Register phenomena and the CoreGram architecture for multilingual grammar development

Stefan Müller, Antonio Machicao y Priemer,
Roland Schäfer, and Felix Bildhauer

1 Introduction

It is not only important what is said but also how it is said. Language users can use and recognize certain registers in communication. For example, people talk differently to a cab driver during a ride, in a job interview, and to their friends in a pub. While the existence of register phenomena is obvious to everybody dealing with language (cf. Paolillo 2000; Bender 2007; Adger 2006; a.o.), there is probably no such thing as a taxonomy of registers for a given language that most researchers would agree on (let alone a "universal" inventory of registers, see Schäfer et al. 2022). Considerable confusion exists regarding the delineation of registers and related categories such as "style" and "genre". Furthermore, the likely fuzzy boundaries between registers make it notoriously difficult to even agree on necessary and/or sufficient conditions (such as the occurrence of particular linguistic features) for category membership (Biber & Conrad 2009; see Argamon 2019). However, it is obvious that all parts of the linguistic system that have been studied in HPSG play a role in modeling register phenomena (Bender 2007: 354). For example, whether reduced forms of words are used or not (phonology/morphology), whether formal or less formal vocabulary is used (lexicon connecting phonology, syntax, semantics), whether complex and elaborated relative clauses are used or not (syntax), whether we use an imprecise expression ("half past 3" vs. 3:32) (semantics, Solt 2015) and even pragmatic issues like the use of metaphors and irony is register-dependent: certain exaggerations are just inappropriate in job interviews.

In a data-driven analysis to be reported elsewhere (Schäfer et al. 2022), we have used Bayesian generative models (Latent Dirichlet Allocation, Blei et al. 2003) to infer clusters of documents (= potential registers) in a large corpus of German based on the distribution of linguistic signs in the documents. While superficially similar to work by Douglas Biber (e.g., Biber 1988; 1995), our approach is entirely different. It is fully probabilistic and allows for many-to-many associations between linguistic signs, registers, and documents, and it does not rely on available a priori register taxonomies. In a further step of manual annotation, we managed to identify situational-functional parameters such as a higher level of education, proximity, or interactivity reliably for the potential registers. We find, for example, that some registers are associated with a high probability of occurrences of adverbs, certain tense forms, or more complex phenomena like passives and clausal pre-fields. It is the purpose of the work presented here to provide an implementation of such findings in a formal grammar.

From the perspective of both grammar theory and psycholinguistics, one overarching question is how variation in grammar (including register variation) is encoded in speakers' grammars, and how speakers use it. Different answers have been given in various frameworks and with respect to diverse sub-components of grammar. One option is to assume that speakers deal with different registers by using a set of distinct grammars or a single grammar with a separate module encoding variation (Yang 2002; Adger 2006). In contrast, it is also conceivable that speakers use a single grammar with all information about the variation encoded in it (Paolillo 2000; Bender 2001; 2007; Pierrehumbert 2008; Hilpert 2013). While it is not certain that such questions can ultimately be answered based one empirical evidence, the goal of this paper is to explore ways in which either approach could be implemented in HPSG.

Using different grammars for different registers bears some similarity to the approach of Søgaard & Haugereid (2007), who propose a grammar for Scandinavian containing subgrammars for Danish,

Norwegian and Swedish. The authors use a LANGUAGE feature that serves to identify the language (or languages) of a linguistic object. Such a model of different languages is not necessarily related to cognitive reality, simply because many speakers only speak one of the three languages. Register variation is fundamentally different in this respect because speakers are always able to understand and produce utterances in various registers.

In what follows, we will compare two potential approaches to register modeling in HPSG: one assumes multiple grammars for multiple registers (Section 2) and one assumes one grammar including information about several registers (Section 3).

2 Multiple grammars for multiple registers

As was pointed out in the introduction, speakers/hearers are able to use and detect various registers. This is reminiscent of multilingualism, and hence an obvious route to take is to have a look at multi-lingual grammar engineering projects with in the HPSG framework and their potential to be adapted to modeling register variation. First, there is the Grammar Matrix (Bender et al. 2002) which, however, produces grammars that are completely independent of each other and do not share any code. Second, there is the CoreGram project (Müller 2015) whose grammars are organized using sets of constraints. Theoretically at least, such sets could also be used to model register phenomena under a multiple-grammars approach.

