The benefit of syntactic vs. linear n-grams for linguistic description

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

Abstract

Automatic dependency annotations have been used in all kinds of language applications. However, there has been much less exploitation of dependency annotations for the linguistic description of language varieties. This paper presents an attempt to employ dependency annotations for describing style. We argue that for this purpose, linear n-grams (that follow the text's surface) alone do not appropriately represent a language like German. For this claim, we present theoretically as well as empirically founded arguments. We suggest syntactic n-grams (that follow the dependency paths) as a possible solution. To demonstrate their potential, we compare the German academic languages of linguistics and literary studies using both linear and syntactic n-grams. The results show that the approach using syntactic n-grams allows for the detection of linguistically meaningful patterns that do not emerge in a linear n-gram analysis, e. g. complex verbs and light verb constructions.

1 Introduction

Linear n-grams in the sense of adjacent strings of tokens, parts of speech, etc. are a very common and successful way of modeling language in computational linguistics. However, linguistic structures do not always work in such linear ways. From a cross-linguistic perspective, some languages are less linearly organized than others. While many (though not all) syntactic structures in English can indeed be described by linear patterns, this is much less true for languages with a more flexible word order and other syntactic properties that induce long distance relations, e. g. German. Still, the linear n-gram approach is quite successful when used for applications in such languages. In the present paper our aim is a slightly different one. We want to employ n-grams not as a means for an application but for linguistic description itself. This requires the language modeling to be more linguistically adequate and interpretable and not just to be a means to an end. We consider the use of *syntactic n-grams* in addition to linear ones to be a possibility to achieve this aim.

In order to motivate our approach, we will first introduce the concept of syntactic n-grams (section 2) and present related work (section 3). Then we will investigate the descriptive benefit of syntactic n-grams by, firstly, looking at theoretical descriptions of non-linear German syntax (section 4.1), and secondly, by investigating empirical consequences of such structures by describing cross-linguistic differences in Universal Dependencies (UD) treebanks, with a special focus on the comparison of English and German (section 4.2).

In the main part of this paper we will present a study of stylistic comparison between different academic disciplines, namely between linguistics and literary studies in German (section 5). To capture these differences, we will compare the frequencies of n-grams between the two disciplines and contrast the results yielded by linear and syntactic n-grams in section 6.

Finally, we will summarize our results in section 7. The analyses show that syntactic n-grams capture relevant structures that would be missed in a purely linear approach, e. g. complex verbs and light verb constructions.

2 Syntactic n-grams

Linear n-gram analysis is an omnipresent method in computational linguistics and has proven to be an easy to implement and highly appropriate approximation of how language works in many ap-

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 4-14,
Pisa, Italy, September 18-20 2017

plications (see Jurafsky and Martin (2014, chap. 4) for an overview).

However, for the linguistic description of language this is often not satisfactory, as the underlying linguistic patterns are not always linear. One possible remedy for this issue is the approach of skip-grams (see e. g. Guthrie et al. (2006)), but they disregard linguistic structures and thus generate a lot of noise. Another approach for overcoming this problem is the use of syntactic n-grams. Instead of following the word order as it appears on the surface, they are based on dependency paths in the sentence.

A simple type of syntactic n-grams relying on unary-branching dependency structures is described by Sidorov et al. (2012):

> [...] we consider as neighbors the words (or other elements like part-of-speech tags, etc.) that follow one another in the path of the syntactic tree, and not in the text. We call such n-Grams syntactic n-Grams (sn-Grams). (Sidorov et al., 2012, 1)

A more sophisticated approach is suggested by Goldberg and Orwant (2013). Their definition augments the one by Sidorov et al. (2012) by including all kinds of n-ary branching subtrees:

> We define a syntactic-ngram to be a rooted connected dependency tree over k words, which is a subtree of a dependency tree over an entire sentence. (Goldberg and Orwant, 2013, 3)

This results in the additional inclusion of n-grams with more than one dependent per head, which is also advocated by Sidorov (2013).[1]

As a base for the more widespread use of syntactic n-grams, Goldberg and Orwant (2013) create a comprehensive database on the basis of the Google Books corpus for general use. In their representation of n-grams, they exclude functional words and include multiple layers of annotation (part of speech, dependency relation, head). In addition, they preserve the information about the word order in the text. Our analysis will be based on the simpler type of syntactic n-grams by Sidorov et al. (2012) (see section 5.2).

[1] Compare also to the concept of catenae presented in Osborne et al. (2012).

3 Related Work

In this section, we will briefly refer to other types of syntactically motivated features and applications they were used in. Then we will look at the use of n-grams and syntactic features in authorship attribution and stylistic analysis.

