

# Implementing dialectal variation in typed feature structure grammars

Anders Søgaard  
Center for Language Technology  
University of Copenhagen  
Njalsgade 80  
DK-2300 Copenhagen S, Denmark  
soegaard@islands-brygge.dk

Petter Haugereid  
Department of Linguistics  
Norwegian University of Science and Technology  
Dragvoll  
N-7491 Trondheim, Norway  
petterha@hf.ntnu.no

**Abstract**

Typed feature structure grammars contain the formal machinery to merge and implement macro- and microperspectives of comparative linguistics, if languages and dialects are seen as a set of types  $\{\tau_1, \dots, \tau_i\}$ , and if  $\{\tau_1, \dots, \tau_i\}$  are used to appropriately cross-classify the set of empirically attested linguistic types, say  $\{\tau_j, \dots, \tau_n\}$ , i.e. such that some (linguistic) type  $\tau_k$  is compatible with a subset of languages, but incompatible with others. The cross-classification constrains the possible design of linguistic grammars in important ways. From a theoretical perspective, it is interesting that the type hierarchy of a grammar  $\Gamma_\Lambda$  indirectly represents the genealogy and typology of the language  $\Lambda$ , but from a more practical point of view, it is also interesting to note how this grammar design facilitates easy grammar engineering for related languages in the long run, e.g. if you have a grammar for language  $\Lambda_i$  and want to write a grammar for  $\Lambda_j$ , you only have to add the types in  $\Lambda_j - \Lambda_i$ , if the types of  $\Lambda_i \cup \Lambda_j$  are compatible with both language types  $\tau_{\Lambda_i}$  and  $\tau_{\Lambda_j}$ . This also reduces redundancies in various multilingual applications, e.g. machine translation systems.

## 1 Introduction

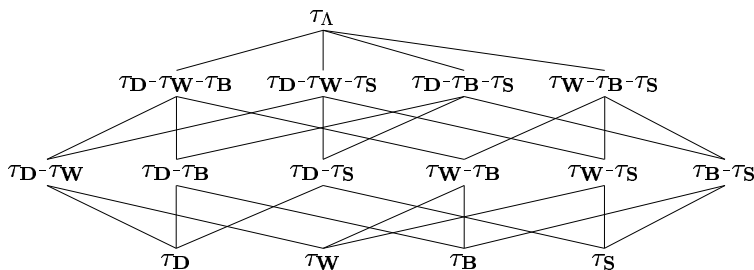
The LKB system (Copestake, 1999) is an environment for constructing and evaluating bidirectional grammars, which support the Head-Driven Phrase Structure Grammar formalism (Sag and Wasow, 1999). The formalism relies on the logic of typed feature structure grammars (Carpenter, 1992; Copestake, 2000), a logic of some complexity and to discuss it in any detail would distort the focus of the paper. The reader is instead left with the references above and a short, informal introduction:

- (i) A type hierarchy can be loosely defined as a tuple  $\langle \top, \sqsubseteq, \{\nu_1 \dots \nu_n\}, \perp \rangle$ , where  $\top \sqsubseteq \nu_i$  and  $\nu_i \sqsubseteq \perp$ .
- (ii) The fundamental notion in the logic is the join operation ( $\bowtie$ ), which finds the greatest lower bound of two nodes, say  $\nu_i$  and  $\nu_j$ . Df.  $\nu_i \bowtie \nu_j = \nu_i$  and  $\nu_i \sqsubseteq \nu_i \bowtie \nu_j$ .
- (iii) If two types share no greatest lower bound, i.e.  $\nu_i \bowtie \nu_j = \perp$ , they are said to be incompatible.

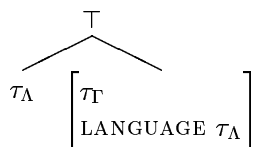
Subsumption on feature structures is similar to standard unification; cf. (ii). A phrase structure grammar (a set of rewriting rules and a lexicon) is defined on  $\langle \top, \sqsubseteq, \{\nu_1 \dots \nu_n\}, \perp \rangle$ .

How is a multilingual grammar constructed in this framework? Let a set of nodes  $\{\nu_1, \dots, \nu_i\}$  be a set of language or dialect types  $\{\tau_1, \dots, \tau_i\}$  and another set of nodes  $\{\nu_j, \dots, \nu_n\}$  correspond to the linguistic types  $\{\tau_j, \dots, \tau_n\}$  of those languages and dialects, where  $\{\tau_1, \dots, \tau_i\} \cap \{\tau_j, \dots, \tau_n\} = \emptyset$ . It is now possible to state that a certain construction or lexical item  $\tau_k$  is accepted in some language  $\tau_h$ , if and only if  $\tau_k \bowtie \tau_h \neq \perp$ .

