NorSource (Norwegian HPSG Resource Grammar) – a brief description

Lars Hellan, NTNU

March 22, 2015

The computational grammar *NorSource* of Norwegian, developed at NTNU since 2002, is an implemented HPSG ('Head-Driven Phrase Structure Grammar' – cf. (Pollard and Sag 1994)) grammar based on the development platform LKB (Copestake 2002). It is a parser for Norwegian (bokmål) text, with information provided in each parse about the morpho-syntactic and semantic structure of the sentence. It combines the roles of a *knowledge system* and a *processing tool*.

HPSG grammars, and LKB grammars in particular, are type-based, and declarative. NorSource uses the architecture of the 'HPSG Grammar Matrix' (Bender 2010, 2012), which underlies a family of current LKB grammars; it includes the semantic representation formalism Minimal Recursion Semantics (MRS; Copestake et al. 2005), which accompanies any parse produced by the grammar, so that from such a representation, sentences of the language can be generated (such a grammar is thus both 'analyzing' and 'generating'). MRS representations to some extent resemble predicate logic formulas, the figure below showing a standard MRS format for the sentence "Gutten kaster ballen", where each formative is represented by a so-called elementary predication (EP):

$$\begin{bmatrix} \text{LTOP}:\text{h1} \\ \text{INDEX e2}\begin{bmatrix}\text{E}\begin{bmatrix}\text{TENSE}:\text{PRES}\end{bmatrix}\end{bmatrix} \\ \\ \text{RELS}:\left\langle \begin{bmatrix}\_\text{def}\_\text{q}\_\text{rel} \\ \text{LBL h5} \\ \text{ARG0 x4} \\ \text{RSTR h6} \\ \text{BODY h7}\end{bmatrix}, \begin{bmatrix}\_\text{gutt}\_\text{n}\_\text{rel} \\ \text{LBL h3} \\ \text{ARG0 x4}\end{bmatrix}, \begin{bmatrix}\_\text{kaste}\_\text{v}\_\text{rel} \\ \text{LBL h8} \\ \text{ARG0 e2} \\ \text{ARG1 x4} \\ \text{ARG2 x9}\end{bmatrix}, \begin{bmatrix}\_\text{def}\_\text{q}\_\text{rel} \\ \text{LBL h11} \\ \text{ARG0 x9} \\ \text{RSTR h12} \\ \text{BODY h13}\end{bmatrix}, \begin{bmatrix}\_\text{ball}\_\text{n}\_\text{rel} \\ \text{LBL h10} \\ \text{ARG0 x9}\end{bmatrix} \right\rangle \\ \\ \text{HCONS}:\langle \text{h6 QEQ h3, h12 QEQ h10}\rangle \end{bmatrix}$$

Figure 1. MRS representation of "Gutten kaster ballen".

The term 'LKB' partly refers to the computational platform so-called, partly to the version of the general HPSG architecture implemented in this platform. In the former sense, the platform, based on lisp, is especially suited for grammar development, with a well-developed interface for exploration and feedback. For processing purposes, it has been supplemented with a much faster parsing system PET (Callmeyer 2002), based on C++, and from 2010 on with ACE,[1] serving for both parsing and generation, and increasing speed by still a significant factor. For its functioning in applications (see below), NorSource until 2011 used LKB and PET, and from 2011 is using ACE as its main processing system. For grammar development, LKB is still the system used.

A demo of the grammar is accessible through
http://www.typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource,
and directly at
http://regdili.hf.ntnu.no:8081/linguisticAce/parse.

The grammar currently is integrated in the following applications and resources:

- A *Norwegian Grammar Sparrer* (see Hellan et al. 2013, and http://regdili.hf.ntnu.no:8081/studentAce/parse). This is an interactive online 'Grammar Tutor' with functionalities as follows (cf. http://typecraft.org/tc2wiki/Classroom:Norwegian_Grammar_Checking). A user writes a putative Norwegian sentence into a window; if grammatical, the system responds that the sentence is

[1] http://moin.delph-in.net/PetTop, http://moin.delph-in.net/AceTop

grammatical, while if ungrammatical, the system informs the user in what respect the string is ungrammatical. For instance, for the ungrammatical string “Mannet smiler”, one gets the feedback “The word “mannet” is of masculine gender, not neuter”. In addition to the error message, the interface window window provides three buttons. *Info* takes one to a detailed instruction about the tool, and *More description* takes one to succinct information about the relevant aspect of Norwegian grammar;[2] by pushing *Generate*, one can get an example of how the intended sentence should be written, viz. “Mannen smiler”.