Dependency-based features have been used for various applications. For example, Snow et al. (2004) use dependency paths between nouns as one feature to extract lexical hypernymy relations. Padó and Lapata (2007) use similar dependency subtrees as a feature to create general semantic space models. Versley (2013) uses subgraphs to describe larger structures, in particular implicit discourse relations in texts.

Syntactic features have also been systematically compared to linguistically less informed features like linear n-grams or bag-of-words approaches. Lapesa and Evert (2017) evaluate the performance of dependency-based and simpler window-based models for computing semantic similarity and find the simpler model to be superior in most cases. Bott and Schulte im Walde (2015) present similar findings when employing syntactically informed features in the task of predicting compositionality of German particle verbs.

Sidorov et al. (2012) use syntactic n-grams in an authorship attribution task. Their syntactic n-grams include the syntactic relation labels only and achieve good results compared to linear n-grams. Stamatatos (2009) gives an overview of the use of other types of syntactic features in authorship attribution. These include for instance syntactic rewrite rules based on phrase structures and syntactic errors. In a more recent study, van Cranenburgh and Bod (2017) successfully quantify the literariness of novels by using, among others, fragments of syntactic constituency trees as features. They stress the fact that these features have the advantage of being more interpretable than others that are not syntactically motivated.

N-gram approaches have also been used for more interpretative analyses in the humanities. Biber et al. (1999) and others investigate academic language with the help of so-called 'lexical bundles'. In literary studies, Mahlberg (2013), among others, uses data-driven 'clusters' for describing the style of Charles Dickens' prose. Both approaches rely on token-based n-grams only and do not make use of syntactic annotation.

Most of the computational linguistics ap-

proaches have in common that they use syntactic n-grams or syntactic subtrees for some practical application. Even stylistic approaches of aim at classifying documents rather than describing them. On the other hand, studies in the humanities that aim at describing and interpreting language tend to use rather simple features that do not include syntactic information. By merging the means of the first with the aims of the second group, we will explore the potential syntactic n-grams hold for the linguistic description of languages.

4 Non-linear structures

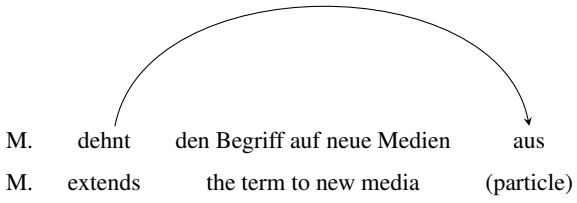
We will at first motivate the need for syntactic n-grams by considering *non-linear structures* in the sense of structures that are expressed in a discontinuous token string. This means that they cannot be captured by regular linear n-grams. In particular, we are interested in structures which occur frequently enough for us to expect them to have an impact on n-gram creation. Section 4.1 gives a theoretical foundation by introducing non-linear syntactic structures from German. Section 4.2 discusses empirical consequences of these properties with a special focus on the comparison of English and German.

4.1 Theoretical foundation

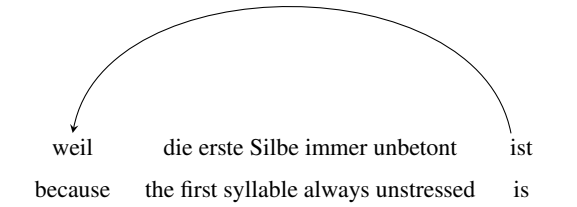
To what extent the syntactic structure of a language is linear is a question of typology and differs widely between languages. The use of n-grams for linguistic applications and analyses is a method that favors languages with dominantly *linear structures*, i. e. structures that are expressed by continuous token strings. German is one example of a language that is rich in non-linear structures.[2]

We will first focus on non-linear structures that are projective, i. e. structures that do not cause dependency paths to overlap. These are commonly discussed under the model of Topological Fields that describes German as using so-called bracketing structures: Once the first part of the bracket is realized, the reader/hearer expects the second part to occur as well (see Kübler and Zinsmeister (2015, 73) or Becker and Frank (2002) for an English description). Three types of these structures can be distinguished:

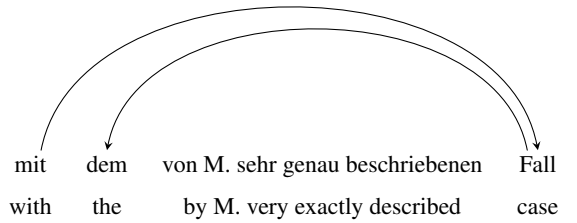
[2]The non-linear characteristics of German are most prominently described and parodied by Mark Twain (1880).



