Needles in Haystacks: Semi-Automatic Identification of Regional Grammatical Variation in Standard German

Tuggener, Don; Businger, Martin

DOI: https://doi.org/10.17885/heiup.361.509

Don Tuggener, Martin Businger

Needles in Haystacks: Semi-Automatic Identification of Regional Grammatical Variation in Standard German

Abstract This paper lays out a semi-automatic approach to identifying regional variation in the grammar of Standard German. Our approach takes as input manually defined templates of grammatical constructions that are automatically instantiated over a corpus collected from regional newspapers. These instantiations are automatically ranked by a metric that quantifies how specific an instantiation is for a region. Ranked lists of instantiations are compiled that contain instantiations specific to a region and are scanned manually by linguists to identify those that denote grammatical variants of Standard German. This approach enabled us to discover variants that so far have not been documented. With respect to research on variation within standard languages as seen from a more general perspective, we aim to contribute towards research strategies that clearly rely on empiricism rather than on intuition or bias.[1]

Keywords Association measures, corpus-driven approaches, diatopic variation, grammatical variation, standard language

1 Introduction

Varieties of a language can display differences in usage at any linguistic level, e.g. pronunciation, grammar, vocabulary or spelling. Variation regarding a feature of one of these linguistic levels—an intralinguistic feature—can correlate with extralinguistic factors, i.e. diastratic, diachronic, diaphasic or diatopic factors.

1 This paper received the support of the *Swiss National Science Foundation (SNSF)* and of the *Austrian Science Fund (FWF)*; grant numbers: *SNSF* 100015L_156613; *FWF* I 2067-G23. We would like to thank Gerard Adarve, Nicole Zellweger, Regula Gass, Reinhard Kunz, Marek Konopka and an anonymous reviewer for their help or comments on earlier versions of this paper.

This paper focuses on the correlation between grammatical variation and the diatopic dimension. Nevertheless, the approach and the methods laid out below are, in principle, applicable to any linguistic variation phenomena that correlate with features pertaining to any extralinguistic dimension.

This work is part of the project *Variantengrammatik des Standarddeutschen* ("Regional Variation in the Grammar of Standard German", cf. http://variantengrammatik.net/) which aims to identify and document grammatical variation in Standard German based on a regionally balanced corpus. For a detailed description of the project design, see Dürscheid and Elspaß (2015). We advocate an approach where language norms constituting a standard language—Standard German in our case—are to be reconstructed based on actual language usage; see Elspaß and Dürscheid (2017) for an extensive discussion on the term *Gebrauchsstandard*, i.e. 'standard language as it is used', and its interpretation in the context of the research project. The project will primarily result in an open-access website that compiles the project's findings and that serves as a searchable database of grammatical variation of Standard German (Dürscheid et al. in prep.).

The corpus compiled for this research project consists of texts from 68 online newspapers that were crawled for approximately one year, thus representing the German *Gebrauchsstandard* from all countries of Europe where German is used as an official language, divided into 15 regions (see Figure 1) based on the "Variantenwörterbuch" (first edition 2004 [= Ammon et al. 2004] and second edition 2016 [= Ammon/Bickel/Lenz et al. 2016], see e.g. map for Germany on p. LIII). The corpus contains roughly half a billion words distributed over 1.5 million articles which have been automatically processed with computational linguistics software (most importantly lemmatization, part-of-speech tagging, morphology, and dependency parsing). This corpus constitutes the basis for our experiments.

Clearly, reading a large text corpus like ours to discover regional grammatical variants is cumbersome and infeasible. Thus, the appeal of (semi-)automated methods that promise to alleviate much of the work is strong. A key interest of this contribution is thus to determine how well automatic and statistical methods from corpus and computational linguistics can assist grammarians in identifying regional grammatical variants. We propose a processing pipeline in which expert linguists and automatic ranking algorithms work together and evaluate how fruitful this collaboration is (Figure 1).

We proceed as follows. In section 2, our semi-automatic approach to identifying regional grammatical variants is described in detail and is compared to related work. In section 3, we examine selected examples of the results and discuss them in the context of recent research on grammatical variation within Standard German. The paper concludes with a summary (section 4).

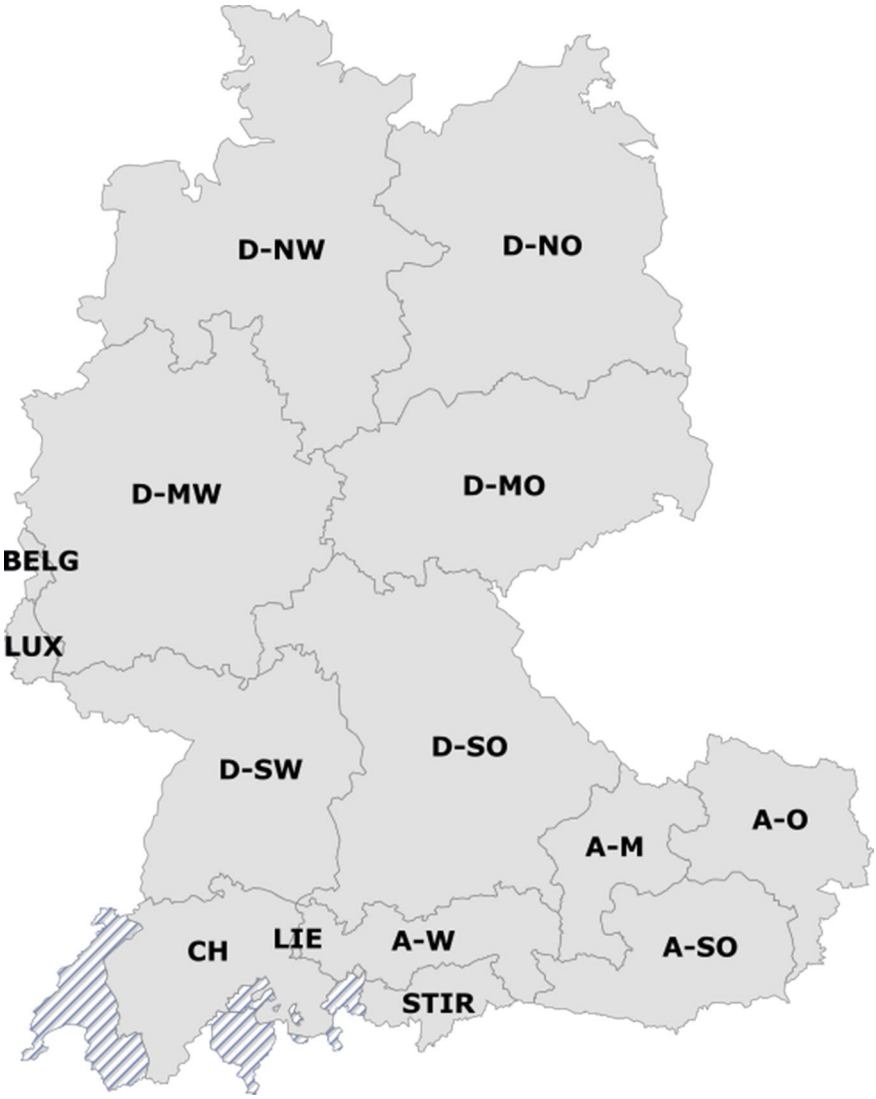

Figure 1: European countries and regions with German as an official language, with subregions.

