Larsson, T. (2016). The introductory *it* pattern: Variability explored in learner and expert writing. *Journal of English for Academic Purposes*, 22, 64–79.

Larsson, T. (2017a). *The importance of*, *it is important that* or *importantly*? The use of morphologically related stance markers in learner and expert writing. *International Journal of Corpus Linguistics*, 22(1), 57–84.

Larsson, T. (2017b). A functional classification of the introductory *it* pattern: Investigating academic writing by non-native-speaker and native-speaker students. *English for Specific Purposes*, 48, 57–70.

Larsson, T. (2018). Is there a correlation between form and function? A syntactic and functional investigation of the introductory *it* pattern in student writing. *ICAME Journal*, 42(1), 13–40.

Larsson, T. & Kaatari, H. (2019). Extraposition in learner and expert writing: Exploring (in)formality and the impact of register. *International Journal of Learner Corpus Linguistics*, 5(1), 33–62.

Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.

Lee, D. (2001). Genres, registers, text-types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3), 37–72.

Michigan Corpus of Upper-level Student Papers (MICUSP). Ann Arbor: The Regents of the University of Michigan. Corpus compiled at the University of Michigan in 2009. http://micusp.elicorpora.info/about-micusp

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Houndmills: Palgrave Macmillan.

Olohan, M. (2003). How frequent are the contractions? A study of contracted forms in the Translational English Corpus. *International Journal on Translation Studies*, 15(1), 59–89.

Paquot, M., Hasselgård, H., & Oksefjell Ebeling, S. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead* (pp. 377–388). Louvain-la-Neuve: Presses universitaires de Louvain.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107–118). London: Longman.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, 7, 140–162.

Smith, E. L. (1986). Achieving impact through the interpersonal component. In B. Couture (Ed.), *Functional approaches to writing* (pp. 108–119). London: Frances Pinter.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Appendix

The confidence intervals to accompany the effects plot in Section 4.1.

Adverb* vs. *comp				
	2.5	**%**	**97.5**	**%**
(Intercept)	0.5990038		1.31892986	
BASE_FORMINTE	−1.1090264		0.06979327	
BASE_FORMLIKE	−0.9878003		0.02531584	
BASE_FORMPOSS	−2.5889198		−1.68176518	
BASE_FORMPROB	−5.7837768		−4.53556525	
Adverb* vs. *NP+PP				
	2.5	**%**	**97.5**	**%**
(Intercept)	0.3435477		1.0902888	
BASE_FORMINTE	−4.2312818		−1.7395862	
BASE_FORMLIKE	−1.8331458		−0.6494819	
BASE_FORMPOSS	−1.8303167		−0.9525997	
BASE_FORMPROB	−5.2090350		−4.0508970	

Address for correspondence

Tove Larsson
Department of English
Uppsala University
Box 527
SE-75120 Uppsala
Sweden
tove.larsson@engelska.uu.se