(a) Example of a finite particle verb



(b) Example of a subordinate clause with the verb in final position



(c) Example of a noun phrase

Figure 1: Examples of non-linear structures in German

Main clauses. In main clauses, several types of complex verbal structures lead to non-linearity:

- full verbs complemented by auxiliary and/or modal verbs,
- copula verbs complemented by predicatives,
- light verb constructions,
- finite particle verbs.

In all of these verb constructions, the finite part of the verb will be in second position while the other verbal elements are in final position. The number of phrases in between, in the so-called middle field, is theoretically unlimited. Figure 1a shows an example of the particle verb *ausdehnen* ('to extend') with the finite verbal part *dehnt* in second position and the separated particle *aus* in sentence-final position.

Subordinate clauses. This bracketing structure is opened by the phrase-initial subjunction and closed by the finite and non-finite verb forms that are in sentence-final position (see example in Figure 1b).

Noun phrases. Finally, German also has non-linear structures similar to English: The noun phrase is opened by a determiner (or indirectly by a preposition) and closed by the noun itself. In between, the phrase can be extended by mainly adjective phrases. Additionally, the German noun phrase can comprise structures in pre-nominal position that would be placed post-nominally in English as shown in the example in Figure 1c.

Maier et al. (2014) present additional discontinuous structures that are characterized not only by the distance between their elements, but also by non-projective dependencies, i. e. by crossing dependencies: "extraposition, a placeholder/repeated element construction, topicalization, scrambling, local movement, parentheticals, and fronting of pronouns" (Maier et al., 2014, 1). However, these structures are much rarer than the projective non-linear ones described above and are not expected to be reflected in the frequency data of the n-gram analysis.

In the light of the example of German we have seen that there are languages with many non-linear structures that do not have an equivalent in English.

4.2 Empirical consequences

In order to empirically demonstrate and quantify the degree to which languages make use of non-linear structures and describe their nature, we focus on the distance between head and dependent in dependency annotated data in terms of surface tokens. For a cross-linguistic comparison we use the training data of Universal Dependencies 2.0 (Nivre et al., 2017). Table 1 shows the median and mean distance and standard deviation between head and dependent in several languages[3]. Punctuation and the root were excluded from the calculation. A distance of 0 means that head and dependent are directly adjacent.

First, we can see that even in English – the language most applications were primarily developed for – head and dependent are often non-adjacent. On average, 1.77 words are in between head and dependent. Second, it becomes clear that the distances vary greatly also within languages, with Arabic and Persian having a very high standard deviation of 6.78 and 5.09, respectively. Even though one should bear in mind that some differences might be due to the language-specific implementations of the Universal Dependencies, we can assume that there are in fact differences between the languages.

[3]The sample of languages is only a subset of more than 50 languages available in UD.

	median	mean	sd
Persian	0	2.62	5.09
German	1	**2.28**	**4.02**
Arabic	0	2.14	6.78
Dutch	1	2.06	3.54
English	1	**1.77**	**3.32**
French	1	1.71	3.92
Russian	1	1.70	3.51
Swedish	1	1.70	4.79
Czech	1	1.70	3.24
Turkish	0	1.69	3.46
Italian	1	1.68	4.12

Table 1: Distance between head and dependent in UD treebanks (without punctuation and root)

Figure 2 exemplary shows the distribution of the distances of the part of speech `sconj` (= subordinating conjunctions) to its head in more detail. Here, the differences between the languages are more pronounced than with other parts of speech. Turkish and Arabic do not have this part of speech. With a median of six (marked by the black line inside the box), German features the highest distance, followed by Persian, another verb-final language, and Dutch, which is similar to German in this respect.

In the remainder of this paper we will focus on German as an example of a language in which the average distance is significantly higher than in English[4] and more variable.

[4]t = 42.998, df = 386460, p-value $<$ 2.2e-16

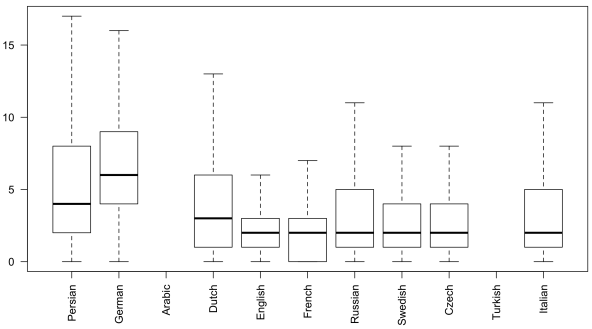


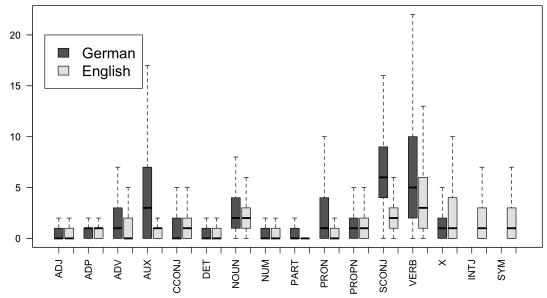
Figure 2: Distance to head of words with the part of speech `sconj` in all languages

