# Approximating Style by N-gram-based Annotation

**Melanie Andresen** and **Heike Zinsmeister**
Universität Hamburg
Institute for German Language and Literature
Germany
{melanie.andresen, heike.zinsmeister}@uni-hamburg.de

## Abstract

The concept of style is much debated in theoretical as well as empirical terms. From an empirical perspective, the key question is how to operationalize style and thus make it accessible for annotation and quantification. In authorship attribution, many different approaches have successfully resolved this issue at the cost of linguistic interpretability: The resulting algorithms may be able to distinguish one language variety from the other, but do not give us much information on their distinctive linguistic properties. We approach the issue of interpreting stylistic features by extracting linear and syntactic n-grams that are distinctive for a language variety. We present a study that exemplifies this process by a comparison of the German academic languages of linguistics and literary studies. Overall, our findings show that distinctive n-grams can be related to linguistic categories. The results suggest that the style of German literary studies is characterized by nominal structures and the style of linguistics by verbal ones.

## 1 Introduction

The concept of style is hotly debated in theoretical as well as empirical terms. From an empirical perspective, the key question is how to operationalize style and thus make it accessible for annotation and quantification. Many recent definitions of style focus on this aspect, resulting in very general definitions:

> Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively. (Herrmann et al., 2015)

This is a good starting point and for many studies focusing on applications such as authorship attribution or author profiling, this concept of style is perfectly sufficient. However, when the aim of investigation is interpretation rather than application, these 'formal features' need to meet additional requirements.

Most importantly, the features need to be interpretable by human readers, which is not strictly true for features like character-n-grams. Also token-based n-grams can be difficult to interpret, as they do not necessarily correspond to an actual phrase. To give a meaningful description of a language variety's style, we need to map the features to linguistic categories and, if possible, also offer independent, non-linguistic explanations. For the former purpose, we suggest an annotation task with multiple annotators that ensures a certain degree of intersubjectivity.

In the present study, this process is exemplified by a comparison of the German academic languages of linguistics and literary studies. It is part of a bigger research project that aims at describing the stylistic differences between the two disciplines. We consider this research question relevant because the two disciplines are often subsumed under one study program (e. g. *German Studies*). While this suggests a very close relationship between linguistics and literary studies, they differ in many respects.

Our analysis is based on features that are not initially linguistically motivated, but widely used: n-grams based on tokens and part-of-speech (pos) annotation. We complement them by more linguistically informed syntactic n-grams (Sidorov et al., 2012; Goldberg and Orwant, 2013). The core of our study is the following annotation experiment: After determining distinctive n-grams automatically based on frequencies, we give the most distinctive 260 token n-grams and 160 pos


*Proceedings of the Workshop on Stylistic Variation*, pages 105–115

n-grams to three annotators. They annotated whether they found the n-grams to be interpretable and, if yes, what kind of linguistic category they could derive from the n-grams.

The paper is structured as follows: Section 2 gives an overview of work in computational stylistics relevant to our study. Section 3 gives a short overview of linguistic as well as non-linguistic properties of linguistics and literary studies to which we will relate our results. We present the study's setup in section 4 by describing our data and how n-grams were generated (section 4.1) and ranked (section 4.2). Section 4.3 gives a detailed account of the annotation scheme and process. In section 5, we present the results of the annotation experiment and relate them to non-linguistic properties of the two disciplines. Section 6 discusses our study's implications.

## 2 Related Work

In this section we give an overview of studies in computational stylistics, focusing on those interested in linguistically interpretable features.

Boukhaled et al. (2015) differentiate between two methodological types of computational stylistics: 1) the *classification approach* that uses linguistic features to confirm or question a grouping of texts based on non-linguistic features, e. g. author or genre, and 2) the *hermeneutic approach*[1] identifying relevant linguistic features that serve as a data-driven starting point for human interpretation.

Most work has been done adopting the first approach, dominated by studies on the task of authorship attribution as described in the survey by Stamatatos (2009). The huge variety of features presented here refers to all kinds of language aspects that are meaningful to a greater or lesser extent, seen from a linguistic point of view. The use of a character-based data compression model is an extreme case of a linguistically uninformative method. Especially syntactic features, on the other hand, potentially contain valuable stylistic information. Hirst and Feiguina (2007) is an example of such a study that is based on bigrams of syntactic labels.

