Stylistic Differences between Closely Related Disciplines: Metadiscourse in German Linguistics and Literary Studies

Melanie Andresen[1]
Universität Hamburg

Heike Zinsmeister[2]
Universität Hamburg

Abstract

The disciplines of linguistics and literary studies are often considered similar, as they are for instance part of one common study program. However, there are many differences between the two disciplines that concern not only the object of study, but also research methods and writing styles. Consequently, students of, for example, German Studies need to adapt to two academic languages at once. Therefore, we aim to describe the stylistic differences between the languages of the two disciplines. Our study is based on a data-driven n-gram analysis of German PhD theses that reveals a more intense use of metadiscourse in linguistics when compared to literary studies. In the light of these results, we carry out a more in-depth study of metadiscourse in the two disciplines, focusing on the expressions *im Folgenden* ("in the following") and *zusammenfassend* ("summarizing"). We find that literary scholars use both of the above expressions less frequently than linguists. We suggest that this might be due to different aesthetic demands and more influence of English academic language on German linguistics. Also, a higher proportion of the instances found in literary studies turned out to be intertextual rather than metatextual. We therefore argue that it is important to inspect the data behind quantitative results in detail.

Keywords
Metadiscourse • Academic writing • German • Linguistics • Data-driven analysis

1 **Correspondence to:** Melanie Andresen, Universität Hamburg, Institut für Germanistik, Grindelallee 46, 20146 Hamburg Germany. Email: melanie.andresen@uni-hamburg.de

2 Heike Zinsmeister, Institut für Germanistik, Universität Hamburg, Überseering 35, Postfach #15, 22297 Hamburg. Email: heike.zinsmeister@uni-hamburg.de

Many universities offer study programs such as German Studies or English Language and Literature. These study programs comprise two disciplines, linguistics and literary studies.[3] From this perspective, these two disciplines appear to be very closely related. However, in other respects these disciplines are very distinct. They ask different questions, use different methods and, and this is going to be the focus of this article, they use language in different ways. This means that students of one of the aforementioned study programs are expected to adapt to the writing conventions of both disciplines simultaneously, e. g. when working on written assignments. To assess the difficulty of this task, we approach the following research question: How do German academic texts of literary studies and linguistics differ stylistically?

In order to answer this broad question, we first conduct a data-driven analysis based on the frequency of (co-occurring) words (n-gram analysis) that will be described in Section 3.[4] One of the results of this data-driven analysis is that several patterns that realize metadiscourse emerge as relevant for the distinction between the disciplines. This is taken as a starting point for two hypothesis-driven case studies that focus on the specific text comments *im Folgenden* ("in the following", Section 4) and *zusammenfassend* ("summarizing", Section 5). We compare the frequency of these expressions as well as different types of use and their combination with modals and main verbs. Both case studies confirm the fact that linguistics and literary studies use metadiscourse differently. Section 6 will present conclusions and elaborate on possible explanations for the differences found between the two disciplines.

Previous Work

In this section, we will first situate our object of study typologically and then discuss previous research on disciplinary differences in metadiscourse. There are two main concepts of metadiscourse (Ädel & Mauranen, 2010): The broader definition of metadiscourse by Hyland (2005, p. 37) refers to "the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community". In his typology, the expressions we are interested in are part of interactive metadiscourse, whose function is to "help to guide the reader through the text" (Hyland, 2005, p. 49). This is differentiated from the interactional dimension of metadiscourse with the purpose of "involv[ing] the reader in the text" (Hyland, 2005, p. 49). More specifically, we are interested in so-called frame markers, that inform the reader about content and position of elements in the text such as *finally* or *to conclude* (Hyland, 2005, p. 49).

These are also part of the narrower concept of metadiscourse promoted by Ädel (2006), according to whom metadiscourse "is text about the evolving text, or the

---

3 Sometimes also cultural studies are included as a third discipline.

4 This section is part of a larger research project about the potential of n-grams for describing style. *See* Andresen & Zinsmeister (2017) for more details.

writer's explicit commentary on her own ongoing discourse" (Ädel, 2006, p. 2). This definition focuses on the property of reflexivity. She further distinguishes between metatext and writer-reader interaction. Our topic of interest is metatext that "spells out the writer's (and/or the reader's) discourse acts, or refers to aspects of the text itself, such as its organization and wording, or the writing of it" (Ädel, 2006, p. 36). Ädel explicitly distinguishes metadiscourse from intertextuality: The latter features references to texts as well, but to texts other than the current text (Ädel, 2006, p. 28). This distinction will be important for our analysis in sections 4 and 5.

