A common indexing format for use in central test-suites, lexicons, and more

Lars Hellan

lars.hellan@hf.ntnu.no

Some concerns

The main linguistic contribution of ('deep') computational grammars lies in the systematization and consolidation of agreed-upon insights.

To bring out this contribution, we ought to better expose what the grammars encode, and their potential for playing roles in larger linguistic enterprises, individually or collectively

Accordingly, grammars should be evaluated not only in terms of their (MRS) inputs into language applications, but also in terms of how they actually handle the languages they are designed to analyze - - i.e., their linguistic effort.

Indexing for Facts

Prerequisite:

A reference frame for representing ***facts*** of a language – at any level of generality, but as opposed to properties of grammars.

Such a reference frame can be realized as an indexing system by which language tokens can be annotated and a language as a whole can be summarized.

This can provide a format for cross-linguistic alignment, to expose which grammars deal with which types of facts and in what manner.

For instance:

- To synchronize MRS outputs, it will be helpful to have a standard cross-linguistic system for classifying phenomena according to expected rendering in MRS format.
- To synchronize lexicons, it will be helpful to have a standard cross-linguistic system for classifying ***valence*** types.

If central test-suites of grammars can be indexed to expose such factors in a uniform system, then we can also define cross-linguistic benchmarks for grammars, and thereby progress a bit in terms of ***evaluation*** suitable for 'deep grammars'.

Templates

We illustrate with an indexing system covering types of argument structure, or ***valence types***. Each such type is represented by a ***template***, which is a string of 'fractal' labels each defining an aspect of a valence type, formally corresponding to a subsection of an AVM of type *sign*.

This 'fractal' system constitutes a limited set of formal-conceptual pieces, combinable in many ways, representing a claim as to what are the basic building blocks in the design of verb construction types cross-linguistically. The claim is tested against how one – without forcing the facts – can accommodate valence inventories of languages from any known typology.

Templates

The template format is exemplified below, where the hyphens and underlines can be interpreted as *unification*:

v-tr-suAg_obAffincrem-COMPLETED_MONODEVMNT

Ex.: He ate the cake

A mapping of the constituent labels to AVMs is illustrated in the next slide, such that the string will correspond to the unification of all the AVMs. The ‘result’ of this unification is the AVM in the subsequent slide.

Template to AVM - 1

v-tr-suAg_obAffincrem-COMPLETED_MONODEVMNT

v - - - [HEAD verb]

tr - - -
$$\begin{bmatrix} \text{GF} \begin{bmatrix} \text{SUBJ} \begin{bmatrix} \text{INDX} \boxed{1} \end{bmatrix} \\ \text{OBJ} \begin{bmatrix} \text{INDX} \boxed{2} \end{bmatrix} \end{bmatrix} \\ \text{ACTANTS} \begin{bmatrix} \text{ACT1} \boxed{1} \\ \text{ACT2} \boxed{2} \end{bmatrix} \end{bmatrix}$$

suAg - - - [GF [SUBJ [INDX [ROLE agent]]]]

obAffincrem - - - [GF [OBJ [INDX [ROLE aff-increm]]]]

COMPLETED_MONODEVMNT - - -
$$\begin{bmatrix} \text{ASPECT completed} \\ \text{SIT-TYPE monotonic\_development} \end{bmatrix}$$

Template to AVM - 2

v-tr-suAg_obAffincrem-COMPLETED_MONODEVMNT

He ate the cake

$$\begin{bmatrix} \text{HEAD verb} \\ \text{GF} \begin{bmatrix} \text{SUBJ} \left[\text{INDX } \boxed{1}\left[\text{ROLE agent}\right]\right] \\ \text{OBJ} \left[\text{INDX } \boxed{2}\left[\text{ROLE aff-increm}\right]\right] \end{bmatrix} \\ \text{ASPECT completed} \\ \text{ACTANTS} \begin{bmatrix} \text{ACT1 } \boxed{1} \\ \text{ACT2 } \boxed{2} \end{bmatrix} \\ \text{SIT-TYPE monotonic\_development} \end{bmatrix}$$

Template build-up

Slot 1 consists of a label for *Parts of Speech* of the *head* of the entire construction, including the category of possible *formatives* marked on the head.