Among the linguistically motivated features used in authorship attribution, syntactic n-grams

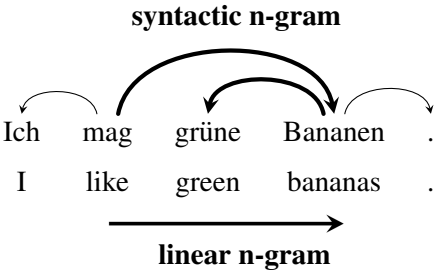


Figure 1: Example of linear and syntactic n-grams: This sentence includes the linear trigram *mag grüne Bananen* and the syntactic trigram *mag>Bananen>grüne*.

are the most promising for our research. Sidorov et al. (2012) suggest a simple concept of syntactic n-grams: Instead of linearly following the text surface as regular n-grams do, syntactic n-grams follow the dependency path in the sentence from head to dependent. Figure 1 shows an example of a linear vs. a syntactic n-gram, spanning the same set of tokens. In contrast to linear n-grams, syntactic n-grams encode syntactically meaningful relations in the sentence. Sidorov et al. (2012) achieve good results in a (non-competitive) authorship attribution task with a model based on syntactic n-grams. Goldberg and Orwant (2013) and Sidorov (2013) augment the concept to n-ary branching subtrees.

The hermeneutic approach is much less prominent than the classification approach and it is dominated by the stylistic investigation of literary works and academic language.

The features used here are primarily token-derived and lexical in nature. A widespread use of this type of analysis working with sequences of words followed upon Biber et al. (1999)'s definition of 'lexical bundles'[2]. This was mainly (but not only) applied to the study of academic language (e. g. Biber et al. (2004); Hyland (2008); Chen and Baker (2010)). Durrant (2015) analyses academic writing by students. By looking at token 4-grams he creates a disciplinary cluster of student writers. Additionally, Durrant interprets the instances found by grouping them into functional categories based on Hyland (2008).

The second field where this type of analysis has proved productive is literary stylistics. Ramsay (2007) bases his analysis of Virginia Woolf on the character-specific frequency of single words.

[1] This approach relates to hermeneutics, the distinctive methodology of interpretation in the humanities, cf. Mantzavinos (2016).

[2] We will not adopt this terminology as we see in section 5 that not all phenomena discovered by this method are lexical in nature.

Mahlberg (2007) looks at frequent token n-grams (using the term 'clusters') that function as a type of signature of characters in Charles Dickens' *Bleak House*. Mahlberg (2013) discusses this in more detail and gives a more comprehensive account of Dickens' fiction. She also gives an overview of the varying terminology (e. g. n-grams, clusters, lexical bundles) and different attempts of using these features for stylistics (Mahlberg, 2013, 48-51).

Far fewer studies use more linguistically enriched features and annotations. Boukhaled et al. (2015) include pos annotations in their investigation of classic French novels. Their features are sequences of pos tags that allow for gaps (so-called skipgrams, Guthrie et al. (2006)). Scharloth et al. (2012) use a similar approach that additionally includes combinations of token, lemma and part of speech to compare the style of two social environments in the late sixties in Germany and successfully relate the resulting linguistic features to social features of these two groups.

We consider our study as following the hermeneutic approach. In contrast to most studies, we include the token and pos level as well as syntactic annotation following Sidorov et al. (2012)'s concept of syntactic n-grams. Additionally, we systematically assess the interpretability of n-gram-based features. For measuring the reliability of the interpretations (Krippendorff, 2013, 267-270), we base this judgment on more than one person and give the task to three annotators, as described in section 4.3.

## 3 Linguistics and literary studies: Linguistic and non-linguistic differences

In this section, we will briefly describe established linguistic and non-linguistic differences between the two disciplines under investigation. We will refer back to these in the interpretation of our own results in section 5.

Academic disciplines are commonly subdivided into hard and soft sciences, which is regarded as a continuum (Biglan, 1973; Hyland, 2004). While linguistics as well as literary studies can clearly be considered disciplines of the soft sciences, most subdisciplines of linguistics tend more to the hard sciences than literary studies does.

Many differences between linguistics and literary studies therefore correspond to the differences between soft and hard sciences, just on a smaller scale. The soft sciences are characterized as being more interpretative, work hermeneutically, show several subjective perspectives and feature plurality of possible objects of study and methods. The hard sciences on the other hand are more analytical, work empirically, have a high agreement on object of study and methods and rely on quantification (e. g. Biglan (1973); Durrant (2015)).

More specifically referring to the two disciplines under examination, Gardt (2007) describes literary studies as focusing on the exemplary analysis of individual objects of study (typically texts) and linguistics as focusing rather on patterns and generalizations. We will come back to these features in the interpretation of linguistic features in section 5.