Which syntactic structures are related to these differences? Figure 3a shows boxplots of the distance distributions between heads and dependents in English and German grouped by the part of speech of the dependent. The most obvious differences relate to the theoretical findings in section 4.1. German verbs and auxiliary verbs show much larger distances from their heads than their English counterparts, as can be expected because of the German bracketing structure. Subordinating conjunctions (SCONJ) show the largest difference in the two languages with the interquartile ranges of their distributions not even overlapping. This reflects the German brackets in subordinate clauses, which result in a large distance between the subjunction and the finite verb of the subordinate clause.

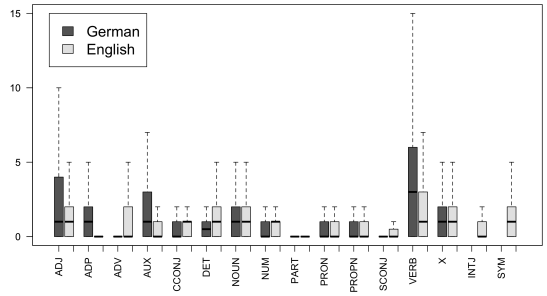
Another clear difference is in pronouns, which are positioned early in the sentence in German (before or immediately after the finite verb, the so-called 'Wackernagel position', Cardinaletti and Roberts (2002, 133)), while their head (usually the main verb) can be sentence-final. Also nouns and adverbs tend to be slightly further away from their head in German than in English. This can probably be attributed to the generally freer word order in German (empirically shown in Futrell et al. (2015)).

Figure 3b shows the same relation from the other direction: The same distances grouped by the part of speech of the head. Again, German verbs and auxiliary verbs prove to be further away from their dependent than the English ones. Adjectives are another notable case. According to the Universal Dependencies' guidelines, adjectives are considered the root of the sentence when they occur in predicative structures (e. g. *This is very easy.*). The copula is one of its dependents, which can again be far away from the predicative adjective in German.

Finally, all of the phenomena described above are also reflected when looking at the distances grouped by syntactic relation: Many of the high-distance relations in German refer to different types of clauses (`acl`, `advcl`, `ccomp`, `csubj`) and complex verbs (`aux`, `compound:prt` (particle verbs)), especially in combination with passives (`csubj:pass`, `nsubj:pass`, `aux:pass`). `mark` is the relation between subjunctions and finite verbs in subordinate clauses. It also features a clear difference



(a) Distance by pos of dependent



(b) Distance by pos of head

Figure 3: Distance between head and dependent in UD treebanks (without outliers)

in distance between the two languages.

This section has shown that the non-linear structures described in section 4.1 have an impact on the distance between head and dependent. It could be demonstrated that these distances are much larger in German dependency structures than in English ones. This means that the modeling of German using only linear n-grams is not fully adequate for its linguistic description. In the next section, we will compare the contribution of syntactic and linear n-grams to a stylistic analysis of German academic language.

5 Study: Disciplinary differences in academic writing style

The following study is part of a larger project on stylistic analysis of German academic texts written in the disciplines of linguistics and literary studies, respectively. This field of research is motivated by the fact that these two disciplines are often combined in one common study program such as German Studies or German Language and Literature. While this suggests that the disciplines are very closely related, writing styles differ widely (see e. g. Afros and Schryer (2009)). We present an attempt to capture these differences by an n-

gram analysis based on linear and syntactic n-grams.

5.1 Data and preprocessing

The study is based on a corpus of 60 German PhD theses, 30 for each of the two disciplines linguistics and literary studies.

All texts were accessible as PDF files. In a first preprocessing step, we converted them to HTML to use the HTML markup for semi-automatically deleting irrelevant parts of the text. In particular, we deleted parts that do not belong to the targeted varieties and often interrupt the running text: tables and figures, footnotes, citations and examples. We also removed all text sequences in parentheses as most of them comprise references, especially in linguistics. Additionally, we excluded sentences with more than 40% of the words in quotes, assuming that they do not represent the target variety either. Other elements we had to exclude manually, e. g. title page, table of contents, and list of references. The resulting plain text version has a total count of 3,579,437 tokens.