2 Semi-automatic identification of grammatical variants

Before turning to our own approach (section 2.2), we briefly discuss relevant related work (section 2.1). Sections 2.3 and 2.4 explain our choice of a suitable ranking metric in detail.

2.1 Related work

One way of discovering grammatical variants is to have speakers from one country or region read newspapers of another country/region and mark the constructions that strike them as 'odd'. These constructions are then queried in a corpus and their distributions are analyzed statistically to verify whether there is sufficient support to categorize them as variants. Obviously, this approach is time-consuming and expensive. Another approach is to gather variants previously described in the literature and then query those in a corpus. The obvious drawback of this method is that it does not allow for any new variants to be identified.

The natural appeal of a corpus-driven approach therefore is its ability to overcome the drawbacks of the two methods described above. Firstly, it requires less time for a machine to read through large corpora, and secondly, the machine does not rely (heavily) on a priori assumptions about variation. Clearly, analyzing all random combinations and permutations of lexeme sequences and their various linguistic properties is infeasible even for smaller corpora. Furthermore, one cannot expect all grammatical constructions to show regional variants—on the contrary: we expect most constructions to be distributed homogeneously. Hence, using some initial and loose linguistic intuitions about which phenomena can be expected to show regional variation is a reasonable approach to help reduce search space.

Our work aligns with corpus linguistic research that aims to compare genres, registers, or varieties of languages. One area therein is the comparison of second language learner corpora to native speaker corpora, e.g. Laufer and Waldmann (2011), Cao and Xiao (2013), and Yoon (2016). Another area evolves around grammatically distinguishing the varieties of e.g. English, e.g. Mukherjee and Hoffmann (2006), Mukherjee (2009), and Xiao (2009). In this area, our approach is most closely related to Schneider and Zipp (2013), who also used an automatic dependency parser in their approach. An important advantage of using a dependency parser over so-called 'window-based' methods is that dependency parsing can tackle long-distance dependencies between lexemes that fall out of the window size. Window-based methods slide a window of a predefined size (e.g. two or five consecutive words) over the sentences in the corpus and analyze

the distribution of re-occurring word sequences. We experimented with different window-based approaches, including complex ngrams (i.e. replacing certain lexemes with their part-of-speech tags) along the lines of Bubenhofer (2015), but struggled to find a setup that yielded ranked lists which contained regional grammatical variants.

The aim of Schneider and Zipp (2013) was to identify novel combinations of verb and preposition in Indian and Fiji English in the International Corpus of English. They compared a fully manual approach to a semi-manual one. In the fully manual approach, the researcher first queried, on the one hand, a list of prepositions known to be productive and, on the other hand, an additional two prepositions that are commonly assumed to show variation in the literature. The combinations found were then compared to dictionaries that contain known variants, and those not contained in the dictionaries were labeled as unrecorded. The semi-automatic approach used a dependency parser and a metric to rank all found verb-preposition combinations which were then evaluated by the linguist. To automatically obtain ranked lists of verb-prepositions combinations, they scored each lexicalized combination in the Fiji and Indian English subcorpora with an observed over expected count ratio (compared to the BNC corpus). Combinations that were considered "unexpected" by the ratio were ranked high and then manually evaluated by a linguist.

The fully manual approach has the advantage of being highly accurate, i.e. the linguist will only pick those query results which are indeed variants. Clearly, the drawback of this method is that it is time-consuming and requires the researcher to know beforehand which lexical items (in their case a set of prepositions) are assumed to induce variation. The semi-manual, parser-assisted approach, on the other hand, has the advantage of not requiring a priori assumptions about the variation of specific lexical items but proceeds in a theory-agnostic, purely corpus-driven fashion. Its drawback is that automatic parsing yields errors and thus reduces the precision of the approach (returning false positives and missing true positives due to parsing errors).

In contrast to Schneider and Zipp (2013), we do not solely focus on combinations of verbs and prepositions. We are interested in all aspects of verbs and their subcategorization frames. That is, we query verb lemmas and all grammatical functions that they subcategorize for (e.g. direct/indirect objects, prepositional phrases, subclauses etc.). Furthermore, we are interested in word formation phenomena, e.g. the combination of verb stems and prefixes, and whether there are regional preferences for certain combinations. Another important difference of our setting to that of Schneider and Zipp (2013) is that our corpus comprises 15 subcorpora (corresponding to geographical regions), rather than two or three. Hence, computing the Observed-Expected ratio used in Schneider and Zipp (2013) would be computationally expensive, since it requires counting each verb

and preposition both together and separately for each subcorpus and the concatenation of the remaining subcorpora to decide whether a combination of a verb and a preposition is "unexpected". Our ranking metric requires less counting and does not need to partition the subcorpora in a one-versus-the-rest fashion to calculate a score for the specificity of a construction in a certain region.

2.2 Pipeline approach

Accounting for the discussion above, we define the following semi-automatic pipeline to discover novel grammatical variants:

Table 1: Pipeline approach.

1	Identify a general grammatical pattern that is assumed to show regional variation, e.g. verb valency.	Manual
2	Translate the pattern to a path or template construction in the dependency trees annotated in the corpus.	Manual
3	Instantiate the template over the corpus, track counts per region.	Automatic
4	Analyze the distribution of each instantiation with respect to the regions. Return a list of instantiations ranked by their specificity for a particular region.	Automatic
5	Inspect the list and manually distinguish between grammatical, orthographic, and lexical variants (and noise).	Manual

To illustrate the process, we walk through the following example: In step 1, we assume that verbs show regional variants with regard to the preposition that they subcategorize for. We formulate the template: verb + preposition (step 2), i.e. only the part of speech of the two items as well as their dependency relation (the preposition is governed by the verb) are specified. Next, in step 3, we automatically extract all lexicalized instantiations of the template from the dependency trees in the corpus, counting their occurrence per region. The following sentence is an example of an instantiation:

(1) Zunächst setzte sich Borna über Turbine Leipzig durch […][2]
first VERB REFL *Borna* over *Turbine Leipzig* VERB-PREFIX
'First, *Borna* won against *Turbine Leipzig* […]'

2 http://www.lvz.de/Region/Borna/Zwei-Heimsiege-Aufstieg-und-Belohnungsspiel (10 February 2017).