These non-linguistic features naturally lead to stylistic differences between disciplines, which have been extensively researched so that our overview has to remain illustrative. For instance, Hyland (2004) looks at disciplinary differences along the hard sciences vs. soft sciences continuum. He describes, among other results, that the disciplines vary in their citation practices: The soft fields use more citations than the hard fields and use different types of reporting verbs (Hyland, 2004, 24-29). An analysis of evaluation practices in reviews shows that while the hard fields use more praise, the soft fields use more criticism (Hyland, 2004, 49).

Biber and Gray (2016) investigate academic English in contrast to other registers and with regard to disciplinary differences. They make a distinction between phrasal (e. g. complex noun phrases) and clausal (e. g. subordination) complexity and find that the natural sciences rely more heavily on the former while the soft sciences prefer the latter.

Afros and Schryer (2009) investigate promotional metadiscourse in linguistics and literary studies and find that the style of literary studies sometimes resembles literary texts and addresses aesthetic values of the research community.

When referring to these previously found differences, we have to bear in mind that almost all studies are based on the English language. While many aspects can be expected to be cross-linguistically valid, we know that different (academic) languages have different properties. For instance, Siepmann (2006) gives a summarizing account of differences between the academic writing of English, French and German.

## 4 Study

We proceed by presenting our data and the way we generated n-grams in section 4.1, our ranking procedure in section 4.2 and the annotation scheme and setup in section 4.3.

### 4.1 Data and n-gram generation

The present study is based on a corpus of 60 PhD theses. The choice of this text type was motivated by the fact that it serves as a 'gateway genre' (Demarest and Sugimoto, 2014, 3), granting access to the academic world, and is therefore expected to highly conform to the disciplinary norms. Additionally, it is a text type that has about the same status in all disciplines. However, we have to be careful about generalizing the results to academic language in general. We created two subcorpora:

- **subcorpus of linguistics**: 30 PhD theses comprising 1,427,758 tokens,
- **subcorpus of literary studies**: 30 PhD theses comprising 2,151,679 tokens.

Sections that do not belong to the register under investigation or that interrupt the text were extracted semi-automatically: footnotes, citations, examples, tables, figures, title page, table of contents, reference section etc. This preprocessing followed rather simple heuristics and while the results are not perfect, they are sufficient for a quantitative analysis based on this amount of data.

We processed the data using the following tools: the system *Punkt* (Kiss and Strunk, 2006)[3] for tokenization and an off-the-shelf version of MATE dependency parser (Bohnet, 2010) trained on the TIGER Corpus (Seeker and Kuhn, 2012) for lemma, pos and dependency annotation. We evaluated the parser's annotations against a gold standard consensually created by two annotators for a sample of 22 sentences (600 tokens). Given that it is applied to out-of-domain data, the parser performance is good (UAS: 0.95, LAS: 0.93).

We extracted the following data sets from the resulting corpus:

- **linear n-grams** of sizes 2-5 using tokens and pos tags, respectively,
- **syntactic n-grams** of sizes 2-5 using tokens and pos tags, respectively, generated by taking every word of the sentence as a starting point and following the dependency path backwards by *n-1* steps (following the concept of Sidorov et al. (2012)).

[3] http://www.nltk.org/api/nltk.tokenize.html, 19.05.2017

### 4.2 Distinctiveness and collocational strength: n-gram ranking

For further analysis, only n-grams with a total frequency of more than 10 are included. For these n-grams we calculate their relative frequencies in all 60 texts.

In order to rank the n-grams in a way that is meaningful for later interpretation, two measures are of interest: distinctiveness and collocational strength.

First, we want to identify n-grams with a high difference in frequency between the two subcorpora and thus corresponding to major differences between the disciplines. To achieve this, we use the t-test as suggested by Paquot and Bestgen (2009) and Lijffijt et al. (2014). One of the benefits of the t-test is that it takes variation within the corpora into account. Consequently, a single text cannot dominate the overall result.

Second, we include a measure for collocational strength between the elements of the n-gram. This is necessary because the t-test results disregard the influence of significant substructures of an n-gram. Consider, for instance, that the pos tag CARD[4] is much more frequent in linguistics. Also, the bigram CARD ADJA[5] is much more frequent in linguistics. The latter observation does not necessarily mean that this combination is characteristic of linguistics but can be caused by the high difference in frequency of CARD alone.

A measure for collocational strength tells us whether the bigram is more frequent than we would expect given the corresponding unigram frequencies. Evert (2008) gives a comprehensive overview of different measures and their properties. We use the log-likelihood measure described by Dunning (1993).

While this computation is very straightforward for bigrams, the situation becomes more complicated with higher *n*. We follow the approach of Zinsmeister and Heid (2003), who break down triples of verb, adjective and noun into nested binary tuples ((adjective, noun), verb) to maintain a binary structure.

