Veredas: Frame Semantics and Its Technological Applications
Vol. 17, nº 1, 2013

Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project

Lars Borin (Språk-Banken / University of Gothenburg)
Markus Forsberg (Språk-Banken / University of Gothenburg)
Benjamin Lyngfelt (Språk-Banken / University of Gothenburg)

ABSTRACT: The Swedish FrameNet++ (SweFN++) project aims at developing an integrated Swedish lexical macro-resource to be used primarily in language technology R&D to build natural language processing (NLP) applications. Most of the component resources making up SweFN++ are existing digital lexical resources; in their case the central project effort is directed at making them interoperable on as many levels as possible. An important new resource being created in the project is a Swedish framenet. Now a sister project is starting with the aim of adding a Swedish constructicon (SweCxn) to the macro-resource. In this paper, we discuss some theoretical and conceptual issues which have arisen in the course of our work on the SweFN++ and the planning of the SweCxn, in the close encounter between the practical requirements of NLP and the theory and practice of linguistic – lexical and grammatical – description.

KEYWORDS: SALDO; Swedish FrameNet; SweFN; SweFN++; Swedish constructicon; SweCxn; language technology resources; FrameNet; constructicon; construction grammar; lexical resources

Introduction

> *Close encounters of the first, second, and third kind [...] were first described by astronomer J. Allen Hynek 20 years ago. [...] [E]mergency medical physician Steven Greer of Asheville, North Carolina, has recently come up with another category: Close encounters of the fifth kind, in which humans and aliens intentionally communicate [...]* (Paul McCarthy: Close encounters of the fifth kind – communicating with UFOs, *Omni*, December 1992)

Over the last two decades or so, research in language technology (LT)[1] has become increasingly disassociated from the concerns of linguistics (REITER, 2007). According to Wintner (2009, p. 642), there are mainly three reasons for this: (1) "applications that were based on explicit linguistic knowledge didn't scale up well"; (2) "[f]unding agencies (mainly in the U.S.) are motivated by short-term practical goals"; and (3) "[linguistics] focused mainly on syntax (and predominantly on English): and its theory became so obscure, so baroque, and so self-centered, that it became virtually impenetrable to researchers from other disciplines". The somewhat myopic view of linguistics evinced under the third point is understandable, given that this was the kind of linguistics that informed much early research in LT, while the field remained strangely untouched by other, equally vigorous (but perhaps less vociferous) strands of linguistic research, such as typological linguistics, language contact research, sociolinguistics, or lexicography (including lexical semantics), to mention a few that all have generated a substantial body of work simultaneously with – sometimes even well before – the kind of linguistics that Wintner refers to.[2]

[1] Three other terms denoting basically the same field of enquiry as LT are *natural language processing* (NLP), *computational linguistics* (CL), and *(natural) language engineering*. We will use these terms interchangeably as synonymous in this paper.

[2] The same tradition that Wintner alludes to has spawned more recent, less syntax- and English-centered formalisms, such as Lexical-Functional Grammar and Head-Driven Phrase Structure Grammar, but their

This topic deserves a longer and more exhaustive historiographic treatment, which it is not our aim to try to give here. However, it seems to us – looking at, e.g., the proceedings of the long-running series of international conferences on computational linguistics (COLING)[3] held biennially since the 1960s – that except at its very beginning, the interaction of computational and other linguistics has been almost exclusively with the kind of linguistics that Wintner characterizes as syntax-focused and English-centered, a kind that, tellingly, a traditional comparative Semitist once referred to as "the computational paradigm" in linguistics (RUNDGREN, 1982). What is rarely noted is that, even here, the traffic has been largely one-way, since, with some rare exceptions, work in computational linguistics has had very little influence on the development of theory or methodology in general linguistics. However, Reiter (2007) and even more so Wintner (2009) see many opportunities for closer interaction between LT and linguistics, which crucially would not be one-way, but a genuinely synergistic endeavor.

