Site report - *Summer 2017*:

NorSource **and** ***TypeCraft*** **related work**

in the *Research Group in Digital Linguistics, NTNU, Trondheim*

Lars Hellan, Dorothee Beermann, Tore Bruland, Tormod Haugland, Elias Aamot

DELPH-IN Summit, August 07, 2017

Oslo

Activities of last year
1. Inside the grammar, and directly off …

- Maintaining current services (Grammar Sparrer, Grammar Scrabble, …)

- Grammar consolidation:
  - Selected prepositions and adverbs/’particles
  - Selected reflexives
  - Valence-bound locatives

- Creating L2-easy resources from these

Statistics of actions/uses of web services per Aug 03 2017:

stud= Grammar Sparrer, ling=web demo, mlv=MultiVal, tag=Norsource tagger

År	Demo	Antall
2017	tag	160
2017	stud	50441
2017	mlv	215
2017	ling	803
2016	tag	349
2016	stud	67983
2016	mlv	448
2016	ling	1636
2015	tag	231
2015	stud	59631
2015	mlv	588
2015	ling	1222

Activities of last year

2. Creating a Norwegian valence corpus from Norsource

We present a procedure for generating a valence corpus of Norwegian (Bokmål) from a deep grammar using the Leipzig Corpus Collection (LCC). The corpus is presented in the form of IGT (interlinear glossed text) augmented by valence information. As our deep parser we use the computational grammar Norsource (Hellan and Bruland 2015) while our online IGT repository is TypeCraft (Beermann and Mihaylov (2014). The purpose is twofold: (i) making the grammatical information encoded in a deep parser more readily available, and (ii) facilitating a further integration of a deep parser, an online linguistic workbench, and a large corpus of text which together stand for the linkage of linguistic analysis and existing datasets to larger corpus resources.

Typecraft Valence + IGT 'normal form'

String: Jeg vet at hun forbauset Ola

Free translation: I know that she surprised Ola

Morph	Jeg	|vet	|at	|hun	|forbause	|t	|Ola
Citation	|	|vite	|	|	|forbause		|
	|1.SG.NOM	|PRES	|DECL	|3.SG.FEM	|	|PRET	|
	|PN	|V	|COMP	|PN	|V		|Np

vet: SAS: NP+Sdecl
FCT: transWithSentCompl
ConstructionLabel: v-tr-obDECL

forbauset: SAS: NP+NP
FCT: transitive
ConstructionLabel: v-tr

From parse to valence

Whenever a sentence is parsed, this entails – given the HPSG design for full sentence parsing - that each verb in the sentence has been assigned one of the valence frames defined in the grammar for that verb, which means that every pairing of verb with a valence frame in the parse result, and in the import to TC, is a pairing declared as correct by the grammar. In that sense, the valence information supplied for each successful parse is accurate.

Going along with a valence corpus

As linked resource, the corpus goes together with an abstract *Valence Profile*, systematically listing all valence frame types used in the language, and a *Valence Lexicon*, which for each verb of the language lists, as different entries, all the valence frames in which the verb can occur (along with other information); again this information can be exported from the grammar.

From 'deep' grammar to corpus

Valence corpora have often been built manually, or by statistical methods where hand annotation plays a crucial role. In some cases valence corpora, possibly in conjunction with tree-banks, are used in the construction of computational grammars. Here we go the opposite way, exporting information from a deep grammar to an IGT corpus, whereby sentences in the corpus serve as classified examples of the verb valence types of the language.

- English: FrameNet, VerbNet and PropBank (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html),
- German: Evalbu (http://hypermedia2.ids-mannheim.de/evalbu/);
- Czech Vallex (http://ucnk.ff.cuni.cz) ;
- Polish, Walenty (http://clip.ipipan.waw.pl/Walenty; Przepiórkowski & al (2014),
- Cf., Osenova, (2011), Patujek and Przepiórkowski (2016).

Using the parse tree for XML

```
head-subject-rule
 jeg_perspron
  jeg
 head-verb-inf-or-s-comp-rule
  pres-infl_rule
   vite_subord_vlxm
    vet
  head-complementizer-comp-fin-rule
   at_subord
    at
   head-subject-rule
    hun_perspron
     hun
    head-verb-comp-rule
     pret-nonfstr-et_infl_rule
      forbause_tv_vlxm
       forbauset
     sg_def_m_final-full_irule
      sg-masc-def-noun-lxm-lrule
       ordfører_n_masc_nlxm
        ordføreren
```

Using the parse tree for valence extraction

We assign a valence value to every verb occurrence in a sentence. For *vet* a look-up in the verb lexemes file establishes that the identifier in question carries the type *v-tr-obDECL* (cf. the simplified view of a verb entry in (a)), and look-up in a file establishing correspondences between the CL code and the SAS and FCT codes yields (b).

a. vite_subord_vlxm := v-tr-obDECL

b. v-tr-obDECL =>

SAS: "NP+Sdecl";

FCT: transWithSentCompl

From these correspondences the following part of the Figure is established:

vet: SAS: NP+Sdecl

FCT: transWithSentCompl

ConstructionLabel: v-tr-obDECL

The files in which these look-ups are made count 12,000 entries corresponding to (a), and about 400 conversions corresponding to (b).