Given the automatic dependency analysis shown in Figure 2, we extract the following instance tuple:[3] <durchsetzen, über, D-Nordwest, 1> (i.e. <verb, preposition, region, count>).

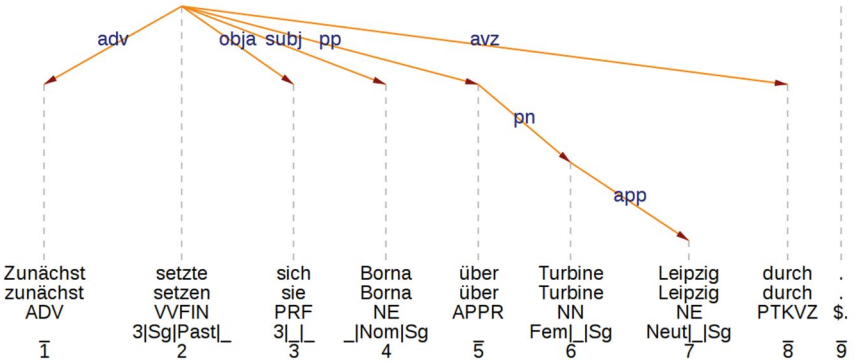

Figure 2: Output of the dependency parser for example sentence 1.

Having traversed all dependency trees in the corpus, the instantiations found are analysed and ranked with respect to their specificity for a region using a metric (cf. section 2.3) in step 4. Our example instantiation from above will rank high in this list because the verb *durchsetzen* 'prevail' commonly subcategorizes for the preposition *gegen* 'against' instead of über 'over' (see discussion of this verb below in section 3.3). Instantiations like *legen+in* ('put+in') will have a low rank, since they occur frequently in all regions.

In step 5, this list is inspected by a linguist to cherry-pick the instantiations that denote grammatical variants. This is necessary because the metric ranks all 'peculiar' constructions high, which means that orthographic (e.g. *ss* instead of *ß* and vice versa) and lexical (e.g. *paraphieren* 'to initial') as well as noise (e.g. verb instances containing encoding errors of *Umlauts*) are ranked high because the metric is not able to distinguish them.

Having outlined the approach, we turn to its core next, the ranking metric.

3 Note that in order to get the correct verb lemma (*durchsetzen*), we have to attach the separable verb prefix (*durch*) to the stem (*setzen*). Otherwise, the instantiation would wrongly be attested to the verb *setzen*. Fortunately, the dependency parser reliably identifies separated verb prefixes.

2.3 Ranking metric

The corpus linguistic literature contains a vast variety of metrics that aim to identify linguistic items that manifest some desirable properties (association, heterogeneous distribution etc.). Providing a comprehensive overview is beyond the scope of this work, and we refer readers to e.g. Evert (2004) and Gries (2008). Instead, we outline the requirements for a metric in our setting and motivate our choice based on them.

In our setting, the task of the metric is to assign a high rank to grammatical variants that occur in a (limited) set of regions. Hence, one criterion for the metric is that a template instantiation should be ranked high if it only occurs in a small number of regions. In other words, the rank of an instantiation should increase with the decreasing number of regions that contain it. Among those instantiations with such limited coverage in the corpus with respect to the regions, we want those to rank high that have a high frequency. We favor high frequency instantiations because we want to avoid the problem of defining an arbitrary minimum frequency threshold for including phenomena in the variation grammar wherever possible. Low-frequency instantiations also often cause problems with low expected values in subsequent statistical analyses (e.g. Chi Square). In addition, favoring high frequency phenomena acts as a natural filter against occasionalisms, typing errors and the like as well as various preprocessing problems, such as encoding errors and faulty dependency parses, which is essential since we work with real-world data and automatic preprocessing.

One metric that perfectly combines both desiderata is *Term Frequency Inverse Document Frequency* (TF IDF), well-known in Information Retrieval. TF IDF is widely used, e.g. for document indexing for search engines. A term is regarded as highly indicative for a document if it occurs frequently in the document (term frequency; TF), but at the same time occurs only in a small number of other documents in a collection (inverse document frequency; IDF). In our setting, we treat the template instantiations as the terms, and the regions as the documents.

More specifically, we calculate the normalized TF of a template instantiation t_i given a region r_j (e.g. <verb, preposition, region, count> = <durchsetzen, über, D-Nordwest, 216>) as:

$$TF = \frac{count(t_i, r_j)}{\sum_{k=1}^{n} count(t_k, r_j)}$$

i.e. by dividing the count of t_i in region r_j by the sum of all counts of all instantiations in r_j. This division normalizes TF to the size of the subcorpus r_j and lets us compare subcorpora of different sizes.

IDF is simply (the logarithm of) the ratio of all regions and the regions r that contain the template instantiation t_i:

$$IDF = \log_2 \frac{count(r_{k...n})}{count(r_{k...n}) \ni t_i}$$

TF IDF is the product of the two, i.e.:

$$TFIDF = TF \times IDF$$

Using this approach, we are able to rank all template instantiations both per region and for all regions combined by creating corresponding ranked lists (one for each region and one for all regions combined).

TF IDF has the advantage that it is relatively cheap to compute compared to other metrics like Observed-Expected ratios or Mutual Information because it does not require access to the counts of the individual components in the constructions (e.g. the separate counts of a verb and a preposition in the subcorpora, which are required by Mutual Information to calculate their association strength).