Slot 2 consists of a label for *valency specification* - like intr (intransitive), tr (transitive), ditr (ditransitive), and varieties thereof.

Slot 3 consists of one or more labels for specification of *syntactic constituents*, identified by their grammatical function (subject, object, etc.).

Slot 4 consists of one or more labels for specification of *participant roles*: agent, theme, instrument etc.

Slot 5 consists of a label for *aspect and aktionsart*, written in CAPS.

Slot 6 consists of a label for the *situation type* of the construction, also written in CAPS.

V-profile for Norwegian

Shown ‘off-slide’ is a condensed version of the indexed test-suite for verb construction types in Norwegian. With differentiating labels only for slots 2 and 3, it unfolds 274 verb valence types in Norwegian (some with more than one sentence). The non-condensed counterpart is the commented test-suite *test-v-stnd* used in Norsource, from a version of July 2011. Nearly all of the types are implemented in the grammar; whether implemented or not, this indexing represents a ‘facts’ benchmark for what the grammar should accomplish for this set of sentences.

Such a list of templates we call a ***v-profile*** of the language in question.

A Norwegian v-profile including specification for slots 4, 5 and 6 is given as Appendix 2 in Hellan and Dakubu 2010 (where a v-profile for the Ghanaian language Ga constitutes Appendix 1).

V-profile for English

Also shown off-slide is a commenced corresponding template list for English – too small to count as a v-profile yet, but serving as a basis of a linguistic contrasting study between the two languages, to be presented at the upcoming SLE 2012 workshop ‘Contrastive Studies in the Valency of European Languages’ (organized by A. Malchukov, J. Barddal, A. Kibort, M. Cennamo, and L. Hellan)

Obviously, although developed independently of ERG or any other implemented English grammar, it could serve as a test-suite for such a grammar, and by its template specifications, induce a benchmark definition.

V-profiles and verb types

V-profiles can be exchanged into verb type inventories

- (i) by mapping the v-profile classification onto a classification using the verb types of the grammar;
- (ii) by defining the verb types of the grammar as carrying the names of the templates of the v-profile themselves.

The latter can be achieved in two ways:

- (iia) by defining each template ’en bloc’ as a type name, with the constituent fractals serving only as reminders to the grammar user of what the types stand for (as currently done for *NorSource*);
- (iib) by having a general type hierarchy accommodating the *fractals*, and deriving the types corresponding to the templates by unification of the constituent fractals (as in the *TypeGram* family of small grammars (cf. tomorrow)).

From verb types to Lexicon

With a verb type inventory reflecting valence, and a verb list of your language, the way is open to compile a verb lexicon with this type of information (preferably added to other information, if the 'list' is already a dictionary).

This was the genesis of the verb parts of the TROLL/ NorKompLex lexicons of Norwegian in the 80ies-90ies, and of a current Ga verb lexicon, see off-slide, both readily convertible into lexicon files of grammars.

Coming the other way, from a lexicon file of a grammar, one obviously can create an interesting lexical resource by spelling out the information encoded.

Multilingual valence bank?

Moreover, whether a set of grammars actually use the same type names for valence information, or align their type inventories through some mapping or script,* one can envisage a constellation of lexical resources from multiple grammars which could relatively easily be turned into a ***multilingual valence bank***.

That could be of interest to the 'Multilingual Europe', for instance.

* Probably the latter. Any grammar(ian) needs a 'private space', some secrecy through non-transparency, and free rule.

The MRS test-suite

The closest we have to such a scenario in the Delph-In world is probably the 'MRS test suite', a multilingual repository of parallel test suites where cross-linguistic correspondence is indicated by the numbering of sentences in the sequence – each language having about 100 sentences. The number of a sentence does not per se indicate a specific grammatical construction type, or a concise semantic interpretation, but a mixture of the two with semantic correspondence as the main factor. Proposed by Ann for English in 2002 or so, it has been very useful for multlingual grammar construction as a practical benchmark.