We tokenized the texts using the system *Punkt* (Kiss and Strunk, 2006)[5] and annotated the sentences with an off-the-shelf version of MATE dependency parser (Bohnet, 2010) trained on the TIGER Corpus (Seeker and Kuhn, 2012). Note that in contrast to the previous chapter, we decided against using Universal Dependencies. As this part of the study deals with German only, we consider the tag set developed specifically for German more appropriate. For the purpose of evaluation, two annotators consensually created a gold standard for a random sample of 22 sentences (600 tokens) against which we compared the parser's output. The parser performance is good (UAS: 0.95, LAS: 0.93), especially given that it is applied to out-of-domain data.

5.2 N-gram generation

We extracted several data sets from the preprocessed corpus:

- **linear n-grams** of sizes 2-5 using tokens, lemmas, pos-tags and dependency relation labels,
- **syntactic n-grams** of sizes 2-5 using tokens, lemmas, pos-tags and dependency relation labels, generated by taking every word of the sentence as a starting point and following the dependency path backwards by *n* steps.

The data set for the present analysis is not sufficiently large to allow for a representation of syntactic n-grams that includes as many annotations as Goldberg and Orwant (2013) used. To avoid issues of data sparsity, only one level of information at a time is included, e. g. token OR lemma OR part of speech OR the dependency relation label. In line with Sidorov et al. (2012), the analysis is restricted to unary syntactic n-grams following only one branch in the syntactic tree.

We exclude n-grams with a total frequency of less than 10 from further analysis. For all the resulting n-grams we calculate relative frequencies in all 60 texts. The difference in frequency between the two subcorpora is assessed based on the t-test as suggested by Paquot and Bestgen (2009) and Lijffijt et al. (2014). Each data set is then ranked according to the t-test's p-values.

6 Results and Discussion

In the analysis, we inspect the degree of overlap between linear and syntactic n-grams in order to assess whether the two types truly give us complementary information (section 6.1). However, our main question is whether both types contribute meaningfully to a linguistic description of the disciplinary differences between linguistics and literary studies. Section 6.2 therefore gives an exemplary interpretation of the most distinctive linear and syntactic 4-grams. On that basis, the final section 6.3 presents an attempt to quantify linguistic interpretability.

6.1 Overlap between linear and syntactic n-grams

In order to first get a general idea of the added value of syntactic n-grams independent of our research question about disciplinary differences, we quantify the overlap between linear and syntactic n-grams. To this end we investigated to what degree the syntactic n-grams correspond to linear n-grams.

We calculated for all four levels (token, lemma, part of speech and dependency relation), to what extent the 200 highest-scoring syntactic n-grams correspond to linear n-grams.[6] For each of the

[5] http://www.nltk.org/api/nltk.tokenize.html, 23.07.2017

[6] With increasing n, the number of n-grams passing the frequency threshold of 10 decreases quickly. Therefore, the number for syntactic token 5-grams is only based on 37 items

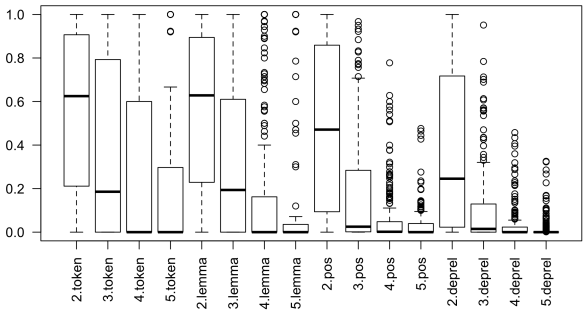


Figure 4: Proportion of syntactic n-grams that correspond to a linear n-gram (by n-gram size and level of annotation)

200 syntactic n-gram types, we checked all corresponding token instances for linearity (score 1) or non-linearity (score 0) and calculated the mean for each type. The resulting value gives us information about the overlap of linear and syntactic n-grams: A score of 1 means that all token instances of the syntactic n-gram are also linear n-grams. A score of 0 means that none of the token instances of the syntactic n-gram correspond to linear n-grams.

Figure 4 shows the resulting distribution of overlap by n-gram size and level of annotation. The proportion of linear n-grams is low, with a mean between 0.36 and 0.57 already for bigrams, depending on the level of annotation. As expected, the proportion of linear n-grams decreases as *n* increases. With every additional transition from one word to the next, the probability of at least one deviation from the linear order rises.