CoreGram uses sets to represent constraints for various languages. For example, when grammars for German, Dutch and Danish are developed, all constraints applying to German and Dutch (Set G-D) are put in one set and all constraints applying to all three languages are put in one set (G-D-D). More specific sets include the more general sets. Hence, all the general constraints from set G-D-D are part of G-D and part of the sets for individual languages, e.g., the set G for German and D for Dutch. This approach can be applied to register phenomena. The most general constraint set would be Set 1 in Figure 1. There would be two subsets Set 2 and Set 3 for two different registers. These subsets include all constraints from Set 1, which is a general grammar of German.

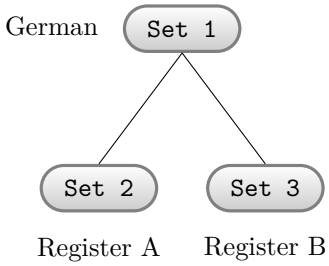

Figure 1: Modeling two registers of German

One could assume that certain words are only available in one of the two registers. For example, *Kohle* is a slang word for 'money' (such as *dough* in English) and one might assign it to Set 3, that is, Register B. Alternatively, one can have all words in Set 1, but add further constraints on the use of *Kohle* in Set 2 and Set 3. The occurrence of register-sensitive linguistic features is usually not a matter of all-or-nothing (see e. g. Biber & Conrad 2009, p. 53f.). Therefore, our approach –in line with the assumptions underlying our exploratory Bayesian analysis reported in Schäfer et al. (2022)– assumes different *probability distributions* of linguistic features in different registers. This can be captured in HPSG by attaching weights or probabilities to register-sensitive entities, including lexemes, inflectional and derivational lexical rules and syntactic schemata.

For instance, in a putative, rather informal register B (Set 3), the word *Kohle* 'money' could have higher probability than the word *Geld* 'money', and it could have a higher probability than the word *Kohle* 'money' in another, rather formal, register A (Set 2). When two or more linguistic objects are combined, the weight/probability of the mother is computed from the weight/probability of the daughters and the register value of the schema/rule that licenses the combination. Unfor-

tunately, the mathematics behind probabilistic HPSG is not completely worked out yet, but there are promising initial ideas (Brew 1995; Abney 1997; Miyao & Tsujii 2008; Guzmán Naranjo 2015).

The approach described above also works for syntactic phenomena, and we now provide an example of that. When we identified registers in our data-driven analysis, the complexity of constituents in clause initial position (in the so-called *Vorfeld*) turned out to be a good indicator of some registers containing mostly educated language. Since German is a Verb Second language, any constituent can occupy the Vorfeld, including full clauses. The syntax of German contains a Filler-Head Schema that is not restrictive as far as the filler daughter is concerned. The actual filler is determined by what is missing in the rest of the sentence. It can be a noun phrase in either of the four grammatical cases, a PP, an adverb, a non-finite verbal projection, or an adjectival projection. This is covered by the fact that the only constraint on the filler daughter that is specified in the Filler-Head Schema is that the LOCAL value of the filler has to match the element in SLASH.

(1) Filler-Head Schema according to Müller (2013: 169):

$$\textit{head-filler-phrase} \Rightarrow \begin{bmatrix} \text{NONLOC}|\text{SLASH} & \langle\rangle \\ \text{HEAD-DTR}|\text{SYNSEM} & \begin{bmatrix} \text{LOC}|\text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{verb} \\ \text{VFORM} \;\; \textit{fin} \\ \text{INITIAL} \;\; + \end{bmatrix} \\ \text{SUBCAT} & \langle\rangle \end{bmatrix} \\ \text{NONLOC}|\text{SLASH} & \langle\, \boxed{1} \,\rangle \end{bmatrix} \\ \text{NON-HEAD-DTRS} & \left\langle \begin{bmatrix} \text{SYNSEM} & \begin{bmatrix} \text{LOC} & \boxed{1} \\ \text{NONLOC}|\text{SLASH} & \langle\rangle \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

We can now use additional constraints on *head-filler-phrase* to encode register knowledge. Assume that we are currently analysing a sentence using a set of constraints that corresponds to a rather formal register (e. g. Set 2 above). If we see a *head-filler-phrase* with a filler-daughter that is a finite verbal projection (a clause), then we know that within the formal register at hand, its probability of occurrence is (relatively) high. Let us assume it is 0.05. In (2), we expand the feature geometry of signs by assuming a REGISTER attribute whose value specifies the type of register (as a value of TYPE) and its probability in this type of register (as a value of WEIGHT).

