Needles in haystacks: Semi-automatic identification of regional grammatical variation in Standard German using a large corpus

Don Tuggener and Martin Businger (Universität Zürich)

We present a semi-automatic approach to identify regional variation in the grammar of Standard German. Our approach takes as input templates of grammatical constructions that are instantiated over a large corpus gathered from regional newspapers.[1] The instantiations are ranked by a metric that quantifies how specific an instantiation is for a given region. For each region, we then create lists of instantiations that are ranked by their specificity for the region. The lists are manually scanned by linguists to cherry-pick instantiations that denote grammatical variants of Standard German. Using this approach, we discover grammatical variants that have not been documented so far.

As input to our processing pipeline, we define templates that capture grammatical constructions which we assume to expose regional variation. For example, we assume that verbs tend to show regional variation w.r.t. the prepositions and grammatical functions that they subcategorize. Hence, we define a template that collects all verb lemmas and their subcategorization frames (prepositions and grammatical functions). This template is then instantiated for each encountered verb during the traversal of the dependency parse of the corpus. Occurrence and frequency of each instantiation are tracked for each annotated region in the corpus.

Next, we require a metric that ranks the instantiations along two axes: a) We want instantiations that occur in few regions to score high (because it indicates that they are regionally confined), and b) among those that occur in few regions, we want high frequent ones to score high. Both these desiderata are captured by a well-known measure from Information Retrieval: Term Frequency – Inverse Document Frequency (TF IDF). To apply TF IDF in our setting, we define the regions to be the documents.

Manual inspection of the ranked lists is necessary, since an instantiation might rank high due to pragmatic reasons. For instance, a verb might often be mentioned in the context of a local event or incident in a given region. Furthermore, since the linguistic processing of the newspaper articles is fully automated, there might be errors that frequently occur in a particular region. Hence, the grammatical variants have to be cherry-picked from the ranked list of instantiations per region, similar to Schneider and Zipp (2013).

In the talk, we will exemplify the construction of templates over the dependency parses and argue for the use of parser-induced collocates over window-based methods to identify regionally distributed grammatical variants. Furthermore, we will outline advantages of the adapted TF IDF measure over related association measures and statistical tests in our setting. Finally, we will present novel grammatical variants identified by our approach that have been undocumented so far in related work.

[1] Our research is based on a large corpus of German that consists of more than 1.5 million texts (approximately 575 million tokens) from 68 online newspapers of all seven European countries where German is the or one of the official languages. The corpus is POS-tagged (TreeTagger, RFTagger) and automatically parsed with the ParZu dependency parser (Sennrich et al. 2013). Most importantly, the corpus is divided into 15 subcorpora representing geographical areas (countries or parts of countries, e.g. Liechtenstein or Western Austria), cf. Ammon (2004).

Data and references

Example Results:
Our approach has proven successful by an array of corpus findings in the areas of word formation and subcategorization of verbs. Here, grammatical variation that is not documented so far in widely used dictionaries and established reference works is of particular interest.

- With regard to word formation, the verbs *sich aufrappeln / sich berappeln* 'to pull oneself together' and *anpöbeln / bepöbeln* 'to accost, to verbally abuse' display regional variation that is not mentioned in Ammon (2004), duden.de and other relevant sources. According to our corpus findings, *sich berappeln* is used in Germany (alongside with the more frequent verb *sich aufrappeln*), but neither in Switzerland nor in Austria where only *sich aufrappeln* is used. The example illustrates that a (semi-)automatic approach is superior to a manual one with respect to identifying linguistic features of Standard German that are *specific to Germany:* Using the (traditionally prevailing) manual method, words or grammatical features that are used exclusively or mainly in Germany pass unnoticed as regional variants by linguists, due to a widespread bias where the Standard German language of (Northern) Germany is considered as – wrongly so – 'the norm'; as a result, in reference works Germany-specific variants are more rarely marked as national or regional variants than e.g. national variants from Austria (cf. Dürscheid and Sutter 2014).
- An example for variation regarding subcategorization of verbs is *sich durchsetzen* 'to prevail, to assert oneself (against)'. Several prepositions can be used with this verb. The semi-automatic approach identified the preposition *über* as a possible head of a dependent PP, semantically exchangeable with the preposition *gegen* and *gegenüber,* but used mainly in one German subregion. The preposition *über* is not listed for the verb in question in Müller (2013) or in other relevant reference works. Similarly, the verb *verlautbaren* 'to proclaim' can be shown to be used with an object NP significantly more often in Austria compared to other German speaking countries or regions (whereas objects of *verlautbaren* in the form of subordinate clauses can be attested in all regional varieties) – a hitherto unknown fact.

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.

duden.de (13.6.2016)

Dürscheid, Christa/ Sutter, Patrizia (2014): Grammatische Helvetismen im Wörterbuch. In: Zeitschrift für Angewandte Linguistik 60/1. 37–65.

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

Schneider, Gerold/ Zipp, Lena (2013): Discovering new verb-preposition combinations in New Englishes. In: Studies in Variation, Contacts and Change in English. Volume 13.

Sennrich, Rico/ Volk, Martin/ Schneider, Gerold (2013): Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2013. Hissar, Bulgaria. 601–609.