Our approach comprises the following steps:

[4] Cardinal number. The tagset used here is Schiller et al. (1999).

[5] Adjective in attributive position

1. For each n-gram that is found to be distinctive by the t-test, we generate all possible sub-n-grams contained in the n-gram. For instance, for a distinctive 4-gram, all trigrams, bigrams and unigrams contained are generated.

2. Each list of sub-n-grams is reduced to those sub-n-grams which show a significant difference between the subcorpora themselves and thus are possible candidates for causing the significance of the original n-gram alone.

3. For each of these distinctive sub-n-grams, we calculate the collocational strength between this sub-n-gram and the rest of the original n-gram.

A low log-likelihood ratio indicates that the combination of the two elements does not occur more often than expected. Consequently, it is just one of the elements that causes the distinctive effect. We exclude n-grams from the ranking that contain a combination of elements with a log-likelihood ratio below a threshold of 50.

### 4.3 Annotating n-grams

The n-gram generation and ranking can be automated to a high extent and is consequently highly replicable. For the following step of interpretation this is much less the case.

Our annotation process aims at objectifying the interpretation of n-grams as far as possible. To this end, the resulting n-grams are annotated by three annotators according to an annotation scheme that was developed in the process of annotating the data (Pustejovsky and Stubbs, 2012, 109).

The n-grams we include in the annotation tasks vary in three dimensions: They are either linear or syntactic n-grams, they are of a size between 2 and 5 and they are either based on tokens or on pos labels.

The sample of token n-grams was taken as Table 1 summarizes: For the n-gram sizes 2-5, we chose at least the 20 highest-scoring linear and syntactic n-grams. If more than 20 instances crossed the significance threshold of p=0.01 in the t-test, the sample size for that group was raised to 40 instances, giving a total sample size of 260 items.

The sample for pos n-grams comprises again the 20 highest-scoring items in our ranking of linear and syntactic n-grams for n=2-5, resulting in 160 items in total. One difference to token n-grams is the fact that pos n-grams are more abstract and consequently more difficult to interpret for human annotators. The annotators are therefore provided with five token realizations of the pos n-gram at hand for illustration. These are randomly chosen from the subcorpus of the discipline in which the n-gram is more frequent. In all annotation tasks, the annotators are not provided with any contexts the n-grams appear in as these can be quite divers and our objective was to judge the interpretability of n-grams as such.

| | | n-gram type | |
|---|---|---|---|
| | | linear | syntactic |
| n-gram size | 2 | 40 | 40 |
| | 3 | 40 | 40 |
| | 4 | 40 | 20 |
| | 5 | 20 | 20 |

Table 1: Number of instances per category in the sample of token n-grams

We present two annotation tasks: One classifying the structures in the n-gram as nominal, verbal or clausal and a second classifying them as carrying lexical or grammatical information (for token n-grams only).

First, we want to know whether the n-grams capture linguistically interpretable structures and if yes, what kind of structures we find for the two disciplines. Our first annotation scheme is roughly based on Biber et al. (2004, 381)'s 'structural types of lexical bundles' and comprises the following categories:

1. This n-gram contains a verbal structure (V).

2. This n-gram contains a nominal structure (N).

3. This n-gram contains a clausal structure (subordination) (C).

4. This n-gram contains a verbal structure that also indicates a clausal structure (subordination) (V_C).

5. This n-gram does not contain any of the above-mentioned structures (other).

By this annotation scheme we hope to achieve a high degree of abstraction that leads us to a very general characterization of the disciplinary writing styles.

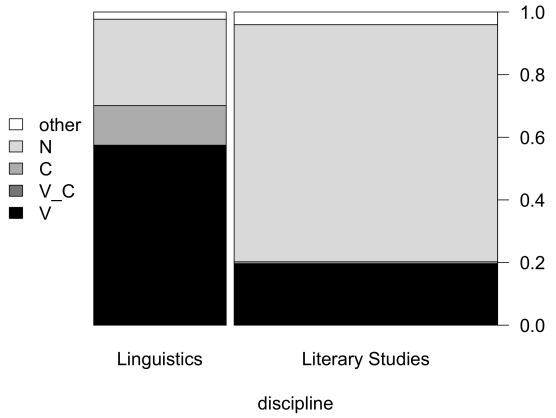


Figure 2: Annotation of structural types of token n-grams dependent on discipline, n=260 (note that the category V_C did not occur here)

