Report for PARSEME Short Term Mission,
reference code: COST-STSM-ECOST-STSM-IC1207-230315-055491.
by
Lars Hellan, NTNU, and Petya Osenova, IICT-BAS
Trondhein/Sofia 21.04.2015

From 23 March to 30 March the mission took place at IICT-BAS, Sofia, with the grantee Lars Hellan as visitor and Petya Osenova and Kiril Simov as hosts, in pursuing the objectives stated in the mission proposal:
The aim of the mission is to enhance the Norwegian-Bulgarian valence lexicons, used in the language deep grammars, with MWEs of various types. Special focus will be put on the following linguistic features: the degrees of compositionality and fixedness. As results we expect: A meaningful comparison of MWEs in the two languages and enrichment of the lexicons with new lexical entries, which would approve the accuracy of the grammars, and enhance the expressiveness and accessibility of the verb information in the valency databases of the languages..
In addition to intensive working sessions every day, Prof. Hellan also gave two presentations, one in the institute (IICT) and one for the BulTreeBank project group.

Background

We had at our disposal the following prerequisites: the *Norsource*[1] and *BURGER*[2] *HPSG*[3] grammars; valency lexicons as part of the grammars; the *BulTreeBank*[4], the annotation platform *TypeCraft*[5]; the online multilingual valency lexicon *MultiVal*[6]; a partial inventory of MWE types in each language; valency frames from BulTreeBank and from Norsource. We had decided to focus on verb-headed MWEs in both languages, and among them, on light verb constructions and semi-fixed idioms, since Bulgarian lacks phrasal verbs. Our approach would include the idea of using *catena*[7] in the encoding of MWEs in the lexicon and in the data.

Results

1. An appropriate encoding of MWEs in the TypeCraft system.
The chosen encoding is illustrated below, for the Norwegian MWE *stå på agendaen* ('be on the agenda'; see (2a) below for a more detailed view of the TypeCraft annotation interface):

(1)	**stå**	**på**	**agendaen**
	HEAD.CAT1	OBL.CAT1	POBJ.CAT1
	stå	on	agenda-the

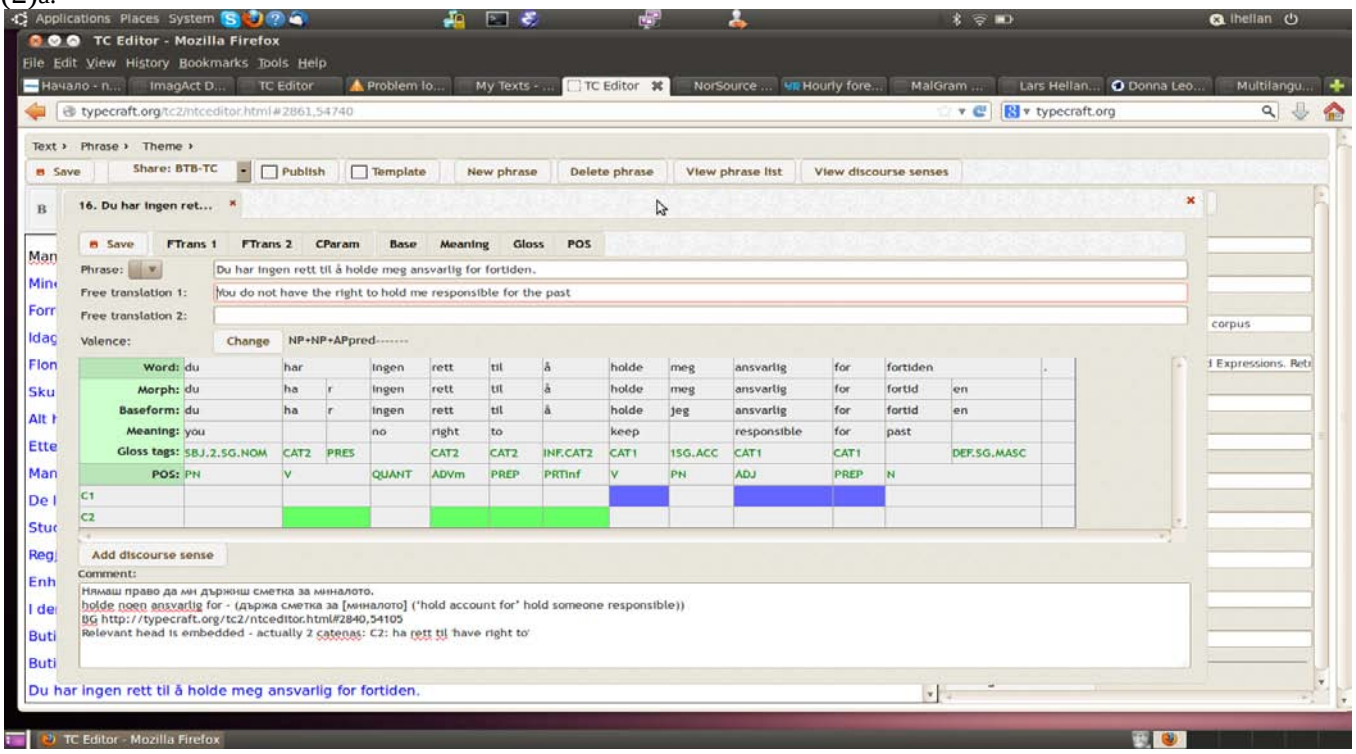
The encoding marks the parts of the catena (CAT1) over grammatical roles (HEAD, OBL(ique), POBJ (prepositional object)). Annotated examples including this information will be used for enriching the lexicons in grammars and in the MultiVal lexicon.