In this spirit, we discuss some issues which have arisen in our work on a Swedish lexical macro-resource, designed and compiled primarily with LT applications in mind, but with strong roots in traditional linguistic description – we are interested in finding solutions that are motivated both from an LT and a theoretical-linguistic viewpoint – and their potential theoretical and conceptual repercussions for linguistics.

1. *E pluribus unum*: The Swedish FrameNet++ and the Swedish constructicon

The Swedish FrameNet++ (SweFN++) project (BORIN et al., 2010) is many things simultaneously. Its main goal is the creation of an integrated lexical macro-resource for Swedish to be used as a basic infrastructural component in Swedish LT research and in the development of NLP applications for Swedish. At the time of writing, the project is half-way through the funding period. It is now being joined by a sister project aiming to build a Swedish constructicon (SweCxn; LYNGFELT et al., 2012). The specific objectives of the SweFN++ project are (1) to link a number of existing free lexical resources – both in-house and external, both modern and historical – into an integrated lexical macro-resource; (2) to create a full-scale Swedish FrameNet with at least 50,000 lexical units and fully integrated into the macro-resource; and (3) to develop methodologies and workflows which make maximal use of LT tools and large text corpora in order to minimize the human effort needed in the work.

To this macro-resource the SweCxn project aims to add a constructicon (see section 3 below), thereby addressing the question of how to account for linguistic patterns that are too specific to be attributed to general grammatical rules but too general to simply include as lexical units. Most of the constructions described in SweCxn are partially schematic, i.e., they typically consist of both variable and lexically fixed constituents.

The macro-resource is topologically a hub-and-spokes structure. There is one primary lexical resource, a pivot, to which all other resources are linked. This is SALDO (BORIN; FORSBERG; LÖNNGREN, 2008, forthcoming), a large (127K entries and 1.9M wordforms),

impact on computational linguistics has been fairly insubstantial, partly because of the "statistical turn" that the field has experienced more or less simultaneously with the advent of these formalisms (Wintner's first point above).

[3] The COLING proceedings from 1965 onwards are accessible online through the ACL Anthology <http://www.aclweb.org/anthology-new/>.

freely available (under a Creative Commons Attribution license) morphological and lexical-semantic lexicon for modern Swedish. It has been selected as the pivot partly because of its size and quality, but also because its form and sense units are identified by carefully designed unique persistent identifiers (PIDs) to which the lexical information in other resources are linked.

The standard scenario for a new resource to be integrated into the macro-resource is to (partially) link its entries to the sense PIDs of SALDO. This cannot be done automatically on the level of word senses in the general case. However, like many other linguistic phenomena, the distribution of senses over citation forms in lexical resources is roughly Zipfian (MOON, 2000; BORIN, 2010); see section 3.1 below. Thus, the vast majority of the lemmas are monosemous, reducing the sense mapping problem to the much simpler problem of pairing up forms between lexical resources. Doing this typically has the effect that the ambiguity of a resource becomes explicit: the bulk of the resources associate lexical information to part-of-speech-tagged base forms, information not always valid for all senses of that base form. This is natural since most of the resources have initially been created for human consumption, and a human can usually deal with this kind of underspecification without problem. Some of these ambiguities can be resolved automatically – especially if information from several resources are combined – but in the end, manual work is required for complete disambiguation.

The macro-resource also includes historical lexical resources (BORIN; FORSBERG; KOKKINAKIS, 2010; BORIN; FORSBERG, 2011), where the starting point is four digitized paper dictionaries: one 19th century dictionary, and three Old Swedish dictionaries. To make these dictionaries usable in a language technology setting, they need morphological information, work that was initiated in the CONPLISIT project for 19th century Swedish (BORIN; FORSBERG; AHLBERGER, 2011) and in a pilot project for Old Swedish (BORIN; FORSBERG, 2008, 2011), and which is now being continued in an ongoing project aiming at creating a diachronic BLARK[4] for Swedish (BORIN; FORSBERG; KOKKINAKIS, 2010; ADESAM; AHLBERG; BOUMA, 2012; AHLBERG; BOUMA, 2012). Linking SALDO to the historical resources is naturally a much more complex task than linking it to the modern resources, especially when moving far back in time. The hope is that a successful (but possibly partial) linking will make it possible to project the modern lexical-semantic relations onto the historical resources, so that, e.g., a framenet-like resource for Old Swedish becomes available for use.