Additionally, there is a tendency of decreasing linearity with increasing abstractness from token to lemma to part of speech and dependency relation. One particular combination of tokens can be exclusively realized linearly but a lemma comprises several different token combinations, which will not all be realized linearly. With increasing abstractness, more heterogeneous cases are subsumed under one label, making purely linear instances less and less likely.

However, it has to be borne in mind that syntactic n-grams with more than one branch were not included. These might correspond to linear n-grams to a higher degree, resulting in a higher overlap between the two types of n-grams. In the present analysis, linear n-grams cover some structures that correspond to syntactic units, but are not captured by our narrow approach to syntactic n-grams. Consequently, the widening of our realization of syntactic n-grams is advisable in future work.

6.2 Interpretation of linear and syntactic 4-grams

We will now focus on the possibilities of interpreting linear and syntactic n-grams in order to draw conclusions about linguistic properties of the German academic languages of linguistics and literary studies. In this section, we discuss one example in detail while the next section will present possibilities of quantifying these interpretations on a larger scale. The focus will be on token n-grams as they can easily be read by humans. Especially longer part-of-speech sequences (like ART-NN-APPR-PPOSAT-NN[7]) are quite abstract and require a person with experience with the tag set and possibly a set of example instances (see Andresen and Zinsmeister (2017) for an attempt to include these).

Table 2a and Table 2b show the 15 highest-scoring 4-grams for the linear and the syntactic data set, respectively. These are the n-grams with the highest difference in frequency when comparing the disciplines. In addition to the n-gram, an approximate translation into English is provided. Given the fragmentary nature of n-grams, these translations are sometimes based on additional assumptions about the context and do therefore only represent one of several possible meanings. The row color indicates in which discipline the n-gram is more frequent: n-grams more frequent in literary studies are colored gray, those more frequent in linguistics white.

Among the linear n-grams in Table 2a, structures following a comma dominate the ranking. This can be explained by the fact that the beginning of subordinate clauses is grammatically restricted to some specific patterns. Because of the grammatical gender in German, some structures reoccur in several similar forms. Many patterns that are significantly more frequent in literary studies indicate relative clauses (rank 3, 4, 5, 7, 8 and 12). For linguistics this is only true for

that do not necessarily achieve low p-values in the t-test. Also, the syntactic token 4-grams and linear token 4-/5-grams are partially based on items that do not pass the level of significance (p=0.001).

[7]The tag set used here is the STTS (Schiller et al., 1999). This sequence corresponds to article – noun – preposition – possessive pronoun in attributive position – noun, e. g. *the name of his mother*.

rank	linear n-gram	literal translation	comment
1	, die bei der	, that.3SG.F/3PL at the	
2	davon aus , dass	expect that	fragment of: expect that the
3	, das in der	, that.3SG.N in the	
4	, in der er	, in which he	
5	, der sich von	, that.3SG.M it.REFL of	
6	aus , dass die	out, that the.3SG.F/3PL	fragment of: expect that the
7	, in dem sie	, in that3SG.M/N she/they	
8	, in dem sich	, in that3SG.M/N it.REFL	
9	bei der Auswahl der	in the selection of	
10	, ob es sich	, whether it it.REFL	
11	, bei denen sich	, at which it.REFL	
12	, der sich in	, that.3SG.M it.REFL in	
13	, sich in die	, it.REFL in the	
14	aus sich selbst heraus	out of it.REFL	
15	, die sich auf	, that.3SG.F/3PL it.REFL on	

(a) Linear token 4-grams

rank	syntactic n-gram	literal translation	translation
1	und>können>werden>.	and>can>be>.	and can be. (passive)
2	rückt>in>Vordergrund>den	bring>to>fore>the	bring to the fore
3	rückt>in>Nähe>die	bring>in>proximity>the	bring sth. closer to
4	ist>in>Lage>der	is>in>condition>the	is capable of
5	im>als>im>auch	in>as>in>also	in X as well as Y
6	bei>als>bei>auch	at>as>at>also	at X as well as Y
7	kann>werden>gelesen>als	can>be>read>as	can be read as
8	werden>erläutert>im >Folgenden	is>explained>in the>following	In the following, ... is explained
9	ist>in>Regel>der	ist>in>rule>the	is generally
10	war>in>Lage>der	was>in>condition>the	was capable of
11	und>kann>nicht>mehr	and>can>not>anymore	and can no longer
12	zu>Beginn>Jahrhunderts >des	at>beginning>century>the	at the beginning of the century
13	werden>vorgestellt>Im >Folgenden	is>presented>in the>following	In the following, ... is presented
14	in>Hälfte>Jahrhunderts>des	in>half>century>of the	in the ... half of the century
15	stellt>in>Mittelpunkt>den	puts>in>center>the	centers/focuses on