However, one downside of TF IDF is that in the IDF calculation, the dispersion of an instantiation (i.e. t_i) is not taken into account. This means that looking up the number of regions that contain t_i does not account for how well t_i is supported in those regions. For example, t_i might only occur once in a comparably large subcorpus, but with high frequency in three smaller subcorpora. However, all these occurrences are weighted equally. Conversely, another template instantiation t_k might occur frequently in the larger subcorpus and only once in each of the three smaller subcorpora. For both t_i and t_k, the IDF value will be the same, since they occur in an equal number of regions. However, their dispersions or distributions in the subcorpora are vastly different, and we would like our metric to reflect that. Thus, we introduce a notion of dispersion to the TF IDF calculation by multiplaying it with the *DISP* parameter, which is based on the count distributions, more specifically their residuals, and calculated as follows:

$$residual(t_I, r_j) = \frac{observed(t_i, r_j) - expected(t_i, r_j)}{\sqrt{expected(t_i, r_j)}}$$

$$DISP(t_i, r_j) = residual(t_i, r_j) - mean(residuals(t_i, r_{j...n}))$$

That is, we subtract the mean of all of t_i's residuals from that of the current region r_j. Note that if t_i's residual in r_j is above the mean, this yields a positive number and vice versa. Hence, all template instantiations whose residual in a given region is below the mean of all its residuals will render the TF IDF score

negative for that region and will rank it low in the list of specific constructions. Conversely, all instantiations with a positive difference to the residuals' mean will get a boost in the ranking. Our final metric then simply consists of:

$$TFIDFDISP = TF \times IDF \times DISP$$

There are other noteworthy metrics that rank construction in relation to the heterogeneity of dispersion. A whole family of statistical tests can serve as such a metric, e.g. Chi Square. One common problem of these tests (which we also encountered during preliminary experiments) is that they tend to yield high significance levels for low-frequency phenomena in large corpora (Gries 2008). Since we are interested in highly frequent phenomena, this is a clear disadvantage. The same applies to (Pointwise) Mutual Information-based metrics. An interesting, intuitive and easily computed metric of dispersion is presented in Gries (2008), called *deviation of proportions*. It is also based on normalized values for observed and expected frequencies and their differences, similar to our *DISP* parameter. We will empirically compare our metric to the unaltered version of TF IDF and Gries' *deviation of proportions* (Gries DP henceforth) in the next section.

2.4 Comparison of metrics

In this section, we compare the ranked lists that emerge when we apply the three ranking metrics outlined above, i.e. TF IDF, TF IDF DISP, and Gries DP to a set of instantiated templates. The instantiations that we rank stem from the combination of two verb-related templates, i.e. verbs and the (lexicalized) prepositions they subcategorize for,[4] and verbs and the (unlexicalized) grammatical functions in their subcategorization frame.[5] We compare the lists by assigning the top 100 instantiations in each to five categories: grammatical variants (which we are interested in), lexical variants (interesting, but not in our focus), *ss/ß* alternation (irrelevant in our case), non-variants (instantiations that are overrepresented in some area of the corpus due to the sampling process, e.g. *sich qualifizieren* 'to qualify' with reflexive morpheme *sich* is ranked high because of oversampling of the sports section), and preprocessing/encoding errors (noise in the corpus). The distribution of the instantiations over these categories can then serve as an estimate of how fruitful it is for a researcher to manually scan each list in terms of the number of returned novel variants, which serves as an evaluation.

4 An example instantiation is: *ersuchen + um* 'to request sth.'.

5 E.g.: *beantragen* 'to request, to apply for' + *dative object* or *beantragen + accusative object.*

As mentioned above, we are looking for phenomena that feature a solid support in the corpus and are thus interested in high frequency instantiations. To evaluate how well the metrics perform in this regard, we count how many of the top 100 instantiations in each list have a frequency of at least 10 occurrences. Note that for Gries DP, all counts in the corpus are considered, while for the TF IDF metrics only the counts in the respective region where an instantiation was ranked high are taken into account (thus the overall occurrences are even higher). To our surprise, we found that in the Gries DP list, only 1 of the top 100 instantiations has a corpus frequency of at least 10, while the top 100 lists created by TF IDF and TF IDF DISP feature 81 and 80 instantiations respectively, with a frequency over 10 in the region where they were ranked high. The Gries DP metric seems to suffer from oversensitivity to low count phenomena, at least in our setting.[6] Since we deem instantiations with a count below 10 as not sufficiently supported in the corpus, we removed all instantiations with a frequency below 10 from the Gries DP list, and then again took the top ranked 100 among the remaining instances for the further comparison.

Next, we analyze the top 100 ranked instantiations of the categories introduced above.

Table 2: Category breakdown per metric.

	Gries DP	TF IDF	TF IDF DISP
Preprocessing / encoding errors	19	31	32
ss/ß alternation: *begrüssen, begrüßen* 'to welcome'	28	43	27
Lexical variants: *paraphieren* 'to initial'	27	11	21
Non-variants	11	8	7
Grammatical variants	15	7	13

As shown in table 2, the filtered Gries DP list and the TF IDF DISP list return 13 to 15 instantiations that denote grammatical variants, while the TF IDF list only contains 7. TF IDF also returns the most *ss/ß* alternations (43), which the added DISP parameter is able to reduce (to 27). Gries DP is most robust against ranking preprocessing and encoding errors, but returns more lexical and non-variants than TF IDF DISP.

An interesting question is whether the different metrics return an overlapping set of instantiations in their top 100 lists or whether they favor different

6 An issue in the calculation of Gries DP in this respect is that it takes the absolute value of the differences between observed and expected values. Low count instances with a high negative difference to the expected value (which are based on normalized subcorpora sizes) therefore drastically increase the sum of the differences. Furthermore, the metric does not take into account the overall frequency of an instance, unlike TF IDF.

instantiations. We measure the overlap of the instantiations in each list in a pairwise manner in table 3.

Table 3: Pairwise overlap in the ranked lists.

Gries DP ∩ TF IDF	19
Gries DP ∩ TF IDF DISP	20
TF IDF DISP ∩ TF IDF	73

Clearly the ranked lists of the TF IDF metrics are more similar to each other than to the Gries DP list. Yet more than 25% of the instantiations in their lists are unique. Compared to the Gries DP list, there is little overlap with the TF IDF metrics. This suggests that the two approaches are complementary. Indeed, if we combine all the grammatical variants found in the three top 100 lists, we obtain a total of 22 unique grammatical variants.

One aspect that distinguishes the variants found in the Gries DP list and the TF IDF lists is their average frequency in the corpus compared to the average frequencies of the variants in the respective regions where the TF IDF metrics found them, as shown in table 4.

Table 4: Average frequency of variants found per metric.

	# Variants	Avg. frequency
Gries DP	15	42 (whole corpus)
TF IDF (region)	7	77 (region)
TF IDF DISP (region)	13	138 (region)

The table shows that the variants found in the Gries DP list have a much lower frequency compared to the TF IDF based variants. Furthermore, half of the 15 variants in the Gries DP list have a frequency below 15. Given a corpus of over half a billion tokens, the question arises whether such counts provide enough support to claim a variant.