But it carries no concise indexing of the kind here considered, neither for syntactic nor semantic properties.

Annotating MRS testsuite

It might indeed be useful to index each MRS test-suite according the system here presented (or some alternative system, reasonably aligned), to expose the morpho-syntactic particulars of the examples from each language, along with a semantic indicator of what the example is supposed to instantiate in terms of MRS structure.

Saying ’morpho-syntactic’, we envisage an indexing using not only the template system described above, but standard interlinear glossing as well.

To make these annotated versions of the MRS test-suites searchable in a common format, ***TypeCraft*** will be a convenient medium.

Template + IGT – Ga 1

Going beyond English examples, it is easily seen that a valence template is preferably is combined with a ‘normal’ glossing. Ex. From Ga:

v-ditr-obPostp-suAg_obEndpt_ob2Mover-PLACEMENT

Amɛ-wo	***tsɔne***	***lɛ***	***mli***	***yɛlɛ***
3P.AOR-put	vehicle	DEF	inside	yam
V	N	Art	N	N

‘They put [vehicle’s inside] [yam]’ =‘They put yams in the lorry.’

Template + IGT – Ga 1

v-ditr-obPostp-suAg_obEndpt_ob2Mover-PLACEMENT

Amɛ-wo	***tsɔne lɛ mli***	***yɛlɛ***
3P.AOR-put	[vehicle DEF inside] Ob	[yam] Ob2

- **ditr**: double object construction;
- **obPostp**: the First Object is a ‘postpositional phrase’, i.e., an NP with a head expressing a spatial domain (a ‘locative noun’ in some terminologies) relative to the item expressed in the Specifier;
- **obEndpt**: the First Object represents the Endpoint of a movement;
- **ob2Mover**: the Second Object represents the Mover of a movement;
- **PLACEMENT**: The situation type expressed is one of *placement*.

Template + IGT – Ga 2

v-tr-suPossp_obIDsuSpec-suBPsuSpec_suLocus_obExp-EXPER

Mi-hiɛ	***di***	***mi***
1S1.POSS-face	black	1S1
N	V	Pron

“My face blackens me” =‘I am dizzy’

- **suPossp**: the Subject is a possessive phrase (NP with an NP specifier)
- **obIDsuSpec**: the Object is (referentially) **ID**entical to Specifier of the Subject
- **suBPsuSpec**: the Subject is (referentially) a **B**ody**P**art of the Specifier of the Subject
- **suLocus**: the subject expresses the ‘locus’ of the situation.
- **obExp**: the Object expresses an Experiencer.
- **EXPER**: The situation type is one of *experiencing* (someone having an experience).

SVC and Situation type

‘Global’ situation type may be distinct from ‘local’ sit-type:

Akan: (1)

svAspID-v1tr-v1obIDv2su-v1suAg_v1obEjct-v2tr-
v2suTh_v2obEndpt

*K**ofi***	***to-o***	***ne***	***nan***	***wɔ-ɔ***	***Kwame***
Kofi	throw-CMPL	3Poss	leg	pierce-CMPL	Kwame
N	V	Pron	N	V	N

‘Kofi kicked Kwame’

v1oblDv2su = v1’s object IDENTICAL to v2’s subject

SVC and Situation type

‘Global’ situation type may be distinct from ‘local’ sit-type:

Akan: (2)

svAspID-v1tr-v1obIDv2su-v1suAg_v1obEjct-v2tr-v2suTh_v2obEndpt

-**CONTACTEJECTION**

*K**ofi***	***to-o***	***ne***	***nan***	***wɔ-ɔ***	***Kwame***
Kofi	throw-CMPL	3Poss	leg	pierce-CMPL	Kwame
N	V	Pron	N	V	N

‘Kofi kicked Kwame’

SVC and Situation type

‘Global’ situation type may be distinct from ‘local’ sit-type:

Akan: (3)

svAspID-v1tr-v1obIDv2su-v1suAg_v1obEjct-v2tr-v2suTh_v2obEndpt-CONTACTEJECTION-**LAUNCHERv1su_MOVERv1ob_TARGETv2ob**