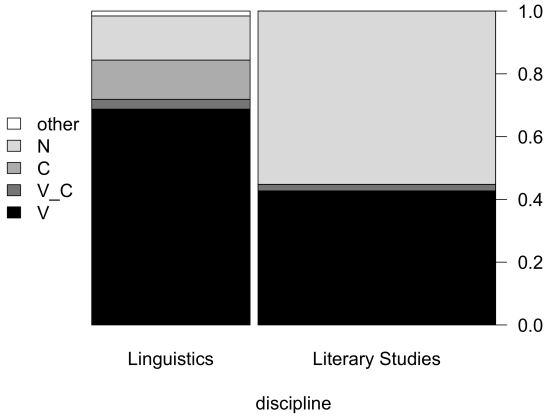


Figure 3: Annotation of structural types of pos n-grams dependent on discipline, n=160

For the sample of token n-grams, we made an additional distinction between lexical and grammatical information. This distinction allows for a general assessment of the nature of the differences between the disciplines. These two types of information contribute to style in different ways. For lexical items, it remains to be seen whether they sometimes reflect topic rather than style. The annotation follows these categories:

1. This n-gram contains a (complex) lexical unit (LEX) or overlaps with one (LEX-P).
2. This n-gram contains a grammatical structure (GRAM) or overlaps with one (GRAM-P).
3. This n-gram contains a structure that is ambiguous between lexical unit and grammatical structure (LEX-P_GRAM-P).
4. This n-gram does not contain a (complex) lexical unit or grammatical structure (NONE).

For categories 1 to 3, the annotators were asked to additionally provide the lexical unit or grammatical structure they were thinking of (e. g. relative clause). This results in very concrete phenomena and can be considered the most fine-grained annotation category. At the same time, a generalizing, quantified evaluation of the results is more difficult due to the diversity of phenomena. For the annotation of pos n-grams the differentiation between lexical units and grammatical structures does not apply, as pos tags do not directly refer to the lexical level. Therefore, the annotators are only asked for a label for the grammatical structure represented in the n-gram.

## 5 Results and Discussion

We will first present the results of the first annotation task about nominal and verbal structures (section 5.1). This is followed by the results related to lexical and grammatical phenomena in token n-grams (section 5.2), and finally by the analysis of these phenomena based on pos tags (section 5.3).

### 5.1 Nominal vs. verbal style

For the first annotation scheme differentiating nominal, verbal and clausal structures, the three annotators reached an inter-annotator agreement of 0.83 for the annotation of 260 token n-grams, measured by Fleiss' Kappa (Fleiss, 1971). Figure 2 displays the results. In the horizontal dimension we can see the two disciplines. The bars' widths show how many of the distinctive n-grams are more frequent in linguistics and literary studies, respectively. We can see that about two thirds of the n-grams in the sample are more frequent in literary studies than in linguistics. In the vertical dimension, the proportion of the annotation categories is displayed. The distinctive n-grams for the style of literary studies are dominated by nominal structures (in light gray) while verbal structures (in black) are more characteristic of linguistics. The data reveal a significant difference between the disciplines (Fisher's test, $p<0.001$).

For the annotation of pos n-grams, the annotators reached a slightly lower inter-annotator agreement of 0.69. This could be expected as pos n-

grams require more interpretation. When comparing the disciplines, we get a result similar to the token level: In Figure 3 we can see the distribution of nominal, verbal and clausal structures in pos n-grams across the disciplines. Even though the difference is less pronounced than in Figure 2, the difference between the disciplines is also highly significant (Fisher's test, $p<0.001$).

In the pos n-grams of both disciplines, verbal structures account for a higher proportion than on the token level. This shift emerges as many token instances belong to the same pos pattern, and are mapped to only one pos instance when abstracting from token to part of speech.

To summarize, we found that verbal structures are more characteristic of linguistics and nominal structures of literary studies. Assuming that our nominal structures correspond to Biber and Gray (2016)'s phrasal complexity, this result is in opposition to their observation that the hard sciences rely more on phrasal complexity than the soft sciences. We surmise that this might be due to the fact that the latter study is based on English data only. German literary studies is firmly rooted in the German academic tradition, which might result in this deviation from the English-based expectations.

Furthermore, we can see that among the sample of most distinctive structures in both figures, about two thirds are more frequent in literary studies than in linguistics. The interpretation of this fact is not straightforward and requires a careful review of the underlying patterns (e. g. their absolute frequencies and textual functions) that is beyond the scope of the current paper.

When interpreting these frequencies, we have to keep in mind that (slightly less than) half of the structures under investigation are syntactic n-grams. The dependency path through a sentence always starts with a finite verb and is relatively short in total. Consequently, most of the larger syntactic n-grams include the finite verb at the root, leading to the classification of the structure as verbal. Consequently, verbal structures are much more frequent among syntactic than linear n-grams (Fisher's test, $p<0.001$). However, this applies to both disciplines and token as well as pos n-grams in the same way. For a more comprehensive comparison of linear and syntactic n-grams, see Andresen and Zinsmeister (2017).