Fandrych and Graefen (2002) discuss the phenomenon at hand under the label "text comments" (without reference to the term metadiscourse). They compare the use of text comments between German and English research articles and suggest a functional typology to further differentiate subtypes. Among other aspects, they consider whether the expression at hand has a forward or a backward orientation, i. e., whether it refers to a part of the text that appears later or earlier in the text. The examples in our case studies in sections 4 and 5 both have a forward orientation, with *zusammenfassend* ("summarizing") being slightly more complex (see Section 5).

There has not been much research on metadiscourse in the two disciplines under investigation. Afros and Schryer (2009) follow the metadiscourse concept posited by Hyland (2005). They compare promotional (meta)discourse between linguistics and literary studies by analyzing rhetorical strategies of how authors publicize their own work in scholarly texts. They find more "pathos appeals" in literary studies, which in this case means that they address aesthetic values of the community. They state that texts in literary studies are sometimes even "transcending borders with literary genres" (Afros & Schryer, 2009, p. 63). Haggan (2004) compares titles of literary studies, linguistics and science. She finds that those in literary studies follow aesthetic principles and often present "an elegant puzzle […] solvable only by reading the paper" (Haggan, 2004, p. 305), rather than just giving information.

Hyland (2005) looks at a wider spectrum of disciplines and finds that "the more discursive 'soft' fields such as applied linguistics[5] employ more metadiscourse overall" (Hyland, 2005, p. 57) in comparison to "hard" fields such as biology. However, for the specific group of frame markers there is no clear tendency between the fields (Hyland, 2005, p. 162).[6]

Most of these studies on metadiscourse are about English academic language only. German academic language has received much less attention (not to speak of many other languages here). While we would generally expect a high transferability between English and German, many studies have shown cultural differences

[5] Note that applied linguistics is the softest discipline in Hyland's investigation. In our study, we focus on the soft disciplines only and consequently linguistics is the "harder" discipline in relation to literary studies.

[6] Note that this finding by Hyland (2005) is based on textbooks.

in academic writing. For instance Clyne (1987) describes German as using more impersonal structures, hedges, nominalizations and syntactically complex structures when compared to English academic language.[7]

In our study, we focus on German academic language and aim at broadening the knowledge about disciplinary differences in the use of metatext. More specifically, we focus on the humanities disciplines linguistics and literary studies. We assume that literary studies is the "softer" of the two disciplines and that it is more firmly rooted in the German academic tradition. This tradition is characterized as valuing theory and membership to schools of thought very highly. This is accompanied by a language that is intended to challenge the reader intellectually and not to maximize understandability (Clyne, 1987).

Data and Data-driven Analysis

In this section we present our data and the n-gram analysis as the first step of our approach. The data used for the present study is a corpus of 60 German PhD theses submitted at German universities, a subcorpus of 30 texts from each of the two disciplines linguistics and literary studies. The texts were accessible as PDF files and were in a first step converted to HTML. The HTML markup was used to semi-automatically extract parts of the text that do not belong to the targeted variety or interrupt the text: tables and figures, footnotes, citations and examples. The resulting plain text version was the input for the n-gram analysis (for more details on the preprocessing see Andresen & Zinsmeister, 2017).

An n-gram analysis is a data-driven approach that was developed in computational linguistics to model characteristics of a language in a bottom-up way (Jurafsky & Martin, 2009). For the purposes of a linguistic study, this method has the advantage of not requiring any hypotheses about the object of investigation. Instead, noteworthy features (in a quantitative sense) of the texts are identified statistically. The building block of this analysis is an n-gram, which is a sequence of *n* elements, where *n* can be any number, usually between 1 and 5. The elements can be, for instance, characters or words or parts of speech. In the present case, the elements under examination

Table 1
N-grams in the Example Sentence I will go hiking (Ignoring Punctuation)

<s> I will go hiking. </s>		
n=1	unigrams	<s> – I – will – go – hiking – </s>
n=2	bigrams	<s> I – I will – will go – go hiking – hiking </s>
n=3	trigrams	<s> I will – I will go – will go hiking – go hiking </s>
…	…	…

---

are words in the sense of surface-based tokens. Table 1 presents all the n-grams of different sizes that are part of the sentence *I will go hiking* as an example. The items *<s>* and *</s>* mark the beginning and end of the sentence, respectively. They are treated just like words.