2. Some close encounters between LT and linguistic description

Even though our own background is more in linguistics than in computer science, and even though many of the existing resources being integrated into the macro-resource are originally traditional lexicographic products, this work has forced us to take a fresh look at a number of linguistic issues, both from the point of view of a 'no-loopholes-allowed' formalization and from the point of view of NLP. Here, we discuss three such issues, on which some central design decisions hinge: Zipfian distributions in language and the status of law-like generalizations of the kind traditionally bandied about in linguistics, concretely illustrated with inflectional paradigms, as well as the competence-performance distinction

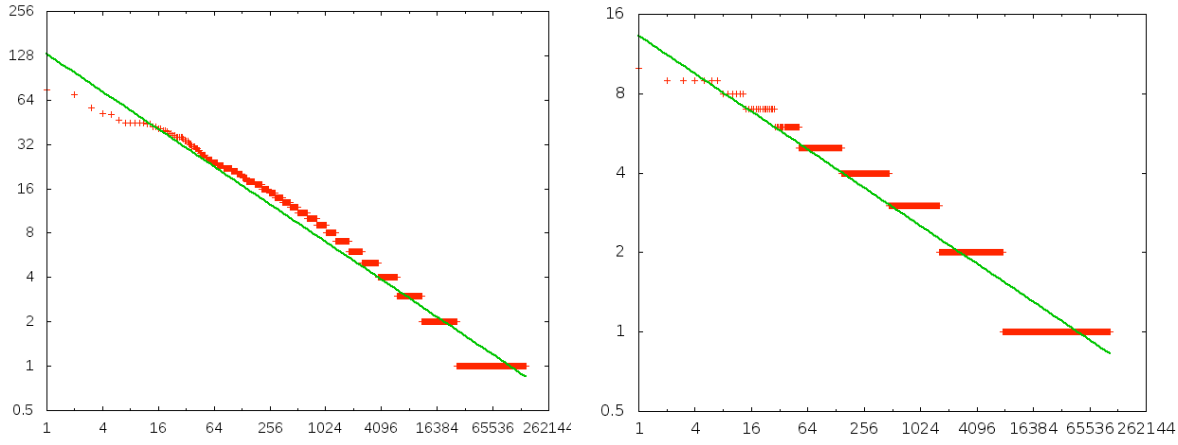
[4] BLARK stands for *Basic Language Resource Kit* and is characterized as "the minimal set of language resources that is necessary to do any precompetitive research and education at all" (KRAUWER, 2003, p. 4).

(section 2.1); practical and theoretical aspects of word-sense granularity in lexical description (section 2.2); and the treatment of multi-word expressions (section 2.3) and constructions (section 3) in LT and linguistics.

2.1. Zipf's law and linguistic description

In designing LT resources and applications, we cannot ignore Zipf's law (ZIPF, 1949), which rears its head in all kinds of contexts where large volumes of linguistic data are to be processed and described exhaustively. Very abstractly, Zipf's law says that there will be a few classes (e.g., corpus word types) with a large number of members and many classes with only one member (*hapax legomena* in a corpus), and everything in between. Distributions of linguistic phenomena are "heavy-tailed" (JÄGER, 2012); they typically display a "large number of rare events" (BAAYEN, 2001).