2. Annotation of 30 parallel (BG-NO) sentences containing MWEs in TypeCraft.
The annotation includes: catena over the MWEs, free translation into English and morphosyntactic information in the form of interlinear glossing. The English translation constitutes the common

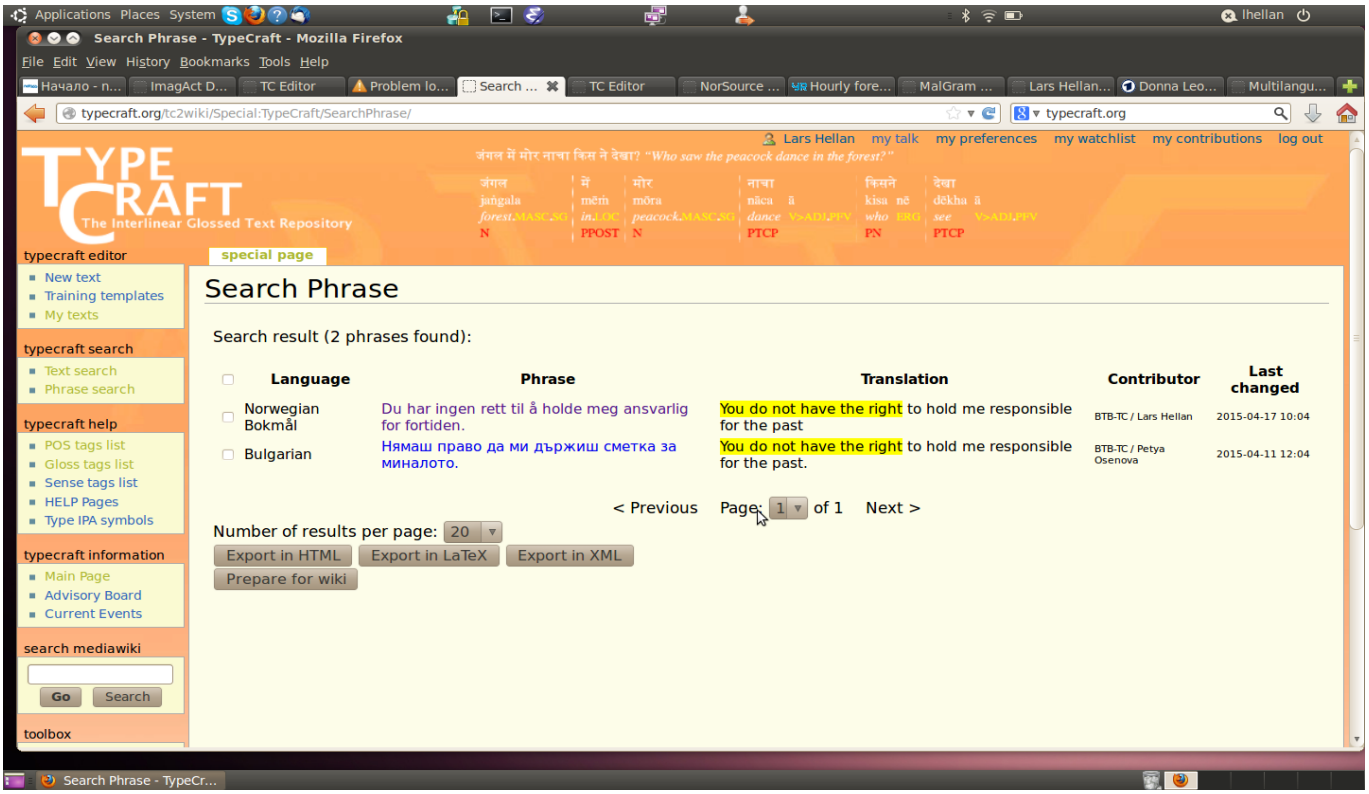
[1] Norsource: http://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource ; Beermann and Hellan 2004, Hellan 2008.
[2] See Osenova 2011.
[3] For introduction to the general framework, see Pollard and Sag 1994, and for the implementation format of the grammars used, the LKB platform, see Copestake 2002. The platform is generally maintained through the DELPH-IN consortium, cf. http://moin.delph-in.net/
[4] See Simov et. al 2005
[5] http://typecraft.org/tc2wiki/Main_Page; Beermann and Mihaylov 2013.
[6] http://regdili.hf.ntnu.no:8081/multilanguage_valence_demo/multivalence ; Hellan et al. 2014.
[7] See Simov and Osenova 2014