Our n-gram analysis consists of three steps: First, we count all possible n-grams in each subcorpus; second, we determine the difference in frequencies between the two subcorpora for each n-gram; and third, we rank the n-grams according to this difference (biggest differences are ranked topmost). The rationale behind this procedure is that n-grams that are more frequent in one of the subcorpora contribute to the distinctive characterization of this subcorpus' discipline. Hence, for the comparison of linguistics and literary studies, we want to know which n-grams show the biggest differences in frequencies between the two subcorpora.

There are many different ways of quantifying this difference in frequency with varying advantages and disadvantages (see Rayson (2003) for an overview). The measure for comparison used here is the log-likelihood measure as presented by Dunning (1993). If the log-likelihood ratio is 0, there is no difference in frequency. The higher the ratio, the clearer is the difference between the two groups. A log-likelihood ratio of 10.83 corresponds to $p < .001$.[8] Theoretically, there is no upper limit for the possible values.[9]

Table 2 shows an example result of such an analysis: The ten most distinctive trigrams that are more frequent in linguistics than in literary studies. They are ranked by their log-likelihood score, starting with the most distinctive instances. Even though n-grams are often fragmentary by nature and consequently not fully translatable, Table 2 also gives an approximate English translation. We can see some complete phrases like *in Bezug auf* ("with regard to") and *in der Regel* ("generally speaking"). Other phrases remain fragmentary, as they are longer than three words. For instance, the trigrams on ranks 3 and 5 are both part of the larger structure *in der vorliegenden Arbeit* ("in the present text"). Semantically the results show some general patterns that are more common in the linguistics subcorpus like *in Bezug auf* ("with regard to"), but also more specific patterns like *die Ergebnisse der* ("the results of"). The latter correspond to the different methodologies of the disciplines as linguists are more likely to report empirical studies. The former are harder to explain functionally and might be due to stylistic preferences developed in the community.

---

8   http://ucrel.lancs.ac.uk/llwizard.html, 07.08.2017

9   One relevant property of log-likelihood is the fact that it is based on word frequencies for each subcorpus as a whole. Consequently, if a word is extremely frequent in one text, this can affect the overall result. This has to be kept in mind, but is not a major problem for the current study as the method is used for hypothesis generation only.

Table 2
Most Distinctive Trigrams More Frequent in the Linguistics Subcorpus

Rank	LLR	German Trigram	English Translation
1	262.16	in Bezug auf	with regard to
2	239.69	<s> bei der	<s> At the
3	236.91	der vorliegenden Arbeit	the present text
4	204.73	in der Regel	generally speaking
5	160.46	in der vorliegenden	in the present
6	156.01	Rahmen der vorliegenden	course of the present
7	155.31	im Hinblick auf	with regard to
8	154.09	Bezug auf die	regard to the
9	150.42	die Ergebnisse der	the results of
10	147.15	<s> bei den	<s> at the

Inspecting high-ranking instances across n-gram sizes, it is striking that many of the patterns more frequent in linguistics are related to metatextual expressions. Table 3 presents the most important instances.[10] The first column indicates the size of the n-gram, the second column gives the n-gram's rank in the corresponding list. All of these (sub)patterns function as text comments, informing the reader where some information was or will be presented. Some are very global and provide information about the text as a whole (*im Rahmen der vorliegenden Arbeit*, "in the present text"). Others have a more narrow scope, referring to one specific section, most commonly the next section (*im Folgenden*, "in the following" and *in Kapitel*, "in chapter"). The sparse use of metatext in literary studies is plausible under the assumption that literary scholars prefer aesthetic principles to facilitating understanding for the reader (cf. Section 2).

Table 3
High-ranking N-grams Related to Metatext

Size	Rank	German N-gram	English Translation
2	10	der vorliegenden	the present
2	19	in Kapitel	in chapter
3	2	der vorliegenden Arbeit	the present text
3	27	<s> im Folgenden	in the following
4	1	im Rahmen der vorliegenden	in the present
4	11	<s> im Folgenden werden	in the following […] will be
4	12	<s> zusammenfassend lässt sich	summarizing it can be […]
5	1	im Rahmen der vorliegenden Arbeit	in the present study
5	2	<s> im Folgenden werden die	in the following the […] will be

The n-gram analysis shows that several text comments are more frequent in the linguistics subcorpus than in the literary studies one. We want to stress the fact that this result emerges from a data-driven analysis that is not targeted at metatext specifically. This leads to two conclusions: First, the use of metatext is an important difference between the two disciplines that should and will be examined more closely in the following. Second, metatext (in linguistics) is realized in a very formulaic way,

---

10 The decision which n-grams have a metadiscoursive function is based on the judgment of one person only.

frequently using the same patterns. A text function can be very frequent in a text, but if it were always realized in different words, a word-based n-gram analysis would not be able to detect it.