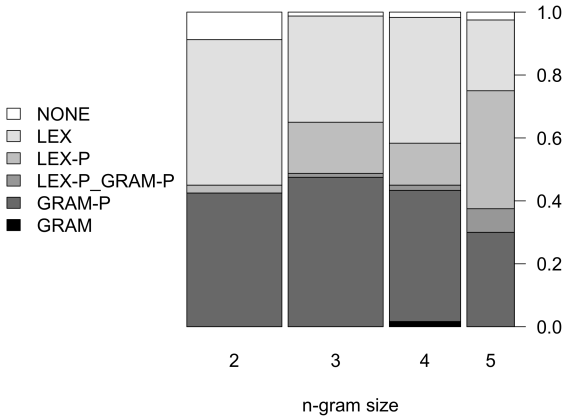


Figure 4: Annotation of information in token n-grams dependent on n-gram size, n=260

### 5.2 Lexical and grammatical structures in token features

The application of the second annotation scheme, labeling structures as being mainly characterized by grammatical or lexical properties, was more controversial. The inter-annotator agreement is 0.48 and shows that the data is rather ambiguous in terms of the annotated categories. At the same time it indicates the limits of n-gram interpretability: n-grams can invite multiple interpretations that have to be verified carefully. In these annotations, there initially were 20 instances where all three annotators chose different categories. These instances were discussed by two annotators who then agreed on one category. The results presented in the following are based on a majority vote.

In Figure 4, we present the results grouped by n-gram size. The bars' widths reflect the subsample sizes presented in Table 1. For the lexical categories, we can see that with increasing n-gram size the label LEX (in the lightest gray) tends to decrease while LEX-P (directly below) is increasing. This is understandable as LEX-P also covers structures that comprise more than the lexical item itself. With an increase in size, we are more likely to include more than the lexical unit in the n-gram. Also, the proportion of grammatical structures (darkest gray and black) drops slightly for larger n-grams. Ususally grammatical structures are signaled by only few items on the language surface, such as a comma and a subordinating conjunction for an embedded clause, whereas lexical units tend to extend over many words. The category NONE (in white) is most frequent among n-grams of size 2, indicating that this size is too

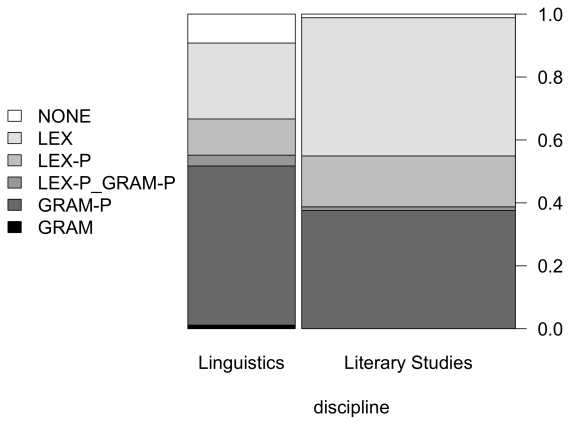


Figure 5: Annotation of information in token n-grams dependent on discipline, n=260

small to fully capture many phenomena.

Figure 5 shows the distribution across the two disciplines. We can see that, even when only taking the rather coarse-grained annotation labels (LEX, GRAM etc.) into account, we find significant differences between the disciplines (Fisher's test, $p<0.001$). Generally speaking, there are more grammatical patterns distinctive for linguistics and more lexical patterns distinctive for literary studies. When assessing this difference, we have to keep in mind that many of the linguistic phenomena are binary in nature, with one of the variants being more easily detectable by an n-gram analysis. For instance, the grammatical phenomenon 'passive voice' is more frequent in linguistics. The logical consequence is that active voice is more frequent in literary studies. However, only the high frequency of passive voice is visible in the data, as it is realized by a rather stable pattern of auxiliary verbs. This problem of detectability is especially pervasive for grammatical phenomena as they often require the realization of one of a set of options.

In addition to assigning these categories, the annotators provided the lexical or grammatical structure they derived from the n-gram. Here, the annotation is increasingly interpretative. At the same time, clearer differences between the disciplines emerge.