Kofi	***to-o***	***ne***	***nan***	***wɔ-ɔ***	***Kwame***
Kofi	throw-CMPL	3Poss	leg	pierce-CMPL	Kwame
N	V	Pron	N	V	N

‘Kofi kicked Kwame’

Citumbuka (Malawi)

v-ditrOblApCs-oblCsu_obAobl-suCsr

Tumbikani wa-ka-mu-phik-isk-ir-a Temwa nchunga kwa Mary

Tumbikani 1SM-pst-1OM-cook-Caus-Appl-fV Temwa beans 'to' Mary

'Tumbikani made Mary cook beans for Temwa'

Causative and Applicative morphology

$$
\begin{bmatrix}
\mathrm{HEAD} & \mathrm{verb} \\
\mathrm{GF} & \begin{bmatrix}
\mathrm{SUBJ} & \begin{bmatrix} \mathrm{INDX} & \boxed{1} \begin{bmatrix} \mathrm{ROLE} & \mathrm{causer} \end{bmatrix} \end{bmatrix} \\
\mathrm{OBJ} & \begin{bmatrix} \mathrm{INDX} & \boxed{3} \begin{bmatrix} \mathrm{ROLE} & \mathrm{benefactive} \end{bmatrix} \end{bmatrix} \\
\mathrm{OBJ2} & \begin{bmatrix} \mathrm{INDX} & \boxed{2} \begin{bmatrix} \mathrm{ROLE} & \mathrm{theme} \end{bmatrix} \end{bmatrix} \\
\mathrm{OBL} & \begin{bmatrix} \mathrm{GOV} & \begin{bmatrix} \mathrm{INDX} & \boxed{4} \begin{bmatrix} \mathrm{ROLE} & \mathrm{agent} \end{bmatrix} \end{bmatrix} \end{bmatrix}
\end{bmatrix} \\
\mathrm{ACTNTS} & \begin{bmatrix}
\mathrm{PRED} & \mathrm{cause} \\
\mathrm{ACT1} & \boxed{1} \\
\mathrm{ACT2} & \begin{bmatrix}
\mathrm{ACT1} & \boxed{4} \\
\mathrm{ACT2} & \boxed{2} \\
\mathrm{ACTobl} & \boxed{3}
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

Indexing and 'Meta-grammar'

Any framework of formal grammar is, strictly speaking, 'an indexing system by which language tokens can be annotated and a language as a whole can be summarized' (cf. slide 3) – the 'summary' is the grammar of the language, and the 'annotations of tokens' are sentence analyses/parses produced by the grammar.

So, what has been addressed now is not the attainment of special contact with a 'Sprache an sich', but rather a wrapper around grammars which does essentially what grammars do, but in an alternative way so as to provide a reference frame for reflecting grammars.

Such a concept is not new – 'metagrammar' being one notion used – and here we have addressed a practical implementation of it.

Selected references

- ***Beermann, D. and Mihaylov, P (2009). TypeCraft – Glossing and Databasing for Linguists. Proceedings of the 23th* Scandinavian Conference of Linguistics, Uppsala, Sweden, October 2008.**
- ***Copestake, A. (2002). Implementing Typed Feature Structure Grammars. CSLI Publications, Stanford.***
- ***Dakubu, M.E.K. (2000). Ga-English Dictionary with English-Ga Index. Accra: Black Mask Publishers. Pp. 226. [2nd edition, revised and expanded, to appear 2009]***
- ***Dakubu, M.E. K. (2008). The construction label project: a tool for typological study. Presented at West African Languages Congress (WALC), Winneba, July 2008.***
- ***Hellan, L. (2008). Enumerating Verb Constructions Cross-linguistically. COLING Workshop on Grammar Engineering Across Frameworks (GEAF). Manchester. (http://www.aclweb.org /anthology-new/W/W08/#1700).***
- ***Hellan, L. and Dakubu, M.E.K. (2010). Identifying Verb Constructions Cross-linguistically. SLAVOB series 6.3, Univ. of Ghana.***