Another downside of Gries DP is that it does not indicate directly which subcorpora (in our case regions) drive a high deviation of proportions,[7] if one is found, while the TF IDF-based measures can return ranked lists for any partition of the subcorpora or the whole corpus. Hence, based on the TF IDF measures, we can easily investigate instantiations that are specific to a given region or country.

After the comparison of the metrics, we now turn to some examples of newly discovered grammatical variants.

7 One could look at high positive differences between observed and expected, though.

3 Result examples: unknown grammatical variants

This section aims to illustrate the potential of the method by focusing on a small selection of results. After some initial remarks on the state of research and an overview of the results, we turn to specific examples from the areas of word formation and valency that we found using our approach.

3.1 Grammatical variation at different linguistic levels

Grammatical variation phenomena can be assigned to either morphology or syntax. In the field of morphology, we find areal (regional) variation in terms of both word formation and inflection. A vast array of morphological variants has been documented in the first and second edition of the *Variantenwörterbuch* (Ammon et al. 2004 and Ammon/Bickel/Lenz et al. 2016 respectively), which is undoubtedly the most comprehensive reference work on linguistic variation in the (written) German standard language to date. The *Variantenwörterbuch* aims primarily to document lexical variation, but it also includes variation phenomena in inflection (e.g. plural forms of nouns) and in word formation. As for syntax, the *Variantenwörterbuch* documents some variation with regard to valency, but syntactic phenomena are not taken into account systematically. This reflects the fact that research on variation within Standard German has traditionally focused on the lexicon and on morphology, rather than on syntax (cf. Niehaus 2015).

The semi-automatic approach outlined above is inherently not restricted to 'one-word-phenomena'. It has proven to be successful with a range of corpus findings in relation to word formation and valency (subcategorization). Overall, besides reproducing 23 previously known variants (i.e. documented in the *Variantenwörterbuch* or in at least one other relevant reference work for Standard German grammar, cf. examples below), we were able to discover 30 previously undocumented variants. In the next section, we present examples of areal grammatical variation still undocumented in relevant reference works. These phenomena were detected by using the pipeline approach described in section 2.

3.2 Word formation

The reflexive verbs s*ich berappeln* and *sich aufrappeln* both mean 'to stand up again' and, in a more figurative sense, 'to pull oneself together'. The key difference between the two verbs is a morphological one: while the verb *berappeln* has the unstressed, inseparable prefix *be*, the verb *aufrappeln* has the stressed and separable prefix *auf*. As is shown on the map in Figure 3, s*ich berappeln* is

found only in newspapers in Germany (and, occasionally, Belgium and Luxembourg). It is not attested in corpus texts from Austria or Switzerland. Note that *sich berappeln* is not mentioned as a regional variant in the *Variantenwörterbuch* (neither in Ammon et al. 2004 nor in Ammon/Bickel/Lenz et al. 2016). Neither does duden.de,[8] among the most widely used online works of reference, mention any regional restrictions on the use of *sich berappeln*. One might wonder if this 'gap' is purely accidental or can be attributed to a larger fundamental factor. We argue for the latter in the following section.

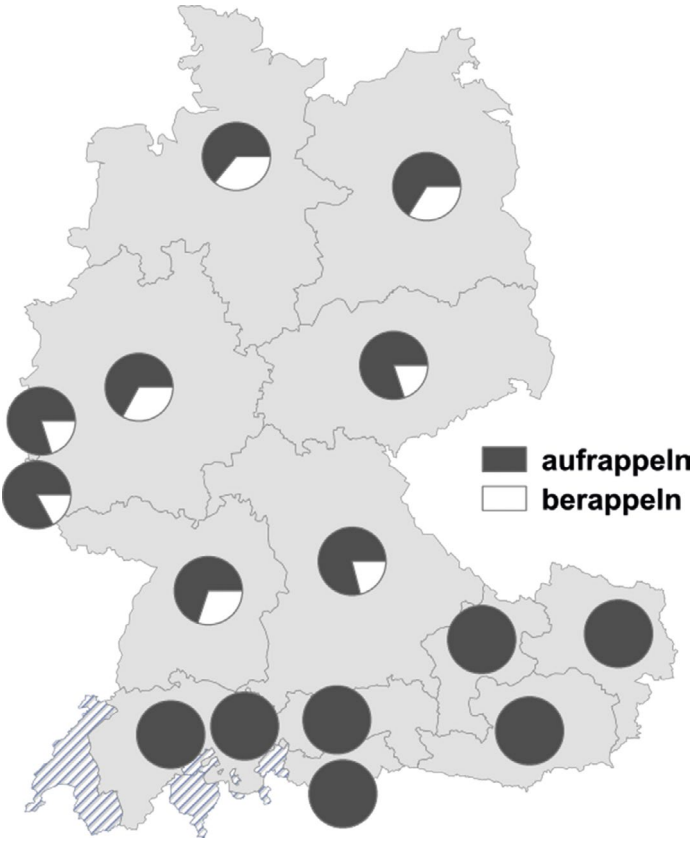

Figure 3: Distribution of *sich berappeln* vs. *sich aufrappeln.*

If only the—traditionally prevailing—manual method is used, linguistic features of Standard German that are used exclusively or mainly in Germany tend to pass unnoticed as regional variants by linguists (cf. Dürscheid and Sutter

8 duden.de (9 February 2017).

2014). This is due to a widespread bias in which the Standard German language of (Northern) Germany is thought to define the (only) norm (Schmidlin 2011: 208). According to Clyne (2004: 297), the varieties of pluricentric languages like German usually relate asymmetrically, with one variety dominating. Characteristic of such situations is the following, among other things: the dominant (D) variety has more effective political and economic resources for being exported, e.g. by means of reference works (dictionaries, textbooks etc.); users of the D variety may believe that there is no linguistic variation in *written standard* language; users of the D variety, as far as they notice differences between their own D variety and another variety, consider such other varieties as "exotic, cute" and, most importantly, "non-standard" (Clyne 2004: 297). This attitude is the basis of what can be identified as 'ideology of homogenism' (Elspaß and Niehaus 2014).

German as used in Germany clearly plays the role of the D variety. As a result, Germany-specific variants are less frequently marked as national or regional variants in reference works than e.g. national variants as found in Austria. This has been shown by systematic research on numerous grammar reference works (see Dürscheid and Sutter 2014 for details).