Section 4 and 5 present case studies on the two text commenting expressions *im Folgenden* ("in the following") and *zusammenfassend* ("summarizing") and investigate their frequency and use. The following research questions will be addressed exemplary: Are there significant differences in the use of metatext between literary studies and linguistics? What additional insights can be gained by inspecting the concrete instances behind the quantitative results in detail?

Hypothesis-driven Analysis 1: *im Folgenden* ("in the following")

Figure 1 shows two box plots of the frequency distribution of *im Folgenden* for each disciplinary subcorpus. The red boxes mark the frequency areas, where 50% of the texts of each discipline can be found. For instance, 50% of linguistics texts use *im Folgenden* between 6 and 27 times. The bold black line marks the median and the black dot the mean; outliers are marked as dots in the upper part. Note that we decided against using relative frequencies here. For comments on text organization, it is an open question whether we would expect them to occur more often the longer the text. Alternatively the frequency could be dependent on the number of sections in the text. Therefore, we will inspect the absolute frequencies here.
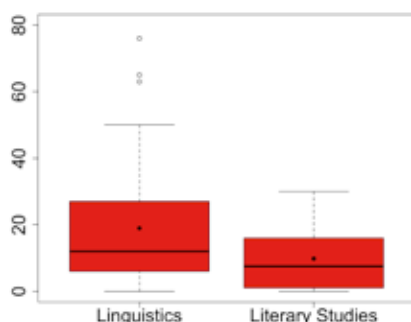

Figure 1. Absolute frequencies of *im Folgenden* ("in the following", n = 30 texts per discipline).

It is visible that the absolute frequencies in linguistics are higher, resulting in a total of 569 instances compared to 294 in literary studies. However, there are several outliers and the boxes also overlap very much, indicating that the frequencies in most texts are in a similar range.

Table 4

Absolute Frequency Distribution of im Folgenden ("in the following")

discipline	mean	sd
Linguistics	18.97	20.17
Literary Studies	9.80	8.95

This is confirmed by the summarizing figures in Table 4: While the difference in mean is considerable, the variance especially in linguistics is also very high, meaning that some authors use *im Folgenden* very often (see the outliers in the upper part of Figure 1) and others hardly at all. Consequently, even though the effect size is high (Cohen's d = 0.59), the difference is not significant (Wilcoxon rank sum test, W = 568.5, *p* = .08).[11]

However, an important distinction that was presented in section 2 has not yet been applied: The pattern *im Folgenden* can be used metatextually, when the text refers to itself, but also intertextually, when the text refers to another text. The following two examples from the corpus illustrate this distinction:

(1) metatextual

Im Folgenden wird auf mögliche Gründe für diese Unterschiede eingegangen.

"In the following, possible reasons for these differences will be addressed."

(Lin_Dui_13)[12]

(2) intertextual

Im Folgenden führt d'Holbach aus, dass […]

"In the following d'Holbach explains […]"

(Lit_Kob_25)

While sentence (1) announces what is to happen in the very same text, sentence (2) clearly refers to another text. In the present discussion of metatext, only the first type is relevant.

To account for this difference in reading, a random sample of 100 sentences per discipline was categorized as metatextual or intertextual[13] and Figure 2 shows the results. As can be seen clearly, the proportion of intertextual instances (marked in black) is much higher in literary studies and a χ^2 test confirms the high significance of the difference (χ^2 = 19.95, df = 1, *p* < .001). An odds ratio of 8.14 shows a very clear effect.

---

11 This means that there might be an effect that cannot be verified given the current sample size and it might be worth looking at a larger sample.

12 The name of a source text is a combination of a short form of the discipline, a short form of the university at which the thesis was submitted and a running number.

13 One of the authors performed the classification. Both authors discussed ambiguous instances.

Thus, we can conclude that linguistics uses *im Folgenden* significantly more often in metatextual function than literary studies does, which confirms our results so far.
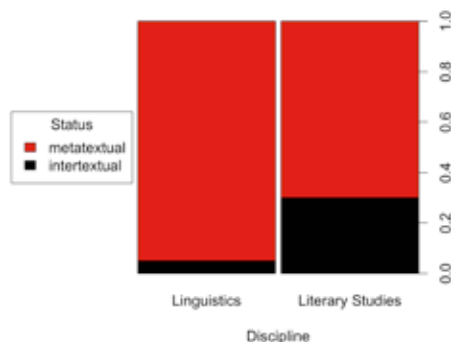

Figure 2. Proportions of metatextual and intertextual use of *im Folgenden* in both disciplines (n = 100 instances per discipline).