In this paper, we confine ourselves to a small set of language and dialect types  $\{\tau_{\mathbf{D}}, \tau_{\mathbf{W}}, \tau_{\mathbf{B}}, \tau_{\mathbf{S}}\}$ , where  $\tau_{\mathbf{D}}$  denotes Standard Danish,  $\tau_{\mathbf{W}}$  denotes Western Jutlandic (a dialect of Standard Danish),  $\tau_{\mathbf{B}}$  denotes Norwegian Bokmål, while  $\tau_{\mathbf{S}}$  denotes Standard Swedish. The types  $\{\tau_{\mathbf{D}}, \tau_{\mathbf{W}}, \tau_{\mathbf{B}}, \tau_{\mathbf{S}}\}$  are implemented in the following type hierarchy under  $\tau_{\Lambda}$  (the supertype):



The basic intuition is clear: The hierarchy allows a linguistic type to be compatible with any subset of  $\{\tau_D, \tau_W, \tau_B, \tau_S\}$ . To minimize the number of rewriting rule,  $\tau_\Lambda$  is given as a restriction on the possible values of a LANGUAGE feature. The alternative would be to implement actual subtypes of  $\tau_\Lambda$  and  $\{\tau_j, \dots, \tau_n\}$  (the set of linguistic types). The final hierarchy has this basic form, where  $\tau_\Gamma : \{\tau_j, \dots, \tau_n\}$ :



## 2 Examples

This section exemplifies the use of the LANGUAGE (or dialect) feature. Since syntactico-semantic types, morphological types and lexical types are implemented in slightly different ways in the LKB system, one example is given for each grammar module. The implementation of possessive constructions in our language sample - i.e. Standard Danish, Western Jutlandic, Norwegian Bokmål and Standard Swedish - will illustrate the use of the LANGUAGE feature in the syntactico-semantic module. The second and third subsections are devoted to, respectively, the implementation of inflectional paradigms and the construction of a multilingual lexicon.

### 2.1 Possessives

Standard Danish, Norwegian Bokmål and Standard Swedish all have a genitival possessive. In the framework described above, this translates into saying that the genitival possessive has a feature structure of the following form:

$$\left[ \begin{array}{l} \textit{genitival-possessive} \\ \text{LANGUAGE } \tau_D\text{-}\tau_B\text{-}\tau_S \\ \dots \end{array} \right]$$

In addition, Norwegian Bokmål has a standard prepositional possessive and the *sin*-possessive, exemplified by (1) and (2):

- (1) *gutten til speiderlederen seiler*  
 boy-DEF to-PREP scout leader-DEF sail-PRES  
 'the scout leader's boy sails'
- (2) *speiderlederen sin gutt seiler*  
 scout leader-DEF PRON-POSS-REFL boy-INDEF sail-PRES  
 'the scout leader's boy sails'

If the standard prepositional possessive is said to be a separate construction, this construction must be marked to be acceptable only in Norwegian Bokmål, i.e. the value of the LANGUAGE feature is  $\tau_B$ . Similarly for the *sin*-possessive.

In Western Jutlandic, a fourth possessive construction is employed:

- (3) *æ pig hend hus*  
 the girl PRON-POSS house-INDEF  
 'the girl's house'

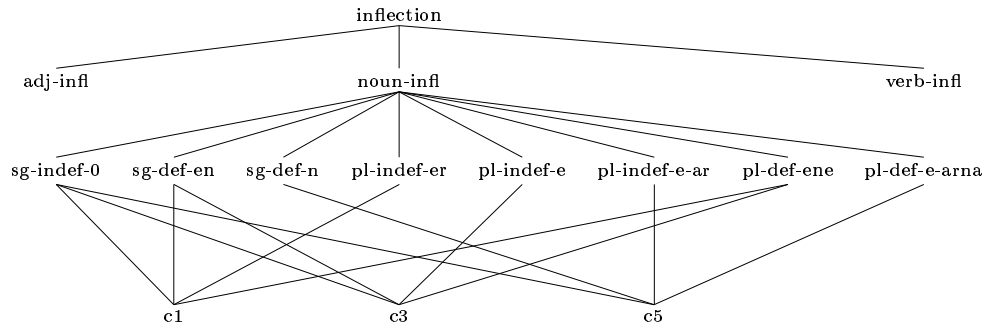
(3) differ syntactically from (2) in that the pronoun is non-reflexive. This short list does not exhaust the number of possessive constructions in Mainland Scandinavian, e.g. Western Bottnian (a dialect spoken in Northern Sweden) has postnominal genitival possessives and even uninflected ones. In addition, the pronominal possessives have been completely ignored.

In sum, there are a number of different possessive constructions in Mainland Scandinavian, some of which are shared by subsets of languages and dialects. The  $\tau_A$  hierarchy allows for sound cross-linguistic generalizations. The basic mechanism is provided by the LANGUAGE feature. Of course, constructions that are only employed by one dialect are dialect-specific; but this does not mean, however, that these types cannot inherit from more common types. In fact, if we adopt standard Head-Driven Phrase Structure Grammar theory, any of the possessive constructions will inherit (at least) from some binary and headed phrasal type. Søggaard and Haugereid (2004) claim that all the possessives (plus a number of related constructions) are instances of the same construction. On this design, the distribution of possessives wrt. languages and dialects becomes an entirely lexical matter.