(2)
$$\begin{bmatrix} \textit{head-filler-phrase} \\ \text{NON-HEAD-DTRS} & \left\langle \begin{bmatrix} \text{SYNSEM}|\text{LOC}|\text{CAT}|\text{HEAD} & \begin{bmatrix} \textit{verb} \\ \text{VFORM} \;\; \textit{fin} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix} \Rightarrow \begin{bmatrix} \text{C-CONT}|\text{REG} & \begin{bmatrix} \text{TYPE} & \textit{2} \\ \text{WEIGHT} & \textit{0.05} \end{bmatrix} \end{bmatrix}$$

A set of constraints corresponding to a different register (e. g. Set 3 above) would contain a different version of this constraint, thus assigning a different probability to REGISTER.

Under this approach, each sentence can be analyzed relative to a particular register grammar (i. e. a particular probability distribution over register-relevant features). In addition to an analysis of the syntax and semantics of the sentence, the weight/probability of the topmost node can then be interpreted as the register score of that sentence, reflecting the probability (and therefore perhaps also the appropriateness) of that sentence with respect to that register.

The downside of this approach is that in order to compare the appropriateness of a sentence across different registers, it is necessary to parse the sentence once for each register, each time using the set of constraints corresponding to the respective register. From a psycholinguistic point of view, it seems rather implausible that humans parse a given sentence using a number of different grammars in parallel. However, under a model that assumes multiple grammars to model variation, we see no way around this. Therefore, we suggest a different approach using one grammar including all information about known registers.

3 One grammar with information about several registers

The alternative approach assumes that there is single grammar enriched with information about any sign's probability distribution across registers. For this purpose, we introduce a REGISTER feature next to PHON and SYNSEM on the outer level of the sign. In comparison to other register approaches in HPSG (cf. Paolillo 2000; Bender 2001; 2007), the values we propose for the REGISTER attribute do not say anything about social meaning, and are therefore not contained within CONTEXT. What our register approach provides is the probability of a sign in all operationalized registers (according to Schäfer et al. 2022). Up to this point, we are agnostic about whether or not a sign has social meaning and how it can be characterized (cf. "*not educated*" in Bender 2007 or "*correct*" in Paolillo 2000) and, in more complex phrases, combined. If present, this information may be stored as part of the CONTEXT attribute.

As in the approach sketched in the previous section, we assume that all signs bear information about registers, thus the REGISTER feature is appropriate for lexemes, for inflectional, and derivational lexical rules as well as for syntactic schemata. In contrast to the multiple-grammar approach, however, all signs carry information about all registers, not only about one particular register. Assuming (for instance) that there are seven registers, the architecture of a sign would look as follows:

(3)
$$\begin{bmatrix} \text{PHON} & \textit{list of phonemes} \\ \text{SYNSEM} & \textit{synsem} \\ \text{REGISTER} & \begin{bmatrix} \textit{reg} \\ \text{REGISTER1} & \textit{value} \\ \text{REGISTER2} & \textit{value} \\ \ldots \\ \text{REGISTER7} & \textit{value} \end{bmatrix} \end{bmatrix}$$

Similarly to the approach using multiple grammars, we need a way to determine the weights/ probabilities of the mother from the corresponding values of the daughters and of the schema/ rule that licenses the combination.[1] In either approach, these computations can be accomplished by a function *reg*. (In the full implementation, *reg* will be interpreted as a Bayesian update function adjusting the probabilities readers/hearers assign to the set of registers.) In contrast to the approach outlined in the previous section, a full representation of a sentence includes weights/ probabilities for each register. Register appropriateness can then be compared across different registers with one parse. For this advantage, the single-grammar approach appears as superior to the multiple-grammars approach in terms of cognitive plausibility.