Another focus of our investigation is on the verbs used in conjunction with *im Folgenden*. Comparing these verbs in both subcorpora, further differences between the disciplines appear regarding modal verbs on the one hand and main verbs on the other hand.

Modal verbs are very frequent in text comments of German academic language, as Fandrych and Graefen (2002) show when comparing the use of text comments in German and English. The most frequently used modal verb in German is *sollen* (approximately "shall"), which "indicates that the impetus for an action is external, i. e. an agent is required to carry out the will of another person or an institution" (Fandrych & Graefen, 2002, p. 32). In practice, it is used as a hedging device (e. g. Graefen, 2000), indicating "a lack of commitment on the part of the speaker with respect to [the] entire proposition" (Prokofieva & Hirschberg, 2014, p. 32).
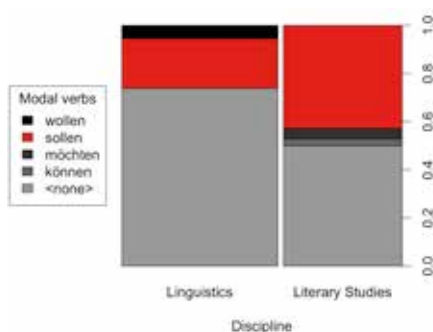

Figure 3. Modal verbs used in metatextual sentences with *im Folgenden* ("in the following", n = 162).

Figure 3 shows the distribution of modal verbs used in sentences with *im Folgenden*. It is a stacked representation of the proportions in which the columns represent the two disciplines. The width of each column is proportional to its discipline's overall frequency counts. Note that *n* is reduced from 200 sentences in the original sample to the 162 metatextual instances only.[14] As literary studies had less metatextual instances, its column is narrower. The two most important groups with respect to modal verbs are sentences without modal verb (segment "<none>") in grey and sentences with *sollen* ("shall") in red. Again, the differences are significant (Fisher's test: $p < .001$): Literary scholars use more modal verbs, especially *sollen*, than linguists. This is in accordance with findings by Hyland (2006, p. 29, among others) that show more use of hedging in the soft disciplines. Even though linguistics would generally be considered a soft discipline, it is less so than literary studies.

The remainder of this section is related to the main verbs used with *im Folgenden*. Table 5 gives an overview of the most frequent verbs for the two disciplines. It is striking that the most frequent verbs in linguistics (*auf etw. eingehen, darstellen, vorstellen,* for translations see Table 5) are communication verbs (also called reporting verbs, among many others by Thompson and Yiyun (1991) and Hyland (2004)), while the most frequent verb in literary studies does not belong to this group (*untersuchen*). This leads us to the hypothesis that linguistics generally uses more communication verbs than literary studies. There is also a theoretical argument for this hypothesis: In linguistics, there is mostly a rather clear distinction between the analysis (as manifest in e. g. data, tables and figures) and the text about this analysis. Literary studies on the other hand use interpretive methods in which this distinction is less clear. The analysis is predominantly manifest in the text itself. We propose that this is why a linguist would rather "present an investigation" in the text while a literary scholar might "investigate" in the text itself.

Table 5

Absolute Frequencies of Main Verbs Used with im Folgenden ("in the following"), Relative Frequencies in Parentheses (Normalized to the Total Number of Sentences)

Verb	Translation	Linguistics	Literary studies
auf etw. eingehen	go into sth.	10 (0.11)	1 (0.01)
untersuchen	investigate	3 (0.03)	10 (0.14)
darstellen	depict	8 (0.09)	3 (0.04)
vorstellen	present	7 (0.08)	2 (0.03)
zeigen	show	6 (0.07)	5 (0.07)
erläutern	explain	5 (0.05)	1 (0.01)
betrachten	consider	4 (0.04)	4 (0.06)
…	…	…	…
	total	92 (1.00)	70 (1.00)

In order to test this hypothesis, we define communication verbs as verbs referring to a situation involving "a speaker S, a listenership H, an utterance with a propositional

14 Of originally 165 metatextual instances, three had to be excluded for the following analysis as they did not have a finite verb.