Among the lexical patterns we found to be more frequent in linguistics "in der Regel" ('as a rule, usually') is very prominent. This corresponds to the initial assumption that in linguistics, generalization plays a bigger role than in literary studies. Patterns like "können zurückgeführt werden auf" ('can be traced back to') show an attempt to give causal explanations. Other words like "Analyse" ('analysis') and "Auswahl" ('selection') mirror the empirical methodology of the discipline. For literary studies, on the other hand, we find many items referring to the temporal dimension: "in dem Moment" ('at that moment'), "in einer Zeit" ('at a time'), "das Ende" ('the end'), "in der ersten Hälfte des" ('in the first half of the'). This characterizes the discipline as being more narrative when referring to the (e. g. temporal) dimensions of the literary object.

Among the grammatical structures literary studies shows a higher frequency of personal pronouns, which is also related to narrative structures and individual objects of study. However, grammatical structures are by far dominated by several patterns introducing relative clauses. This indicates a rather nominal style already found in section 5.1. Interestingly, the relatively few relative clauses more frequent in linguistics all use the relative pronoun "die", which can be feminine but is more likely to be plural. This corresponds to the idea that literary studies rather deals with individuals (mostly male individuals, as the frequencies show) while linguistics deals with groups of phenomena. Other grammatical structures characteristic for linguistics are passive constructions and modal verbs as well as generally more indications of sub- and coordination (structures with "dass", 'that' and "und", 'and').

### 5.3 Lexical and grammatical structures in pos features

For the pos n-grams, the annotation of lexical vs. grammatical phenomena is less meaningful. But again, the annotators were asked to name or describe the linguistic phenomenon they see represented in the n-gram. This proved to be more difficult than for the token annotation. Often the n-grams were annotated with phenomena that could be derived from a single pos tag in the sequence, e. g. all n-grams including the pos tag PRELS[6] were annotated as relative clause, independent of the other tags in the sequence.

However, the following results can be found: Generally speaking, the phenomena mirror the differences between verbal and nominal structures found in section 5.1. More specifically, passives as well as modals and predicatives are more fre-

[6]Relative pronoun

quent in linguistics. For literary studies, complex noun (and prepositional) phrases are more common. In contrast to the results based on the token level, patterns with relative clauses occur in literary studies only. Here, the token level offers an informative differentiation. Many of these noun phrases include possessive pronouns, which are hardly found in linguistics, cf. personal pronouns discussed in the previous section.

## 6 Conclusion and future work

Our study had the aim of determining the potential of n-grams for linguistically describing style. We illustrated this by a study comparing the German academic languages of linguistics and literary studies. By means of an annotation experiment, we could show that most n-grams are interpretable in the sense that they could be related to some linguistic category. However, interpretations become more challenging with increasing n-gram length and abstractness, e. g. when interpreting parts of speech instead of tokens. Additionally, the results we found can clearly be related to non-linguistic properties of the disciplines: e. g. references to empirical methodology in linguistics, narrative structures in literary studies. Overall, the distinctive structures more frequent in literary studies are for the most part nominal. Linguistics, on the other hand, exhibits more verbal and clausal patterns.

These specific results might help scholars and especially students of the disciplines to reflect on and adapt to disciplinary writing conventions. More generally, we hope to have contributed to a better understanding of how n-gram analysis can add to the linguistic description of style. Last but not least, n-grams can serve as a starting point for subsequent in-depth analyses of language and style.

In the future, we intend to refine our method of dealing with the influence of significant substructures. Between some parts of speech there is a general collocation tendency in languages, e. g., in German a determiner and an adjective generally cooccur more often than expected by their unigram frequencies. Our current approach of using a measure of collocational strength, the log-likelihood measure, does not include this information. It requires a more detailed compositional analysis of n-grams to determine to what extend substructures can serve as a proxy of larger n-grams. In addition, it is necessary to decide whether some of the n-grams are related to topic rather than style. This depends on the specific definition of style and the analysis' objective.

In our opinion, the mathematical decisions behind the ranking of n-grams are especially important when an interpretation by humans is intended. When given an n-gram with the information that it is more frequent in one language variety than in another, humans will usually come up with some kind of interpretation of this fact. If the n-gram's rank is more of a mathematical artifact, this can lead to a highly skewed interpretation of the data.

## Acknowledgments

We would like to thank Sarah Jablotschkin for contributing to the manual n-gram annotation, Piklu Gupta for improving our English and the anonymous reviewers for their extensive and detailed comments on our submission. All remaining errors are our own.

## References

Elena Afros and Catherine F. Schryer. 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.

Melanie Andresen and Heike Zinsmeister. 2017. The Benefit of Syntactic vs. Linear N-grams for Linguistic Description. In *Proceedings of the 4th International Conference on Dependency Linguistics (Depling 2017)*. Pisa, Italy.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics* 25(3):371–405.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English language. Cambridge University Press, Cambridge.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow.

