Modelling Register Variation in HPSG

Stefan Müller & Antonio Machicao y Priemer

HU Berlin, Institute for German Language and Linguistics, Syntax Lab

1. Preliminary assumptions

- **Grammatical competence** (GC): Knowledge about *how to combine language specific elements* (word, phrases, etc.) to build form-meaning pairs (in any situation) → **possibility**
- **Register competence** (RC): Knowledge about *in which situations* form-meaning pairs can be used and how in these situations form maps to meaning (e.g. imprecision, metaphors, irony, etc.) → **probability**
- GC and RC are not (always) **acquired** at the same time:
  In many cases, children know *how* to build a grammatical utterance, but they do not know (yet) *when* to use it.
- RC is related to the acquisition of **social skills**, but it is closely **intertwined with GC** at all levels of grammatical description (cf. (1)–(4)).

(1) Phonology:
  a. ich habe …
  b. ich hab …
  'I have …'

(2) Lexical selection:
  a. Herr XY
  b. Digga
  'Mr XY'

(3) Syntax:
  a. Das Kind hat aber nicht mal geweint!
  b. Hat aber nicht mal geweint das Kind! [Facebook]
  'The child has not even cried.'

(4) Semantics: (cf. Solt 2015: 111)
  a. I arrived at 3:53 pm.
  b. I arrived at 4.

- Furthermore, RC is related to **extra grammatical information** (5), to phenomena **combining** different **grammatical levels** (6),

(5) Speakers intentions (pragmatics)
  a. e.g. intentionally violating Grice's maxims to convey information
  b. e.g. using a dialect to convey familiarity

(6) Interface: Phonology–Morphology–Syntax–Information Structure:
  a. Wir haben das gekauft.
  b. Hamwa gekauft.
  'We have bought it.'

2. Which kind of model do we aim for?

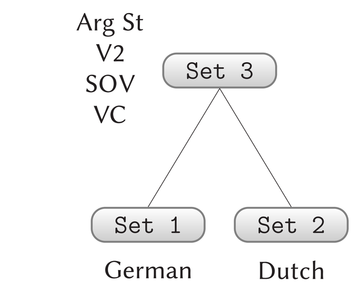
We use the HPSG formalism since … (Pollard & Sag 1994; Müller 2013)

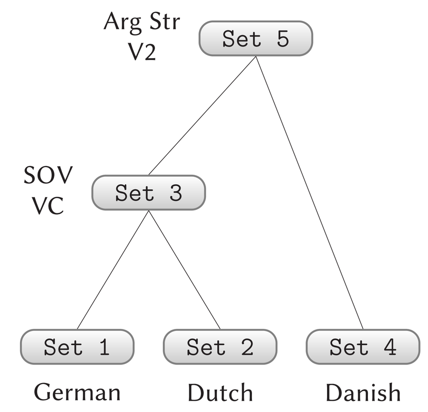
1. it allows us to combine different levels of linguistic description
2. it can integrate extra-grammatical levels of language description to account for the different phenomena related to register variation (cf. Paolillo 2000; Bender 2001; Bildhauer 2007)
3. it can be implemented as a computer processable grammatical model (cf. Müller 2014, 2015)
4. it is a performance-compatible competence grammar (cf. Sag & Wasow 2011; contra Newmeyer 2003; Adger 2006)
5. it can be expanded by probabilistic methods (cf. Brew 1995; Abney 1997; Miyao et al. 2008)
6. it can be exapanded to domains bigger than a sentence (e.g. texts)
7. The model must be able to predict how probable a linguistic sign (e.g. word, sentence, text) is in a specific register (register comparison).
8. The formalism has to allow the combination of REGISTER values of lexemes, phonological or syntactic constraints, etc. to compute the REGISTER value of the linguistic sign at different levels of complexity.

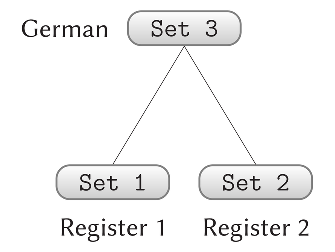
3. Multiple grammars

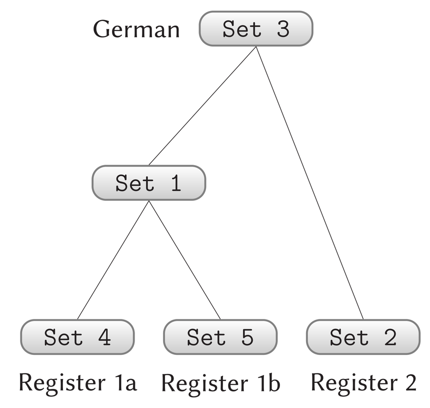
Following the conceptualisation in multilingual grammar engineering (cf. Bender et al. 2002; Müller 2015), one possible way to model registers is assuming a single set of constraints (viz. a grammar) for each register.