content Sa(P) and a complex communicative attitude of the speaker E(S)" (Harras, Winkler, Erb, & Proost, 2004, p. 9, our translation). Their instantiation is based on two resources: The *Handbuch deutscher Kommunikationsverben* ("Handbook of German Communication Verbs", Harras et al., 2004) and GermaNet (Hamp & Feldweg, 1997). GermaNet is a lexical-semantic net similar to the English WordNet (Princeton University, 2010) that gives a semantic classification of words. One of the verbal semantic classes is "verbs of communication". We consider every verb a communication verb that is listed in one of these resources.[15]
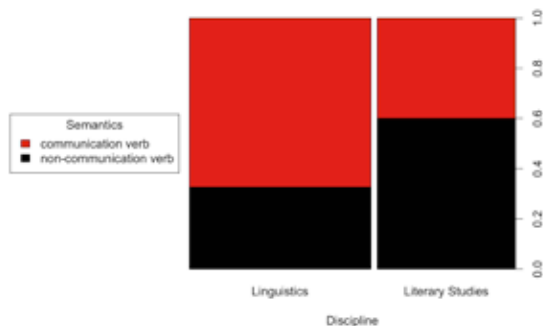

Figure 4. Relation of communication verbs and non-communication verbs in metatextual sentences with *im Folgenden* (n = 162 verbs).

Figure 4 shows the result which indicates a significant difference between the use of communication verbs in metatext between linguistics and literary studies (χ^2 = 11.00, df = 1, *p* = .001, odds ratio = 3.10). The use of communication verbs with *im Folgenden* can be considered a kind of text comment in itself. Fandrych and Graefen (2002) discuss this type as "[i]ntroductory qualification of speech actions". Consequently, it is in line with the other results showing a general tendency for less metatext in literary studies.

Hypothesis-driven Analysis 2: Zusammenfassend ("Summarizing")

For the second case study, the deverbal adverb *zusammenfassend* ("summarizing") was chosen, because it is also very frequent and complementary to *im Folgenden* with respect to the expected position in the text: *Im Folgenden* tends to occur at the beginning of sections and announces something that is still to come, thus being cataphoric. *Zusammenfassend* can in contrast be expected at the end of a text or chapter. However, usually it also refers cataphorically to something that is to come (a summary of what was said before) and we will see below that the two phrases actually cooccur in many sentences.

---

15 This instantiation is only a rough approximation. First, many verbs have several readings and one of them might relate to communication while the other ones do not. Second, German academic language uses many light verb constructions. In these cases, the verb alone might not relate to communication even though the whole light verb construction does. And third, both resources were not developed for academic language specifically.
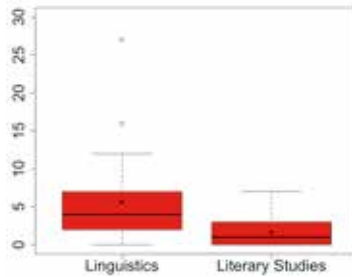
Figure 5. Absolute frequencies of *zusammenfassend* ("summarizing", n = 30 texts for each discipline).

Figure 5 shows boxplots of the frequency distributions of *zusammenfassend* in the texts of both subcorpora. Again, the bold lines mark the medians and the black dots mark the means. Compared to *im Folgenden*, the difference is even clearer as the two boxes hardly overlap. With a total of 167 instances in linguistics and 49 in literary studies, *zusammenfassend* is much less frequent than *im Folgenden*. The numbers in Table 6 confirm the visual impression of a clear difference. There is much less variation than in case study 1, the means differ significantly (Wilcoxon rank sum test: $W = 699$, $p < .001$) and the effect is relevant in practice (Cohen's d = 0.91).

Table 6
Absolute Frequency Distribution of Zusammenfassend ("summarizing")

discipline	mean	sd
Linguistics	5.60	5.80
Literary Studies	1.67	1.94

The difference between metatextual and intertextual use of *zusammenfassend* is not as big as the one in case study 1. The sample contains again more intertextual instances in literary studies than in linguistics and the difference is significant, but the effect is much smaller (Fisher's Exact test: $p = 0.03$, odds ratio = 3.43).

The comparison of modal verbs reveals a notable difference between the disciplines (Fisher's Exact test: $p = 0.017$). The most prominent difference is in the verb *sich lassen* ("can be")[16], which is about twice as frequent in linguistics as in literary studies. Additionally, the use of modal verbs differs markedly from the one of case study 1. 42.2% of all instances do not use a modal verb. The verb *sich lassen* is at the same time the most frequent verb (71 occurrences, 35.7% of all instances) followed by *können* ("can") with 19.1%. Here, the concluding function of *zusammenfassend* makes it likely to cooccur with modal verbs that focus on possibility, more precisely the possibilities opened up by the study.