- An in depth *Multilingual Valence* repository, with aligned valence information for verbs in Norwegian, Spanish, Bulgarian and Ga. (See Hellan et al. 2014, and http://regdili.hf.ntnu.no:8081/multilanguage_valence_demo/multivalence and http://typecraft.org/tc2wiki/Multilingual_Verb_Valence_Lexicon.)
- A POS-tagger, under development, reflecting the lexical inventory of the grammar, useful for lexical acquisition from new text (http://regdili.hf.ntnu.no:8081/webtagger/tagger).

The grammar consists mainly of files for types, rules and lexicons, as follows.
There are four *type* files:

- matrix.tdl, with the types defined in the version 0.6 of the Grammar Matrix (from 2002), with many amendments introduced later;
- norsk.tdl, the main file (about 27 000 types);
- predsort.tdl, devoted mainly to types for spatio-temporal expressions going beyond the most basic combinatorial needs;
- lex-types-v.tdl, encoding a conversion from the 'old' verb types defined in norsk.tdl from 2002 up to 2008, to the 'Construction Labeling' types (cf. Hellan 2008, Hellan and Dakubu 2010).

There are many small size *lexicon* files:

- lex1.close.tdl contains all closed class items except spatio-temporal prepositions and adverbs, and representatives of all subtypes of open class items, including all verb types (1050 entries);
- lex2.open.tdl contains open class items supporting the test files (see below);
- lex3.p-adv-full.tdl contains all spatiotemporal prepositions and adverbs;
- lex4.propn1.tdl and lex4.propn2.tdl contain proper names supporting the test files (lex4.propn2.tdl devoted exclusively to the test file hike.no);
- lex2.semlab.tdl has words with types adapted to special semantic specifications of more experimental nature.

There are also four large lexicon files, for verbs, adjectives and nouns; three of those are amendments of material from *NorKompLeks* (see below), with new inflectional and grammatical information, and one from the project *TROLL* (see below):

- lex4.lrg-v.tdl (about 10 000 entries);
- lex4.lrg-a.tdl (about 10 000 entries);
- lex4.lrg-n.tdl (about 50 000 entries);
- lex3.v-troll.tdl (about 3 000 entries).

Inflectional rules (irules) come in three files, for verbs, adjectives and nouns:

- irules-v.tdl;
- irules-a.tdl;
- irules-n.tdl.

They are supplemented with files for irregular patterns relative to those stated in the files mentioned;

- irregs.tab

Derivational rules (‘lexical rules’) are assembled in the file

[2] In the relevant case, http://typecraft.org/tc2wiki/The_Noun_Phrase_-_Norwegian.

- lrules.tdl.

Phrasal rules (‘syntactic rules’) are assembled in the file

- rules.tdl (about 250 rules actively used).

Systematic test suites, with translations to English, are found in the directory /Tests(combined in the suite ‘massifcentral’):

- test-v-stnd represents all frames corresponding to those defined in lex-types-v.tdl;
- test-np represents most configurations inside NPs;
- test-p represents all uses of spatiotemporal prepositions and adverbs, with comments;
- test-dir represents more complex spatio-temporal constructions, and exemplifies calculation of aspect (for instance, in the web demo MRS seen on the path ‘...INDEX.PATH-TELIC …’);
- test-cmpar represents constructions with comparatives and other degree specifications;
- test-clause represents constructions at clausal level, including wh-dependencies (topicalization, relativization, constituent questions), adverbial distribution, pronominal distribution, passive, reflexives, coordination, punctuation in many environments, apposition, free predicatives, derivational morphology;
- mrs-suite, a counterpart to the same-named suite developed for English.

There are also corpus-based test-suites representing domains of analysis addressed throughout.

In addition to these grammar parts, the Grammar Sparrer resides in a specific ‘sub-grammar’ of types and rules for error representation, in a folder called *MalGram*.