Using the parse tree for POS and GLOSS extraction

Norsource has a lemma-based lexicon, which means that inflectional processing is done via ‘rules’, stated in a form exemplified below (for verbs 22 such rules, for nouns 28, and for adjectives 38).

pret-nonfstr-et_infl_rule :=

%suffix (e a) (e et) (es es) (es edes)

infl-pret-verb-word &

[ARGS <[INFLECTION nonfstr-et]>].

This rule is mentioned in the tree for *forbauset*, reflected in the lines

pret-nonfstr-et_infl_rule

forbause_tv_vlxm

forbauset

stating that the form *forbauset* has been derived from the lemma form *forbause* by the application of this rule.

Using the parse tree for POS and GLOSS extraction

The appropriate GLOSS tag in TC will be PRET, and this is assigned through the mapping rule below to the GLOSS line in TC:

pret-nonfstr_infl_rule = PRET

There are altogether 75 mapping rules from Norsource inflection rules to TC GLOSS tags. Most of them apply simply to rule names, more examples for verbs are given below

- a. ppart-finalstr-dd_infl_rule = PRF
- b. s-passive_s_infl_rule = PRES.PASS
- c. pl_def_m-or-f_light-e_irule = PL.DEF
- d. sg_def_n_light-e_irule = SG.DEF.NEUT

Using the parse tree for POS and GLOSS extraction

GLOSS for constant words

- a. fordi_comp = CAUS
- b. idet_prep-time = TEMP
- c. mer_cmpar-reg = CMPR
- d. seg_refl = 3P.REFL.ACC
- e. en_indef-art = SG.MASC.INDEF

POS for words according to entry suffix, or to word as a whole (constituting most of the 472 mappings to POS tags :

- a. nlxm = N
- b. alxm = ADJ
- c. vlxm = V
- d. dirtel-end-p = PREPdir
- e. reg-p-loc = PREPplc
- f. mer_cmpar-mass = QUANT

Assessments

- The quality of the valence information depends on the quality of the deep parser, that is, a deep parser combines syntactic and semantic parsing with the recognition of predicate-argument structure, and our valence corpus therefore will be only as good as the parser is in handling these grammatical dependencies. Moreover the quality of the corpus depends on the convervsion itself which is not without complexity, so that mistakes could arise, and, per the automatic design, 'infect' a large number of sentences.
- Yet an obvious advantage of the method is that, once analyses are deemed plausible, one can in relatively little time obtain a comprehensive valence corpus.

Assessments

- A valence resource should also include a valence lexicon, where each verb is specified for all the frames in which it can occur, preferably with links to selected examples from the corpus, or to the corpus seach interface. Such a facility is not yet included in TC, but is, apart from corpus links, independently available in MultiVal, which is a parallel valence lexicon for four languages, Norwegian, Ga, Spanish, Bulgarian, with the same labels for valence frame encoding as are used presently.
- A combination of the resources, where also frequency data based on the corpus can be called upon, is an obvious further desideratum.
- Cf. Hellan et al. 2014, http://regdili.hf.ntnu.no:8081/multilanguage_valence_demo/multivalence. The lexicons are based on computational HPSG grammars for the four languages, where the verb lexicons of course by themselves are valence lexicons, just not online searchable.

References

- Beermann, Dorothee and Mihaylov, Pavel, 2014. Collaborative databasing and Resource sharing for Linguists. *Languages Resources and Evaluation* 48. Dordrecht: Springer, 1-23.
- Dakubu, M.E. Kropp and Lars Hellan, 2017. A labeling system for valency: linguistic coverage and applications. In Hellan, L., Malchukov, A., and Cennamo, M (eds) *Contrastive studies in Valency*. Amsterdam & Philadelphia: John Benjamins Publishing Co.
- Goldhahn, D., Eckart, T., Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012.
- Hellan, Lars and M.E. Kropp Dakubu, 2010*. Identifying verb constructions cross-linguistically*. In *Studies in the Languages of the Volta Basin* 6.3. Legon: Linguistics Department, University of Ghana.
- Hellan, L., D. Beermann, T. Bruland, M.E.K. Dakubu, and M. Marimon (2014) *MultiVal*: Towards a multilingual valence lexicon**.** In Calzolari et al. (eds) 2014.
- Hellan, L., Bruland, T., Aamot, E., Sandøy, M.H. (2013): A Grammar Sparrer for Norwegian. *Proceedings of NoDaLiDa* 2013.
- Hellan, Lars and Tore Bruland. 2015. A cluster of applications around a Deep Grammar. In: Vetulani et al. (eds) *Proceedings from The Language & Technology Conference* (LTC) 2015, Poznan.
- Adam Przepiórkowski, El´zbieta Hajnicz, Agnieszka Patejuk, Marcin Woli´nski, Filip Skwarski, and Marek ´Swidzi´nski. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In Calzolari et al. (eds) 2014.
- Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pages 2785–2792, Reykjavík, Iceland. ELRA.
- Osenova, Petya (2011). Localizing a Core HPSG-based Grammar for Bulgarian. In: Hanna Hedeland, Thomas Schmidt, Kai Worner (eds.) *Multilingual Resources and Multilingual Applications*, *Proceedings of GSCL 2011*, ISSN 0176-599X, Hamburg, pp. 175-180.
- Patejuk, Agnieszka (2016). Integrating a rich external valency dictionary with an implemented XLE/LFG grammar. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, Stefan Müller (editors) *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar.* Stanford: CSLI Publications, pp 520-540.