---

16 To be exact, *sich lassen* is not a modal verb. Its function is described as an 'alternative to passive constructions with modal verb' (Duden, 2009, p. 549, our translation), more specific with the modal verb *können* ('can'). Because of this functional similarity to modal verbs and its relevance in the data under examination we include it in the analysis.

Table 7
Absolute Frequencies of Main Verbs Used with Zusammenfassend ("summarizing"), Relative Frequencies in Parentheses (Normalized to the Total Number of Sentences)

Verb	Translation	Linguistics	Literary Studies
festhalten	record	56 (0.35)	10 (0.24)
sagen	say	19 (0.12)	11 (0.27)
feststellen	determine	12 (0.08)	2 (0.05)
darstellen	depict	11 (0.07)	1 (0.02)
…	…	…	…
total		158 (1.00)	41 (1.00)

Table 7 lists the most frequent main verbs in metatextual sentences with *zusammenfassend*. The numbers are striking: The verb *festhalten* amounts to more than one third of all instances in linguistics and about one fourth of those in literary studies. The second verb, *sagen*, is also quite frequent, but far less than *festhalten*. 36 of the instances with *festhalten* also use the modal construction *sich lassen*, resulting in the prototypical sentence beginning in (3):

(3)   Zusammenfassend lässt sich festhalten, dass […]
         In summary, it can be said/recorded that […]

This indicates that academic language and especially linguistics employs very formulaic language for the metatextual purpose of indicating a summary (see for example Oakey, 2002). This is stressed even further by the fact that a considerable amount of sentences with *zusammenfassend* does at the same time use *im Folgenden* from case study 1 or a similar expression (e. g. *wie folgt* ("as follows"), or *folgende* ("following") in attributive position). This can be attested for 38 of all 199 metatextual sentences with *zusammenfassend*, and relativizes our initial assumption about the position of *im Folgenden* and *zusammenfassend* in the text. Ten sentences with *zusammenfassend* refer to a figure or table in the text, indicating that these often have a summarizing function. This type occurs in linguistics only, as tables and figures are rather rare in literary studies.

Conclusions

This study shows that candidates for metadiscourse can be identified automatically, but the retrieved instances should be inspected in detail. Especially the case study on *im Folgenden* showed no significant differences between the disciplines of literary studies and linguistics when considering the surface-based token frequencies only. However, the review of a sample revealed that literary scholars use the expression much more often in intertextual function than linguists, resulting in a relevant difference between the disciplines.

Generally speaking, linguists seem to use more metatext than literary scholars. This has been shown by the results of the n-gram analysis as well as the investigation

of the two examples *im Folgenden* and *zusammenfassend*. However, we need to keep in mind that the n-gram analysis can only capture patterns that are repeatedly realized in the same form. Maybe the metatextual function of orienting the reader in terms of text structure is realized in a more variable way in literary studies.

Putting this global difference aside, the review of those instances that are in fact metatextual revealed additional differences between the disciplines. In conjunction with *im Folgenden*, literary scholars use more modal verbs, especially *sollen*. This can be explained by a general tendency to hedging in the soft, interpretive disciplines. Another difference emerged in the type of main verb used in metatext. For linguistics, a significantly higher proportion of communication verbs was attested, which is in line with the other results as they can be regarded as a type of metatext as well.

We will now briefly address possible explanations for the disciplinary differences attested by our analysis. As the studies by Afros and Schryer (2009) and Haggan (2004) indicate, scholars in literary studies might have higher aesthetic demands with regard to their own texts, making the use of many metatextual comments undesirable. With regard to text comments,[17] Clyne (1987) even considers "embarassment [sic!] about this formal adherence to the conventions of an international journal". To confirm this hypothesis, it would be necessary to conduct interviews with the scholars and let them reflect on their motivations. Another reason can be that the research process in literary studies is much less analytical: The research process is less subdividable into distinct steps and at the same time less universal. This could explain why there are fewer references to such steps than in linguistics. Another explanation might be that the German academic language of linguistics is much more influenced by English academic language. As English tends to use more metadiscourse than German (Siepmann, 2006, p. 143), this can be a cause of its higher frequency in linguistics.