The grammar since its start in 2002 has gone through the following stages:

Phase 1, the *Grounding* phase (2001-04),
Phase 2, the *Semantic Expansion* phase (2005-07),
Phase 3, the *Cross-Linguistic Coding* phase (2008-10), and
Phase 4, the *Interoperability* phase (2010-).

Phase 1 resided in the building of a basic core grammar around the Matrix skeleton (using the Matrix versions 0.1 – 0.6, as they developed; this included the MRS system). This stage included the accommodation of a 80,000 entries lexicon imported from the previously established resources TROLL and NorKompLex[3], where a verb valence code and a code for inflectional paradigms constituted major parts. Main publications from this period are: Hellan and Haugereid 2002, Hellan 2003.

Phase 2 resided in the development of a fine-grained ontology and computing system of spatial and temporal relations, amenable to grammatical systems across languages and typologies, and a detailed semantics of comparative constructions. The grammar was also used as a part of a small Norwegian-Japanese MT system. In this period, the inflectional system was thoroughly revised. Main publications from this period are: Hellan and Beermann (2004), Bermann et al. (2004), Beermann and Hellan (2005).

Phase 3 was devoted to a thorough revision of the valence code, to accommodate a cross-linguistically defined classification system of valence and construction types. Main publications from this period are: Hellan (2008), Hellan and Dakubu (2010)

Phase 4 has resided in the development of the applications mentioned above.

[3] A system built on TROLL and Bokmålsordboka, conducted throughout 1995-2000. Cf. Nordgård 1998.

Background

In addition to the general technical and theoretical background described above, the construction of NorSource has benefitted significantly from previously existing lexical resources, in particular addressing valence. Large coverage Norwegian valence repositories include *TROLL*[4] and *NorKompLex*[5], both existing as text files. In the lexicon project *TROLL*, one combined the following structures:[6]

1. A set of 27 'basic' verb lexeme types, each with a label and an analytic description standing in a 1-to-1-relation to the label: <1, iv>, <2, erg>, <3, exp-iv>, <4, tv>, <5, th-tv>, <6, exp-tv>, …, <27, erg-ditv>. These can be seen as 'verb classes' in the sense later introduced by Levin 1993.
2. A set of 'derivational' and frame alternation rules, producing 'derived' verb lexeme types, each with a label and analytic description as in 1 (e.g., 'object deletion', 'small clause formation', …), bringing the total number of types accounted for in the system to about 150.
3. A dictionary of about 1000 verb lexemes, each ascribed a basic type.

In the subsequent project *NorKompLex*, about 100 of these types were selected in a 'flat' type system, not distinguishing between basic or derived types.[7] With many lexemes thus appearing in multiple entries, and a large import of verbs from *Bokmålsordboka*, the verb part of the NorKomplex dictionary counts around 10 000 verb entries. The NorKompLex dictionary (completed for all parts of speech, inflection codes, lemmas, and with phonetic transcription of all inflrcted forms) has been a vital resource both in commercial and academic applications. For instance, it is an essential basis for *Norsk Ordbank* which has an online search interface for all word classes,[8] however not with valence as a search criterion.[9] It has also served in the build-up of the lexicon of two computational grammars, NorSource as mentioned, and the LFG grammar *NorGram*.[10]

References (1)

Bender, E.M., Scott Drellishak, Antske Fokkens, Laura Poulson. and S. Saleem. 2010. Grammar Customization. In *Research on Language and Computation*, 8(1), 23-72.

Bender, E.M., Sumukh Ghodke, Timothy Baldwin and Rebecca Dridan. 2012. From Database to Treebank: On Enhancing Hypertext Grammars with Grammar Engineering and Treebank Search. In Sebastian Nordhoff (ed) *Electronic Grammaticography*, pp 179-206.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars.* CSLI Publications, Stanford.

Copestake, A., D. Flickinger, C. Pollard, and I. Sag. (2005). Minimal Recursion Semantics: an Introduction. In Research on Language and Computation 3(4), 281-332.

Levin, B. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago,IL.

Nordgård, T.(1998) "Norwegian Computational Lexicon (NorKompLeks)". Proceedings of the 11th Nordic Conference of Computational Linguistics NODALIDA 98. CST, Copenhagen.