(b) Syntactic token 4-grams

Table 2: Highest-scoring token 4-grams for linear and syntactic n-grams (rank based on t-test; gray = n-gram is more frequent in literary studies, white = n-gram is more frequent in linguistics)

rank 1, 11 and 15. Interestingly, all of these use the pronoun *die*, which can be feminine singular, but is more likely to be plural (independent of gender). We might derive the explanatory hypothesis that literary scholars write more about individuals while linguists are rather concerned with groups of phenomena in a generic way. This is in accordance with the intuitive idea of how these disciplines work.

The results for syntactic n-grams in Table 2b are quite different. The most distinctive is a very general complex verb pattern in passive voice with the modal verb *can*, that can be combined with any main verb and is more common in linguistics. There are also some more specific complex verbs that include a main verb (rank 7, 8 and 13). Additionally, there are the light verb constructions *in den Vordergrund rücken* ('bring to the fore'), *in die Nähe rücken* ('bring sth. closer to sth. else'), *in der Lage sein* ('be able to do sth.') and *in den Mittelpunkt stellen* ('focus on sth.'). All of these structures relate to the properties of German described in section 4.1 and would not be detected in a purely linear n-gram approach. Other syntactic n-grams refer to structures that can be captured similarly by linear n-grams, e. g. the syntactic 4-gram *ist>in>Regel>der* corresponds to the linear n-gram *ist in der Regel*. This reflects the findings of section 6.1 showing overlap as well as differences between the two types of n-grams.

6.3 Quantifying linguistic interpretability

Taking these interpretations as a starting point, we made the attempt to quantify the interpretability of linear and syntactic n-grams. Thereby we hope to objectify the n-grams' potential and provide a foundation for a deepened comparison.

A sample of syntactic and linear n-grams[8] was annotated by three annotators according to the following 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 lexicon and grammar (LEX-P_GRAM-P).

[8]For the n-gram sizes 2-5, we chose the 20 highest-scoring syntactic and linear token n-grams, respectively, giving a total sample size of 160 items.

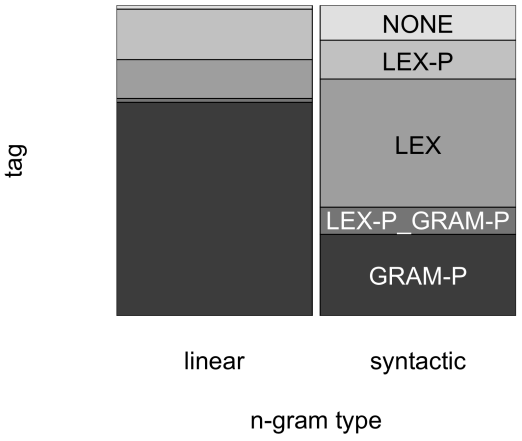


Figure 5: Annotation of information in n-grams dependent on n-gram type, n=160

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.

The annotators reached an inter-annotator-agreement of Fleiss' κ 0.55 which we consider satisfying given the natural ambiguity of the task. After discussing nine elements where no agreement was reached initially, all three annotators agreed on one category for 57% of items. For the rest at least two annotators agreed on one category. The following results are based on a majority vote.

Figure 5 shows the distribution of annotation categories for the two n-gram types. For the linear n-grams, more grammatical phenomena were found, and for syntactic n-grams, more lexical phenomena (especially complete lexical items) were found. The difference is significant with $p < 0.001$ (Fisher's Exact Test), which shows that there are many non-linear lexical items that are detected by the syntactic n-grams only. The number of non-interpretable instances is higher in syntactic n-grams (1 vs. 10 instances). These are e. g. sequences of only one word and the following punctuation or sequences related to specific properties of the annotation scheme.

Regarding the concrete structures observed, there is a clear overlap in lexical phenomena, e. g. the sequence *in der Regel* ('as a rule') is a linear as well as a syntactic n-gram. Syntactic n-grams additionally capture light verb constructions that are non-linear (see section 4.1), e. g.

den<Vordergrund<in<rückt ('bring to the fore'), which might explain the higher proportion of lexical phenomena. In grammatical structures, on the other hand, there is hardly any overlap. While most linear n-grams (35 of 55 grammatical structures in total) capture different types of relative clauses (e. g. the trigram *, die ihm*, 'that [...] him'), among the syntactic n-grams complex verb structures (11 of 19 grammatical structures in total) and phenomena of coordination (5 of 19) dominate.