With regard to the n-gram analysis, we have to bear in mind that the results are highly dependent on the measures used. The current analysis is based on the log-likelihood ratio. A replication of the analysis using Welch's *t*-test instead yielded rather different results. Many metatextual items were ranked much lower. In contrast to the log-likelihood ratio, the t-test takes the n-grams' distribution across the corpus into account. This might mean that there are some linguistics texts in the corpus that make extensive use of metatext and cause the high scores of the log-likelihood ratio. However, our two case studies have shown that there is indeed a significant difference between the disciplines. For the future, we aim at a comprehensive comparison of the results yielded by the log-likelihood ratio and the *t*-test, respectively.

---

17 Clyne (1987) himself uses the term "advance organizers", defining their function very generally as "explain the path and organization of a paper" (p. 229).

With regard to teaching academic writing, we can draw the following conclusions: First, teachers in programs that combine linguistics and literary studies should be aware of the fact that their students have to adapt their writing to both disciplines simultaneously. This can mean that these students sometimes get contradicting information in their classes. This is why, second, teachers should explicitly draw their students' attention to the disciplinary border in their study program and discuss similarities and differences of the disciplines. Addressing differences in language could be fruitfully combined with a discussion of more general differences in what qualifies as knowledge and how knowledge is created in the two disciplines.

Acknowledgements

References

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

Ädel, A., & Mauranen, A. (2010). Metadiscourse: Diverse and divided perspectives. *Nordic Journal of English Studies*, *9*(2), 1–11.

Afros, E., & Schryer, C. F. (2009). Promotional (meta)discourse in research articles in language and literary studies. *English for Specific Purposes*, *28*(1), 58–68. https://doi.org/10.1016/j.esp.2008.09.001

Andresen, M., & Zinsmeister, H. (2017). Approximating style by N-gram-based Annotation. In *Proceedings of the Workshop on Stylistic Variation* (pp. 105–115). Copenhagen, Denmark: Association for Computational Linguistics.

Clyne, M. (1987). Cultural differences in the organization of academic texts. *Journal of Pragmatics*, *11*(2), 211–241. https://doi.org/10.1016/0378-2166(87)90196-2

Duden. (2009). *Die Grammatik: Unentbehrlich für richtiges Deutsch* (8th ed.). Mannheim, DE: Dudenverlag.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Fandrych, C., & Graefen, G. (2002). Text commenting devices in German and English academic articles. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, *21*(1), 17–43. https://doi.org/10.1515/mult.2002.002

Graefen, G. (2000). *"Hedging" als neue Kategorie? Ein Beitrag zur Diskussion*. Retrieved from http://www.daf.uni-muenchen.de/media/downloads/hedge.pdf

Haggan, M. (2004). Research paper titles in literature, linguistics and science: Dimensions of attraction. *Journal of Pragmatics*, *36*(2), 293–317. https://doi.org/10.1016/S0378-2166(03)00090-0

Hamp, B., & Feldweg, H. (1997). GermaNet – a lexical-semantic net for German. In *Automatic Information extraction and building of lexical semantic resources for NLP applications ACL/ EACL-97 Workshop Proceedings* (pp. 9–15). Madrid: European Commission, Directorale General XIII Luxembourg.

Harras, G., Winkler, E., Erb, S., & Proost, K. (2004). *Handbuch deutscher Kommunikationsverben: Wörterbuch*. New York, NY: De Gruyter.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Michigan, MI: The University of Michigan Press.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London, UK: Continuum.

Hyland, K. (2006). Disciplinary differences: Language variation in academic discourses. In K. Hyland & M. Bondi (Eds.), *Academic discourse across disciplines* (pp. 17–45). Bern, Switzerland: Peter Lang.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). London, UK: Pearson Prentice Hall.

Oakey, D. (2002). Formulaic language in English academic writing. A corpus-based study of the formal and functional variation of a lexical phrase in different acedemic disciplines. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 111–129). Amsterdam, NL: Benjamins.

Princeton University. (2010). *About WordNet*. Retrieved from http://wordnet.princeton.edu

Prokofieva, A., & Hirschberg, J. (2014). Hedging and speaker commitment. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland* (pp. 10–13). European Language Resources Association (ELRA).

Rayson, P. E. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison* (Doctoral dissertation). Retrieved from http://eprints.lancs.ac.uk/12287/

Siepmann, D. (2006). Academic writing and culture: An overview of differences between English, French and German. *Meta: Journal Des Traducteurs/Meta: Translators' Journal*, *51*(1), 131–150.

Thompson, G., & Yiyun, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, *12*(4), 365–382. https://doi.org/10.1093/applin/12.4.365