References (2, with L. Hellan as (co-)author)

2014, with D. Beermann, T. Bruland, M.E.K. Dakubu, and M. Marimon (2014) MultiVal: Towards a multilingual valence lexicon**.** *LREC 2014.*

2013 with Tore Bruland, Elias Aamot, Mads H. Sandøy: A Grammar Sparrer for Norwegian. Proceedings of *NoDaLiDa* 2013.

2012: with D. Beermann. Semantics of Spatial Prepositions in the Grammar NorSource. Paper presented at 'Meaning of P', Ruhr-Universität Bochum, Nov 2012.

[4] Cf. Hellan et al. 1989.

[5] Cf. Nordgård 1998.

[6] In terms of semantic specification, both TROLL and NorKompLex use 9 semantic roles.

[7] Of course keeping a process like 'passive' still as a potentiality feature. The verb type labels here have the form intrans1, intrans2, intrans3, …. .

[8] http://www.edd.uio.no/perl/search/search.cgi?appid=72&tabid=1106

[9] As opposed to the Valence repository mentioned above, based on the information in NorSource.

[10] http://clarino.uib.no/iness/xle-web

2010: with Dakubu, M.E.K. (2010). *Identifying Verb Constructions Cross-linguistically*. SLAVOB series 6.3, Univ. of Ghana (http://www.typecraft.org/w/images/d/db/1_Introlabels_SLAVOB-final.pdf).

2008: From Grammar-Independent Construction Enumeration to Lexical Types in Computational Grammars. Paper presented at COLING, Workshop on Grammar Engineering Across Frameworks (GEAF) Manchester, August 2008 (http://www.aclweb.org/anthology-new/W/W08/#1700).

2007a: On 'Deep Evaluation' for Individual Computational Grammars and for Cross-Framework Comparison. In: T.H. King and E. M. Bender (eds) Proceedings of the GEAF 2007 Workshop. CSLI Studies in Computational Linguistics ONLINE. CSLI Publications. http://csli-publications.stanford.edu/

2007b: Representing clause-internal binding in an HPSG/LKB grammar. In Branco, A. (ed) Proceedings from DARC 2007 (Discourse Anaphora Resolution Conference), Lagos.

2006: with Dorothee Beermann. Word Sense and Semantic Disambiguation of Constructions in a Deep Processing Grammar. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Paris, France: European Language Resources Association 2006 ISBN 2-9517408-2-4.

2005a: with Dorothee Beermann. The`specifier' in an HPSG grammar implementation of Norwegian.Proceedings of the 15th NODALIDA conference, Joensuu 2005 ed. by S. Werner Ling@JoY: University of Joensuu electronic publications in linguistics and language technology 1. Joensuu 2006

2005b. Implementing Norwegian Reflexives in an HPSG Grammar. In St. Müller (ed) Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar . CSLI Publications, Stanford. (http://csli-publications.stanford.edu/)

2005c. with Dorothee Beermann. Classification of Prepositional Senses for Deep Grammar Applications. In: Valia Kordoni and Aline Villavicencio (eds.): Proceedings of the 2nd ACL-Sigsem Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, Colchester, United Kingdom, ACL-Sigsem, 2005

2004a: with Dorothee Beermann, Jon Atle Gulla and Atle Prange. Trailfinder: a case study in extracting spatial information using deep language processing. In Ton van der Wouden, Michaela Poss, Hilke Reckman, and Crit Cremers (eds) Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting, pp. 121-131, Leiden, Netherlands, 2004.

2004b: with Dorothee Beermann. A treatment of directionals in two implemented HPSG grammars. In Stefan Müller (ed) Proceedings of the HPSG04 Conference, Katholieke Universiteit Leuven. CSLI Publications, 355-377. (http://csli-publications.stanford.edu/)

2004c: with Dorothee Beermann, Berthold Crysmann, Petter Haugereid, Dario Gonella, Daniela Kurz, Giampaolo Mazzini, Oliver Plaehn, and Melanie Siegel. DEEPTHOUGHT deliverable 5.10. Technical report, The DEEPTHOUGHT consortium.

2003a. NorSource: an introduction. Ms, NTNU.

2003b: with P. Haugereid. The NorSource Grammar - an excercise in the Matrix Grammar building design. In: Emily Bender, Dan Flickinger, Frederik Fouvry, and Melanie Siegel (eds) Proceedings of Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI 2003.

1989, with L. Johnsen and A. Pitz. 1989. TROLL. Ms., Univ. of Trondheim.