denominator for the corresponding Bulgarian and Norwegian sentences. Beneath are screenshots illustrating in (2a) the Norwegian member of such a pair, with MWE annotation following the format shown in (1), and in (2b) the result for a search leading to this pair, prompted by its common meaning, as exhibited in the search interface in TypeCraft:

(2)a.



b.



3. Preliminary annotation of MWEs with catena in lexicons.
As a model we use the way in which verbal MWEs are represented in Norsource, exemplified below with the lexical entry for *alliere* ('ally') for the frame where it selects a reflexive and a PP headed by the preposition *med* (corresponding to English 'ally oneself with'); cf. (3) below. Like all verb entries, this one consists of an orthographic form (conventionally selected to be like the infinitive form), under STEM, a conjugation class under INFLECTION indicator ('nonfstr-te' representing 'non-finally-stressed infinitival form, of the type selecting 'te' as past tense form), and a semantic predicate placeholder (introduced by the path 'SYNSEM.LKEYS.KEYREL.PRED'). In addition the entry has the line 'SYNSEM.LOCAL.KEY-SPEC med-assoc' whereby 'med' is indicated as a selected preposition. (In the source format notation used, ':=' means 'is a subtype of' and '&' means 'unified with'.)

(3)
```
alliere_tr-obl-refl := v-trObl-obRefl_oblN &
  [ INFLECTION nonfstr-te,
    STEM < "alliere" >,
    SYNSEM.LKEYS.KEYREL.PRED "_alliere_v-trObl_rel",
    SYNSEM.LOCAL.KEY-SPEC med-assoc ].
```

The specification in the last line in the lexical entry is integrated into the general type system of the grammar through the type definition below, where the path 'LOCAL.KEY-SPEC _' is defined as having the same value as 'LOCAL.CAT.VAL.ICOMPS <[LOCAL.CAT.HEAD [KEYS.KEY _]]>' (indicated by '#1' in both underscore positions), where the longer path specifies restrictions on a selected PP for the verb frame in question:

(4)
```
trans-obl-synsem-sup := arg1-subj-synsem & arg2-comps-synsem &
  [ LOCAL.CAT.HEAD non-copula,
    LOCAL.CONT.HOOK hook & [ VARG #vrg],
    LOCAL.CAT trans-with-1icomps-pp-cat,
    LOCAL.CAT.VAL.COMPS <[LOCAL.CAT.HEAD nominal ]>,
    LOCAL.KEY-SPEC #1,
    LOCAL.CAT.VAL.ICOMPS <[LOCAL.CAT.HEAD prep-or-adv & [SELECTED +,
                                    KEYS.KEY #1 & index-sita]]>,
    LOCAL.CAT.VAL.ICOMPS <[LOCAL.CAT.HEAD [KEYS.KEY #1 ]]>,
    LOCAL.CAT.QVAL.OBL1.LOCAL.CONT.HOOK.LTOP #4 & handle,
    LOCAL.CAT.QVAL.OBL1.LOCAL.CONT.HOOK.VARG #vrg,
    LOCAL.CAT.VAL.ICOMPS <[LOCAL.CAT.HEAD [SELECTED +]]>,
    LKEYS.KEYREL arg12obl-rel & [ ARGOBLQ #4 ]].
```

Although such general definitions have to be given for all relevant verb frame types (which, super-types included, may reach about 150 in a complete inventory), the strategy is straightforward to implement in any grammar with the feature structure format in question (reflecting the 'HPSG Grammar Matrix' design[8]), such as Norsource and BURGER.

4. Preliminary ideas on data exchange strategy between CLaRK[9] system and TypeCraft:
Since BulTreeBank has been created in the CLaRK system, a two-direction conversion template will be created in order to manage the data between the two systems in XML.

[8] See Bender et al. 2010.
[9] See Simov et al 2001.

For the incremental addition of new data, the template will be used for both purposes: loading the data into TypeCraft, encoding it (with valencies and catena), and bridging it back to the CLaRK system format.
In this way, the communication between the two systems will be ensured for the purposes of control and data syncronization.

Future Work

We plan to:
1. extend the parallel MWE corpus further with data from parallel corpora;
2. enrich the lexicons in the grammars with MWEs;
3. test the format of catena-based description on more data and more MWE types;
4. implement the data exchange strategy mentioned under 'Results', point 4;
5. relate the parallel MWE annotation to the 3D system of verb action representation, *ImagAct*[10];
6. with a view to Machine Translation as a field of application of the encoding of MWEs in corpora and grammars, explore ways of reflecting MWEs in the semantic representation format Minimal Recursion Semantics ('MRS', cf. Copestake et al. 2005), which constitutes part of the 'matrix' architecture.

Joint publications are envisaged reflecting some of these planned activities.
A precise time plan for these activities has not been made yet; most of the cooperation can be done remotely between Sofia and Trondheim, but we envisage applying for a further STM to enhance aspects of the work calling for more direct cooperation.

References

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

Beermann, D. and Mihaylov, P. (2013). Collaborative databasing and Resource sharing for Linguists. *Languages Resources and Evaluation*. Springer, 1-23.

Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L. and Saleem, S. (2010). Grammar Customization. In *Research on Language & Computation*, Volume 8, Number 1, 23-72.

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

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. (2005). Minimal Recursion Semantics. In *Journal of Language and Computation*. 281-332.

Hellan, L. (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).

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

Osenova, Petya Osenova (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.

Pollard, C. and Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago University Press.

Simov, K., Osenova, P., Simov, A. and Kouylekov, M. (2005): Design and Implementation of the Bulgarian HPSG-based Treebank. In Erhard Hinrichs and Kiril Simov, editors, *Journal of Research on Language and Computation*, Special Issue, Springer, 2005, pp. 495-522.

Simov, K and Osenova, P. (2014): Formalizing MultiWords as Catenae in a Treebank and in a Lexicon. In: Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski (eds.) Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), December 12-13, 2014, Tübingen, Germany, pp. 198-207.

Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A (2001). CLaRK - an XML-based System for Corpora Development. In: Proc. of the Corpus Linguistics 2001 Conference, pages: 558-560.

[10] See http://www.imagact.it