4 Conclusion

In this paper, we have discussed multiple-grammar and single-grammar approaches to language-internal variation such as register in HPSG. We showed that an architecture similar to the CoreGram project can be adapted to the development of subgrammars encoding different registers of one language. Due to the probabilistic nature of register knowledge, probabilities of linguistic signs need to be specified in the subgrammars for each register. An alternative single-grammar approach was also sketched, where the discrete probability distributions over the set of registers are stored with each sign. We argued that the single-grammar approach is preferable because it allows us to evaluate the register properties of each sentence with a single parse instead of one parse per register. These fundamental considerations are part of the foundations for a planned long-term project wherein fine-grained register distinctions as discovered in our data-driven work (Schäfer et al. 2022) are implemented in a register-aware probabilistic HPSG.

[1] In other accounts dealing with register connected to social meaning (e.g. Paolillo 2000), the register information of the mother is computed by the set union of the register values of the daughters to see whether an utterance satisfies or not the felicity conditions of the register. In that sense, our approach –if combined with social meaning– can be seen as way to *quantify* to which extent the utterance satisfies the felicity conditions.

References

Abney, Steven P. 1997. Stochastic attribute-value grammars. *Computational Linguistics* 23(4). 597–618.

Adger, David. 2006. Combinatorial variability. *Journal of Linguistics* 42(3). 503–530. DOI: 10.1017/s002222670600418x.

Argamon, Shlomo Engelson. 2019. Register in computational language research. *Register Studies* 1(1). 100–135. DOI: 10.1075/rs.18015.arg.

Bender, Emily M. 2001. *Syntactic variation and linguistic competence: The case of AAVE copula absence*. Stanford University. (Doctoral dissertation).

Bender, Emily M. 2007. Socially meaningful syntactic variation in sign-based grammar. *English Language and Linguistics* 11(2). 347–381.

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk & Richard Sutcliffe (eds.), *COLING-GEE '02: Proceedings of the 2002 Workshop on Grammar Engineering and Evaluation*, 8–14. Taipei, Taiwan: Association for Computational Linguistics. DOI: 10.3115/1118783.1118785.

Biber, Douglas. 1988. *Variation across speech and writing* (none). Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511621024.

Biber, Douglas. 1995. *Dimensions of register variation* (none). Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511519871.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511814358.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3. 993–1022. DOI: 10.1145/1015330.1015439.

Brew, Chris. 1995. Stochastic HPSG. In Steven P. Abney & Erhard W. Hinrichs (eds.), *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, 83–89. Dublin: Association for Computational Linguistics.

Guzmán Naranjo, Matías. 2015. Unifying everything: Integrating quantitative effects into formal models of grammar. In Johannes Wahle, Marisa Köllner, Harald Baayen, Gerhard Jäger & Tineke Baayen-Oudshoorn (eds.), *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, none. Tübingen: Tübingen University. DOI: 10.15496/publikation-8636.

Hilpert, Martin. 2013. *Constructional change in English. Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139004206.

Miyao, Yusuke & Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics* 34(1). 35–80. DOI: 10.1162/coli.2008.34.1.35.

Müller, Stefan. 2013. *Head-Driven Phrase Structure Grammar: Eine Einführung*. 3rd edn. (Stauffenburg Einführungen 17). Tübingen: Stauffenburg Verlag.

Müller, Stefan. 2015. The CoreGram project: Theoretical linguistics, theory development and verification. *Journal of Language Modelling* 3(1). 21–86. DOI: 10.15398/jlm.v3i1.91.

Paolillo, John C. 2000. Formalizing formality: an analysis of register variation in Sinhala. *Journal of Linguistics* 36(2). 215–259.

Pierrehumbert, Janet B. 2008. Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology 7*, vol. 4–1 (Phonology and Phonetics), 101–139. de Gruyter. DOI: 10.1515/9783110197105.1.101.

Schäfer, Roland, Felix Bildhauer, Elizabeth Pankratz & Stefan Müller. 2022. Modelling registers. Ms. Humboldt-Universität zu Berlin.

Søgaard, Anders & Petter Haugereid. 2007. A tractable typed feature structure grammar for Mainland Scandinavian. *Nordic Journal of Linguistics* 30(1). 87–128. DOI: 10.1017/S0332586507001667.

Solt, Stephanie. 2015. Vagueness and imprecision: Empirical foundations. *Annual Review of Linguistics* 1(1). 107–127. DOI: 10.1146/annurev-linguist-030514-125150.

Yang, Charles. 2002. *Knowledge and learning in natural language*. Oxford University Press. DOI: 10.1007/978-1-4419-1428-6_837.