In this context, a second example worth noting is *bepöbeln* in contrast to *anpöbeln* 'to accost, to verbally abuse'. In our corpus, *bepöbeln* is confirmed to be used exclusively in Germany (in all regions except D-southwest; mainly in D-northwest). Again, *bepöbeln*, like *berappeln*, is not mentioned in the *Variantenwörterbuch* (either edition).

To conclude, the two examples, *sich berappeln* and *bepöbeln* indicate that a (semi-) automatic, at least partially corpus-driven, and thus less biased approach is superior to a purely manual one when it comes to identifying linguistic features of the dominant variety of a pluricentric language.

In the next section, we turn to examples of variation in subcategorization frames of verbs.

3.3 Valency

As a first example on valency, let us turn to the reflexive verb *sich durchsetzen* 'to prevail (against)', which can be combined with more than one preposition without difference in meaning (but note the caveat in footnote 10): *gegen,* über and *gegenüber* (meaning 'against'). *Gegen* is, as expected, by far the most frequently used preposition with *sich durchsetzen* in the corpus (black on the map in Figure 4). In contrast, the preposition *über* (gray on the map)—the one preposition that ranked high in combination with *sich durchsetzen* in our metric—is used almost exclusively in the center-east of Germany (one of the six predefined

German subregions) (cf. example (1) in section 2.2). A third attested preposition is *gegenüber*, which is generally rare and not restricted to particular regions (see Figure 4).

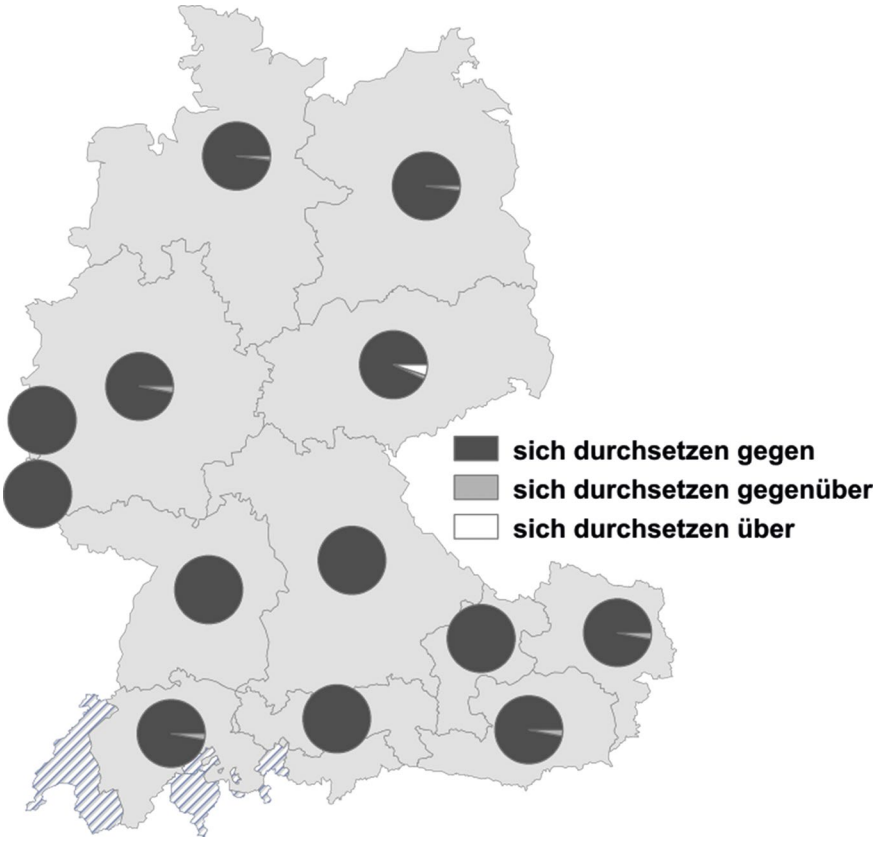

Figure 4: Distribution of *sich durchsetzen gegen* vs. *sich durchsetzen gegenüber* vs. *sich durchsetzen über.*

The verb *durchsetzen* is not listed in the *Variantenwörterbuch* (first and second edition) and it is therefore not possible to find any reference to prepositions selected by this verb there. The *Wörterbuch der Präpositionen* (Müller 2013)[9] has the prepositions *gegen* and *gegenüber* for *durchsetzen*, but not *über*—the one preposition that is of interest here because of its diatopically restricted usage.

9 This dictionary does not consider regional variation, but lists a large number of German verbs, adjectives and nouns with their respective prepositions.

We conclude that it is a hitherto unknown fact that *sich durchsetzen* is used with the preposition *über* in Standard German texts.[10]

A second example of regional variation in subcategorization frames is the verb *verlautbaren* 'to announce (officially), to proclaim'. According to the instantiations found in the corpus, *verlautbaren* ranked high in terms of our metric when governing a direct object NP.

The manual analysis of the phenomenon (in and after step 5, cf. section 2.2) proved to be complex. It is necessary to distinguish between several formal types of objects:

- (A) Nominal and pronominal objects: use of indefinite pronouns like *nichts* 'nothing' or *etwas* 'something' can be confirmed in almost all countries/regions without regional preferences. Examples with objects in the form of indefinite pronouns were therefore excluded (and are not represented on the map in Figure 5). Instead, only examples with a 'full NP'[11] object (including examples with full NP subjects in passive sentences as (2a) below) were counted.
- (B) Object clauses: subordinate clauses introduced by the subjunction *dass* 'that' or object clauses without subjunction (see example (2b)) together constitute one category.
- (C) No object (intransitive): usages of *verlautbaren* without any object at all commonly appear in a subordinate clause headed by *wie* 'as' which depends on the matrix clause (see example (2c), where the matrix clause is left out).

(2) a. Erst am Samstag soll […] das Endergebnis verlautbart werden.[12]
Only on Saturday is-said the final-result announced PASSIVE-AUX
'The result will not be announced until Saturday.'

b. Das Auswärtige Amt verlautbarte, die Echtheit des Videos
The *Auswärtige Amt* announced the authenticity of-the video
werde noch geprüft.[13]
PASSIVE-AUX still verified
'The Federal Foreign Office [of Germany] announced that the authenticity of the video remains to be verified.'

10 It must be noted that 35 out of 36 manually inspected corpus examples of *sich durchsetzen über* (= 97 %) were found in the sports section of the respective online newspapers. No such preference for a specific text type can be observed for *sich durchsetzen* when governing one of the other prepositions (*gegenüber* or *gegen*). Further research as to the (non-) interchangeability of the three prepositions governed by *sich durchsetzen* is necessary.