Together, linear and syntactic n-grams result in an informative comparison of the two disciplines: In literary studies we find many more relative clauses and light verb constructions, while linguistics employs more complex verb forms like passive and modal verbs. A more comprehensive interpretation of these and more data with respect to the disciplinary differences is conducted in Andresen and Zinsmeister (2017).

The annotation experiment shows that linear and syntactic n-grams capture very different phenomena and can complement each other in useful ways. At this point, it is not possible to generalize these results as they need to be verified by analyzing more data of different genres (and languages).

7 Conclusion

The research presented in this paper shows that an analysis based on syntactic n-grams, understood as n-grams following the path of dependency relations in the sentence, can give linguistically meaningful insights in the properties of a language variety. We have demonstrated theoretically and empirically that there are many non-linear structures in languages like German. These are not adequately taken into consideration in a language representation based on linear n-grams only. Through the example of comparing the German academic languages of linguistics and literary studies we showed that linear and syntactic n-grams capture very different linguistic structures. In our exemplary study, especially complex verbs and light verb constructions could not be detected by the linear n-gram analysis.[9] However, the analysis of syntactic n-grams is highly dependent on the quality of the dependency annotation. Also, some structures are frequent only because of specific properties of the annotation scheme. It remains a desideratum for future research to determine the influence of the annotation scheme and the potential of Universal Dependencies to allow for a cross-linguistic comparison of this type of analysis.

For the future, it would be desirable to include syntactic n-grams that take more than one dependent per head into account. Currently, patterns such as a verb and its subject and object or a noun and two modifiers are missed by the syntactic n-grams of our study. The linear n-grams can compensate this only very partially. Also, it should be considered to systematically evaluate the potential of dependency-based annotations in comparison to other syntactic models, e. g. constituency-based models.

[9] Our aim was to increase coverage of phenomena included in the analysis. We do not to automatically distinguish between light verb constructions and free verb-noun associations.

Acknowledgments

We would like to thank Yannick Versley and Fabian Barteld for their very helpful comments on an earlier version of the paper, Sarah Jablotschkin for contributing to the manual n-gram evaluation, and Piklu Gupta for improving our English. 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, January.

Melanie Andresen and Heike Zinsmeister. 2017. Approximating Style by n-Gram-based Annotation. In *Proceedings of the Workshop on Stylistic Variation*, Copenhagen, Denmark, September.

Markus Becker and Anette Frank. 2002. A Stochastic Topological Parser of German. In *Proceedings of COLING 2002*, pages 71–77.

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

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.

Stefan Bott and Sabine Schulte im Walde. 2015. Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs. In *Proceedings of the 11th International Conference on Computational Semantics, IWCS 2015, 15-17 April, 2015, Queen Mary University of London, London, UK*, pages 34–39.

Anna Cardinaletti and Ian Roberts. 2002. Clause Structure and X-Second. In Guglielmo Cinque, editor, *Functional Structure in DP and IP: The Cartography of Syntactic Structures*, volume 1, pages 123–166. Oxford University Press.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala.

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*, pages 241–247, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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.

Dan Jurafsky and James H Martin. 2014. *Speech and language processing*, volume 3. Pearson.

Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, 32(4):485–525, December.

Sandra Kübler and Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury, London, New York.

Gabriella Lapesa and Stefan Evert. 2017. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 394–400, Valencia, Spain, April. Association for Computational Linguistics.

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, December.

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

Wolfgang Maier, Miriam Kaeshammer, Peter Baumann, and Sandra Kübler. 2014. Discosuite - A Parser Test Suite for German Discontinuous Structures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Joakim Nivre, Željko Agić, and Lars Ahrenberg. 2017. Universal Dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a Novel Unit of Syntactic Analysis. *Syntax*, 15(4):354–396, December.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

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*, pages 247–269. Brill, January.

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*, pages 3132–3139, Istanbul, Turkey.

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*, number 7630 in Lecture Notes in Computer Science, pages 1–11. Springer, October.

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.

Rion Snow, Daniel Jurafsky, Andrew Y Ng, et al. 2004. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*, volume 17, pages 1297–1304.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, March.

Mark Twain. 1880. *A Tramp Abroad*. Chatto & Windus, London.

Andreas van Cranenburgh and Rens Bod. 2017. A Data-Oriented Model of Literary Language. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1:1228–1238.

Yannick Versley. 2013. A graph-based approach for implicit discourse relations. *Computational Linguistics in the Netherlands Journal*, 3:148–173.