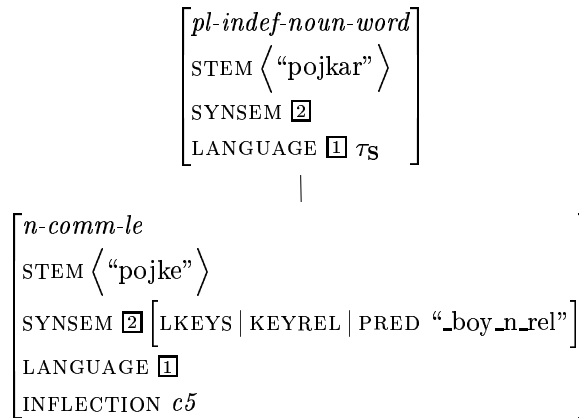
## 2.2 Inflectional paradigms

Open class items are listed in the lexicon as lexemes, while closed class items are listed as words. Each lexeme is specified to be compatible with some inflectional paradigm. The types *c1*, *c3* and *c5* denote such paradigms and constrain the number of inflectional rules that may apply to the lexeme, e.g. the Danish lexeme *dreng* is constrained by the inflectional paradigm *c3* (implemented as the value of its INFLECTION feature), which is compatible with *sg-indef-0*, *sg-def-en*, *pl-indef-e* and *pl-def-ene*. Similarly, Norwegian *gutt* is constrained by *c1*, while Swedish *pojke* is constrained by *c5* (Western Jutlandic does not have definite inflection at all). These constraints provide the following inflections:

SG-INDEF	SG-DEF	PL-INDEF	PL-DEF
dreng	drengen	dreng	drengene
gutt	gutten	gutter	guttene
pojke	pojken	pojkar	pojkane



The name *sg-indef-0* is meant to indicate that no inflection is added to obtain the singular indefinite, etc. The parse tree below illustrates how lexemes undergo inflection:



Some inflectional rules are only found in a subset of  $\{\tau_D, \tau_W, \tau_B, \tau_S\}$ , e.g. *pl-indef-e-ar* is only found in Standard Swedish (in our subset of Scandinavian languages and dialects, that is, the inflectional rule is also found in Nynorsk). Consequently, inflectional rules are restricted wrt. language and dialect. Of course this implies that inflectional paradigms only consist of rules, the LANGUAGE values of which have a greatest lower bound  $\gamma$ , where  $\gamma \neq \perp$ .

### 2.3 Lexical entries

In the case of open class items, lexical entries contain information about which lexeme type the lexemes inherit their most general constraints from. The entries also

specify the phonology, the predicate value of the elementary predicate (and additional lexical semantics) and the inflectional paradigm. These are standard constraints on lexical entries. In addition, they are specified wrt. language or dialect. Consider, for illustration:

$$\left[ \begin{array}{l} n\text{-comm-le} \\ \text{STEM} \langle \text{"dreng"} \rangle \\ \text{SS|LKEYS|KEYREL|PRED} \text{"_boy_n_rel"} \\ \text{LANGUAGE } \tau_{\mathbf{D}} \\ \text{INFLECTION } c\mathcal{B} \end{array} \right]$$

If we conceive of the multilingual lexicon as multiple lexicons in one, the lexical entry for the noun *dreng* specifies the noun to belong to the Standard Danish part of the lexicon. Western Jutlandic has a similar noun *dreng* with similar semantics. In the multilingual lexicon, however, these are split into two entries, since they belong to different inflectional paradigms. In principle, you could underspecify the lexical entry wrt. inflectional paradigm, but this would complicate composition somewhat, i.e. supertypes would have to be introduced on inflectional paradigms in a rather *ad hoc* way.

### 3 Conclusion

Head-Driven Phrase Structure Grammar (and any other formalism conforming to the logic in Carpenter, 1992) is expressible enough to implement macro- and micro-perspectives of comparative linguistics. This is theoretically interesting, since it allows multilingual and typologically adequate grammar engineering. It also means that matrices specifying the common types for sets of languages (in the sense of Bender et al., 2002) can function as start kits for the linguist who needs to write a language-specific grammar for some monolingual application. For multilingual applications, the cross-classification of linguistic types with language or dialect types reduces major redundancies.

## 4 References

- Bender, Emily, 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. *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Taipei, Taiwan.
- Carpenter, Bob. 1992. *The logic of typed feature structures*. Cambridge: Cambridge University Press.
- Copestake, Ann. 1999. *The (new) LKB system*. Stanford: CSLI Publications.
- Copestake, Ann. 2000. Appendix: definitions of typed feature structures. *Natural Language Engineering* 1.1, 1-4.
- Sag, Ivan & Tom Wasow. 1999. *Syntactic theory*. Stanford: CSLI Publications.
- Søgaard, Anders & Petter Haugereid. 2004. A brief documentation of a computational HPSG grammar specifying (most of) the common subset of linguistic types for Danish, Norwegian and Swedish. Submitted to H. Holmboe (ed.), *Nordisk Sprogteknologi 2004*. Copenhagen: Museum Tusulanum.