11 By the informal term "full NP", we refer to a nominal phrase headed by a noun, not a pronoun.

12 http://derstandard.at/1350260818406/Wahlergebnis-fruehestens-am-Samstag (10 February 2017).

13 http://www.schwaebische.de/region_artikel,-Filiz-G-soll-angeblich-freigepresst-werden-_arid,5227115_toid,351.html (22 March 2012).

c. Wie am Wochenende verlautbart wurde, […][14]
As on-the weekend announced PASSIVE-AUX
'As was announced on the weekend, […]'

In the resulting map (Figure 5), *verlautbaren* governing a full noun phrase (i.e. excluding pronouns) functioning as the object (black on the map; cf. example 2a) is contrasted with examples where the verb governs a clausal object or no object at all (white; cf. example 2b/c).

To sum up: in the Austrian regions, examples with full NP-objects constitute between 17 % (A-southeast) and 35 % (A-west). By contrast, in the middle and northern regions of Germany, this phenomenon is rare.

It is therefore possible to surmise that intricate and 'non-intuitive' variation phenomena, like the case of *verlautbaren,* would probably not be detected with a purely manual approach.

Let us now turn to a third example. In the area of verb valency, the diatopically conditioned alternation between reflexive and non-reflexive usage of certain verbs has received some attention in the literature. It has been presumed that speakers and writers of German in Austria tend to often use the reflexive pronoun *sich* with several verbs (Ebner 2008: 44f., Ziegler 2010). Current research has confirmed the alleged tendency to some extent (Dürscheid et al. in prep.). For example, the verb *erwarten* 'to expect' can be used reflexively, i.e. with a reflexive pronoun, in the same meaning as when it is used without a reflexive pronoun:

(3) Was erwarten Sie sich von dem Projekt?[15]
What expect you REFL from the project
'What are you expecting from the project?'

This usage is rare outside of Austria and South Tyrol. One is therefore tempted—based on hypothesis—to search for more instances of reflexive verbs in Austria (and South Tyrol) only. On the other hand, adopting a 'theory-agnostic' approach, like the one advocated in this paper, helps to ensure that no relevant data is overlooked. A case in point is the reflexive use of the verb *ausprobieren* 'to try', which ranked high in our metric when used with a reflexive pronoun.

14 http://www.krone.at/oesterreich/wahlbeteiligung-in-graz-sinkt-seit-1945-kontinuierlich-mangel-an-themen-story-341329 (10 February 2017).

15 http://www.nachrichten.at/oberoesterreich/wels/Gaesterekord-in-der-Vitalwelt-Bad-Schallerbach;art67,1059781 (8 February 2017).

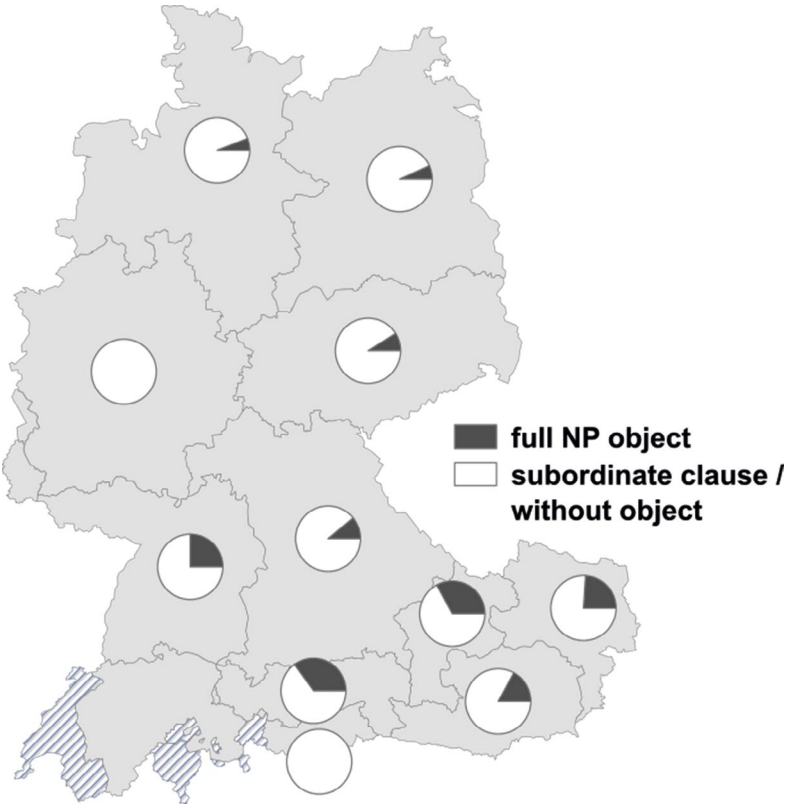

Figure 5: Distribution of *verlautbaren + full NP object* vs. *verlautbaren + with subordinate clause / without object.*

(4) In den Ferienkursen […] können sich Kinder ab zehn Jahren […]
in the holiday courses […] can REFL children from ten years […]
schauspielerisch ausprobieren.[16]
as-actors try-out
'In the holiday courses, children from the age of ten can dabble in acting.'

Sich ausprobieren (in/als) 'to try out something / to give something a try (in/as)' is used almost exclusively in Germany where it is most frequent in the subregions north-east (35 % of all hits in the corpus) and center-east (25 %). It is used less frequently in the other German subregions and in Belgium, and is hardly used in the other German-speaking countries/regions in Europe. To sum up: until very

16 http://www.tagesspiegel.de/berlin/gut-geruestet-fuer-die-freien-tage-unsere-freizeittipps-fuer-die-ferien/7717290.html (8 February 2017).

recently, the diatopically conditioned use of *sich ausprobieren* has not been documented.[17] We consider a semi-automatic approach promising for filling in gaps on the map of regional variation of German or, for that matter, of any language—gaps that tend to be overlooked in purely hypothesis-driven research settings.

From the point of view of variationist linguistics, the valency patterns presented in this section are clearly diatopically conditioned. At the same time, it is worth noting that these results cannot be interpreted in a strictly pluricentric model, i.e. a model where 'national varieties' are constitutive elements. National boundaries *are* an extralinguistic factor that can correlate with the diatopical distribution of variants in a standard language, but, at the same time, variation within or across national boundaries must be included systematically and without bias (cf. Niehaus 2015 as well as Elspaß and Dürscheid (2017) for discussion and references on *pluricentricity* vs. *pluriareality* in German).