Here, we follow the idea of the CoreGram project (Müller 2015) where **grammars are organised as sets of constraints.** Assuming only two grammars (German and Dutch), we will find shared constraints (e.g. w.r.t. argument structure, V2, SOV, etc.). These constraints are organised in a set (Set 3) that is shared by the grammars of both languages. Adding a further language (e.g. Danish), we can organise the constraints shared between Danish and the other two languages in a more general set (Set 5).









In the same way, registers can be organised as sets containing the **constraints for the use of signs** in specific registers. Building on work from the 1st phase of the SFB (cf. Schäfer et al. 2022), we assume different *probability distributions* for linguistic signs in different registers. That is, the GC is constrained in Set 3, but the probabilities for the constraints to be realised in one register or the other are organised in different sets. The probabilities are represented by WEIGHT values that can be assigned to affixes, words, constraints, etc. and by means of register constraints can be **computed together** reflecting the probability of an utterance in a register.

4. One grammar

In an alternative approach, we assume a single grammar. The **feature geometry** of signs is **enriched** by a REGISTER feature, whose value is an AVM with feature-value pairs for every single register. In contrast to other approaches (cf. Bender 2001; Paolillo 2000), the register feature is not inside the pragmatics of the sign, but at the outer level of the sign, since it can **interact with all levels.** In contrast to the multiple grammars approach, all signs carry information about all registers, having the advantage that **register appropriateness** can be **compared** across registers with one parse (in contrast to the multiple grammars approach).

PHON	*list of phonemes*	
SYNSEM	*synsem*	
REGISTER	*reg*	
	REG1	*value*
	REG2	*value*
	…	
	REG7	*value*

Conclusions & open questions

- Cooperation with other projects in order to identify register related phenomena and provide adequate formal analyses for them.
- Discussions about possible models and what they have to perform are needed.
- Mathematics behind probabilistic HPSG has to be worked out in detail.
- Assuming a specific situation and knowledge about the register for such situation: How can we recognise **register violations** and what do they mean (change of grammar vs. inappropriate use)?

(7) Mit Verlaub, Herr Präsident, Sie sind ein Arschloch. [Joscha Fischer, 1984]
  'With respect, Mr. President, you are an asshole.'

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.

Bender, Emily. 2001. *Syntactic variation and linguistic competence: The case of AAVE copula abscence*: Stanford University PhD thesis. `https://faculty.washington.edu/ebender/dissertation/bender_thesis.pdf`.

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.

Bildhauer, Felix. 2007. *Representing information structure in an HPSG grammar of Spanish*: Universität Bremen Dissertation. `http://hpsg.fu-berlin.de/~fbildhau/diss/felix-bildhauer-diss.pdf`.

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.

Miyao, Yusuke, Rune Sætre, Kenji Sagae, Takuya Matsuzaki & Jun'ichi Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In Johanna D. Moore, Simone Teufel, James Allan & Sadaoki Furui (eds.), *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 46–54. Columbus, OH: Association for Computational Linguistics. `http://aclweb.org/anthology/P08-1006`.

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

Müller, Stefan. 2014. Kernigkeit: Anmerkungen zur Kern-Peripherie-Unterscheidung. In Antonio Machicao y Priemer, Andreas Nolda & Athina Sioupi (eds.), *Zwischen Kern und Peripherie: Untersuchungen zu Randbereichen in Sprache und Grammatik*, 25–39. Berlin: De Gruyter. doi:10.1524/9783050065335.25.

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.

Müller, Stefan & Antonio Machicao y Priemer. 2019. Head-Driven Phrase Structure Grammar. In András Kertész, Edith Moravcsik & Csilla Rákosi (eds.), *Current approaches to syntax – A comparative handbook* (Comparative Handbooks of Linguistics 3), Berlin: De Gruyter Mouton. doi:10.1515/9783110540253-012.

Newmeyer, Frederick J. 2003. Grammar is grammar and usage is usage. *Language* 79(4). 682–707.

Paolillo, John C. 2000. Formalizing formality: An analysis of register variation in Sinhala. *Journal of Linguistics* 36(2). 215–259. doi:10.1017/S0022226700008148.

Pollard, Carl J. & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Sag, Ivan A. & Thomas Wasow. 2011. Performance-compatible competence grammar. In Robert D. Borsley & Kersti Börjars (eds.), *Non-transformational syntax: Formal and explicit models of grammar*, 359–377. Oxford: Wiley-Blackwell. doi:10.1002/9781444395037.ch10.

Schäfer, Roland, Felix Bildhauer, Elizabeth Pankratz & Stefan Müller. 2022. Modelling registers.

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