Anthony Biglan. 1973. The characteristics of subject matter in different academic areas. *Journal of Applied Psychology* 57(3):195–203. https://doi.org/10.1037/h0034701.

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference*

*on Computational Linguistics (COLING 2010)*. Beijing, China.

Mohamed-Amine Boukhaled, Francesca Frontini, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2015. Computational study of stylistics: A clustering-based interestingness measure for extracting relevant syntactic patterns. *International Journal of Computational Linguistics and Applications* 6(1):45–62.

Yu-Hua Chen and Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14(2):30–49.

Bradford Demarest and Cassidy R. Sugimoto. 2014. Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology* pages 1–14. https://doi.org/10.1002/asi.23271.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1):61–74.

Philip Durrant. 2015. Lexical Bundles and Disciplinary Variation in University Students' Writing: Mapping the Territories. *Applied Linguistics* pages 1–30. https://doi.org/10.1093/applin/amv011.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, De Gruyter, Berlin, Boston, volume 2 of *Handbooks of Linguistics and Communication Science*, pages 1212–1248.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382. https://doi.org/10.1037/h0031619.

Andreas Gardt. 2007. Linguistisches Interpretieren. Konstruktivistische Theorie und realistische Praxis. In Fritz Hermanns and Werner Holly, editors, *Linguistische Hermeneutik: Theorie und Praxis des Verstehens und Interpretierens*, Niemeyer, Tübingen, number 272 in Reihe Germanistische Linguistik, pages 241–261.

Yoav Goldberg and Jon Orwant. 2013. A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 241–247.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*. pages 1–4.

Berenike Herrmann, Karina van Dalen-Oskam, and Christof Schöch. 2015. Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory* 9(1):25–52. https://doi.org/10.1515/jlt-2015-0003.

Graeme Hirst and Ol'ga Feiguina. 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing* 22(4):405–417. https://doi.org/10.1093/llc/fqm023.

Ken Hyland. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. The University of Michigan Press, Michigan.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27(1):4–21. https://doi.org/10.1016/j.esp.2007.06.001.

Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics* 32(4):485–525. https://doi.org/10.1162/coli.2006.32.4.485.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, third edition.

Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* pages 1–24. https://doi.org/10.1093/llc/fqu064.

Michaela Mahlberg. 2007. Corpus stylistics: Bridging the gap between linguistic and literary studies. In Michael Hoey, Michaela Mahlberg, Michael Stubbs, and Wolfgang Teubert, editors, *Text, Discourse and Corpora. Theory and Analysis*, Continuum, London, Studies in corpus and discourse, pages 217–246.

Michaela Mahlberg. 2013. *Corpus Stylistics and Dickens's Fiction*. Number 14 in Routledge advances in corpus linguistics. Routledge, New York.

Chrysostomos Mantzavinos. 2016. Hermeneutics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Winter 2016 edition.

Magali Paquot and Yves Bestgen. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Andreas H. Jucker, Daniel Schreier, and Marianne Hundt, editors, *Corpora: Pragmatics and Discourse*, Brill, pages 247–269. https://doi.org/10.1163/9789042029101_014.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., Sebastopol, CA.

Stephen Ramsay. 2007. Algorithmic Criticism. In *A Companion to Digital Literary Studies*, Blackwell Publishing, Malden, MA, number 50 in Blackwell companions to literature and culture, pages 477–491.

Joachim Scharloth, Noah Bubenhofer, and Klaus Rothenhäusler. 2012. Andersschreiben aus korpuslinguistischer Perspektive: Datengeleitete Zugänge zum Stil. In Britt-Marie Schuster and Doris Tophinke, editors, *Andersschreiben. Formen, Funktionen, Traditionen*, Erich Schmidt Verlag, Berlin, number 236 in Philologische Studien und Quellen, pages 157–178.

Anne Schiller, Simone Teufel, Christine Thielen, and Christine Stöckert. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. Stuttgart, Tübingen.

Wolfgang Seeker and Jonas Kuhn. 2012. Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey, pages 3132–3139.

Grigori Sidorov. 2013. Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction. *International Journal of Computational Linguistics and Applications* 4(2):169–188.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2012. Syntactic Dependency-Based N-grams as Classification Features. In Ildar Batyrshin and Miguel González Mendoza, editors, *Advances in Computational Intelligence*, Springer, number 7630 in Lecture Notes in Computer Science, pages 1–11. https://doi.org/10.1007/978-3-642-37798-3_1.

Dirk Siepmann. 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.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3):538–556. https://doi.org/10.1002/asi.21001.

Heike Zinsmeister and Ulrich Heid. 2003. Significant triples: Adjective+ noun+ verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003), Budapest*.