4 Conclusion

This paper presented a semi-automatic method to identify regional grammatical variants. We discussed our pipeline approach that combines linguistic expertise and automatic ranking metrics and showed that it yields a fruitful combination in the sense that we discovered a reasonable number of (novel) variants while not having to go through too much noise (e.g. preprocessing errors) in the generated lists. We proposed an extended version of TF IDF which returned the most usable ranked lists containing variants with a substantial frequency in our corpus, while other metrics produced fewer variants or variants with less support in the corpus.

A theory-agnostic, (at least partially) data-driven approach like the one being put forward here is especially valuable in a field where ideologically colored discussions are common, even among linguists:

> "Offensichtlich wird die Diskussion um die Rolle der Arealität in der deutschen Standardsprache […] bisher eher politisch-ideologisch geführt" (Niehaus 2015: 138).
> 'It seems that the role of areality in the German standard language has been discussed in a rather political-ideological manner so far.'

17 The reflexive use of *ausprobieren* is mentioned neither in duden.de (last accessed: 8 February 2017)—as opposed to *sich versuchen in/als* '=' that is used in all German-speaking countries/regions, which *is* mentioned—nor in *Duden Zweifelsfälle* (2016), and *sich ausprobieren* was also not entered in the first edition of the *Variantenwörterbuch* (Ammon et al. 2004). However, it has been included in the second edition, where it is marked as "D", i.e. as a variant of Germany as a whole (Ammon/Bickel/Lenz et al. 2016: 69).

This paper contributes towards overcoming the lack of empiricism in research on variation within standard languages (cf. Niehaus 2015: 139).

References

Ammon, Ulrich et al. 2004. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol.* Berlin/New York: de Gruyter.

Ammon, Ulrich, Hans Bickel and Alexandra N. Lenz et al. 2016. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen.* 2., völlig neu bearbeitete und erweiterte Auflage. Berlin/Boston: de Gruyter.

Bubenhofer, Noah. 2015. Muster aus korpuslinguistischer Sicht. In Christa Dürscheid and Jan Georg Schneider (eds.), *Handbuch Satz – Äußerung – Schema*, 485–502. Berlin/New York: de Gruyter.

Cao, Yan and Richard Xiao. 2013. A multi-dimensional contrastive study of English abstracts by native and non-native writers. *Corpora* 8.2: 209–234.

Clyne, Michael. 2004. Pluricentric Language. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier and Peter Trudgill (eds.), *Handbooks of linguistics and communication science.* Vol. 3: Sociolinguistics. 2., completely revised and extended edition, 296–300. Berlin/New York: de Gruyter.

duden.de A Website of Bibliographisches Institut GmbH. (8 February 2017).

[Dudengrammatik]. 2016. *Duden. Die Grammatik. Unentbehrlich für richtiges Deutsch.* 9., vollständig überarbeitete und aktualisierte Auflage. Berlin: Dudenverlag (= Duden 4).

[Duden Zweifelsfälle]. 2016. *Richtiges und gutes Deutsch.* 8., vollständig überarbeitete und erweiterte Auflage. Berlin: Dudenverlag (= Duden 9).

Dürscheid, Christa et al. In prep. *Variantengrammatik des Standarddeutschen. Ein Online-Nachschlagewerk.* Verfasst von einem Autorenteam unter der Leitung von Christa Dürscheid, Stephan Elspaß und Arne Ziegler.

Dürscheid, Christa and Stephan Elspaß. 2015. Variantengrammatik des Standarddeutschen. In Roland Kehrein, Alfred Lameli and Stefan Rabanus (eds.), *Regionale Variation des Deutschen – Projekte und Perspektiven*, 563–584. Berlin/Boston: de Gruyter.

Dürscheid, Christa and Patrizia Sutter. 2014. Grammatische Helvetismen im Wörterbuch. *Zeitschrift für Angewandte Linguistik* 60.1: 37–65.

Ebner, Jakob. 2008. Österreichisches Deutsch. Eine Einführung. Mannheim etc: Dudenverlag.

Elspaß, Stephan and Christa Dürscheid. 2017. Areale Variation in den Gebrauchsstandards des Deutschen. In Marek Konopka and Angelika Wöllstein (eds.), *Grammatische Variation – empirische Zugänge und theoretische Modellierung*, 85–104. Berlin/Boston: de Gruyter (= Jahrbuch des Instituts für Deutsche Sprache 2016).

Elspaß, Stephan and Konstantin Niehaus. 2014. The standardization of a modern pluriareal language. Concepts and corpus designs for German and beyond. *Orð og tunga* 16: 47–67.

Evert, Stefan 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, University of Stuttgart.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13.4: 403–437.

Laufer, Batia and Tina Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61.2: 647–672.

Mukherjee, Joybrato. 2009. The lexicogrammar of present-day Indian English. In Ute Römer and Rainer Schulze (eds.), *Exploring the Lexis-Grammar Interface*, 117–135. Amsterdam: John Benjamins.

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27.2: 147–173.

Müller, Wolfgang. 2013. *Das Wörterbuch deutscher Präpositionen. Die Verwendung als Anschluss an Verben, Substantive, Adjektive und Adverbien*. Berlin/Boston: de Gruyter.

Niehaus, Konstantin. 2015. Areale Variation in der Syntax des Standarddeutschen. Ergebnisse zum Sprachgebrauch und zur Frage Plurizentrik vs. Pluriarealität. *Zeitschrift für Dialektologie und Linguistik* 82.2: 133–168.

Schmidlin, Regula. 2011. *Die Vielfalt des Deutschen. Standard und Variation. Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*. Berlin/New York: de Gruyter (= Studia Linguistica Germanica 106).

Schneider, Gerold and Lena Zipp. 2013. Discovering new verb-preposition combinations in New Englishes. *Studies in Variation, Contacts and Change in English* Vol. 13. http://www.helsinki.fi/varieng/journal/volumes/13/schneider_zipp/ (3 February 2017).

variantengrammatik.net Website of the research project *Regional Variation in the Grammar of Standard German* (14 February 2017).

Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28.4: 421–450.

Yoon, Hyung-Jo. 2016. Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing* 34: 42–57.

Ziegler, Arne. 2010. ›Er erwartet sich nur das Beste …‹ Reflexivierungstendenz und Ausbau des Verbalparadigmas in der österreichischen Standardsprache. In Dagmar Bittner and Livio Gaeta (eds.), *Kodierungstechniken im Wandel. Das Zusammenspiel von Analytik und Synthese im Gegenwartsdeutschen*, 65–81. Berlin/New York: de Gruyter (= Linguistik – Impulse und Tendenzen 34).