If we plot rank and frequency in a log-log coordinate system, with a perfect Zipfian distribution the points should form a downward-sloping straight line. Figure 1 shows the number of senses per base form in Princeton WordNet 3.0 – PWN (FELLBAUM, 1998) – and the SweFN++ pivot lexicon SALDO, together with a best-fit Zipfian distribution line. Even though the most polysemous base form in PWN has an order of magnitude more senses than the most polysemous base form in SALDO, the distributions are very similar and approximately Zipfian.



1a 1b

Figure 1: Senses per base form in PWN 3.0 (panel 1a) and SALDO (panel 1b)

Since these distributions are pervasive in language, it will come as no surprise that the inflectional component of the Swedish lexical macro-resource shows clear evidence of this. The inflectional component is structured into *inflectional patterns*, or *paradigms*, characterized as sets of bundles of morphosyntactic features conventionally expressed inflectionally in a language. For instance, Swedish nouns conventionally (maximally) express the following bundles – or combinations – of morphosyntactic features (illustrated with the word *bil* 'car'):

bil – singular indefinite nominative	*bilar* – plural indefinite nominative
bilen – singular definite nominative	*bilarna* – plural definite nominative
bils – singular indefinite genitive	*bilars* – plural indefinite genitive
bilens – singular definite genitive	*bilarnas* – plural definite genitive

There may be subclasses or individual cases of the main parts of speech which express fewer – or, in rare individual cases, more – such combinations, e.g. nouns appearing only in the singular or only in the plural.

If we want to be able to use the inflectional component for automatic text analysis, we cannot ignore singleton "paradigms", i.e., "inflectional classes/patterns" with only one member. In true Zipfian fashion, they make up a large chunk of all inflectional classes, and at least some of them belong to words with high text frequency. At present there are well over one thousand different inflectional patterns represented in the lexicon (1,329 on the most recent count). Among these are many singleton "patterns". In many cases, these are the irregular words of traditional grammar. Surprisingly often, however, the source of plenty is another, viz. variation. We often find that a particular combination of morphosyntactic features – a particular position in a paradigm – for a word or small group of words can be filled by more than one form, i.e. realized in more than one way. Such cases are legion, e.g., the three alternative forms *himmeln*, *himlen*, *himmelen* 'heaven sg def nom' (citation form *himmel*), of a word which in all other respects follows the inflectional pattern designated as "nn_2u_nyckel", which includes words like *nyckel* 'key', *åker* 'field', *öken* 'desert', *hummer* 'lobster'. This pattern allows only for the first of the three variants shown above for the singular nominative definite form of *himmel*, viz. the form made by affixing an *-n* to the citation form.

There is an interesting theoretical issue lurking here. For Wurzel (1989, p. 57), an inflectional class (which he uses as a technical term) must have more than one or even just a few members, although he is not prepared to commit himself to a specific lower limit. A practically useful computational lexicon should in any case specify the morphological behavior of individual words as accurately as possible. In the macro-resource, this behavior is encoded uniformly for all words – in the form of a unique identifier for each inflectional pattern[5] – i.e. in the lexicon we do not make a distinction between inflectional classes and individual cases in Wurzel's sense. This task is relegated to the computational morphological component, where a mapping is made between the inflectional patterns and regular, subregular and idiosyncratic inflectional descriptions. However, it is not difficult to get a picture of which inflectional patterns are general and which idiosyncratic. As expected from a Zipfian distribution, a small number of patterns account for the majority of entries. There are 25 inflectional patterns each accounting for more than 1,000 entries in the morphological component of the macro-resource, which together cover 75% of all entries.

Against this background, what is the theoretical status of linguistic generalizations? Do they belong mainly to linguistic intuition – but not to the language 'system' – and emerge out of a human propensity for (extrinsic) generalization? This could also account for the noted unreliability of elicited grammaticality judgments (cf. SCHÜTZE, 1996), and the difference

[5] The identifiers were designed to have some internal structure for the benefit of humans working with the lexicon. We cannot for reasons of space go into any details here, but just to give the reader a flavor of how identifiers are built up: The identifier "nn_3u_film" conveys the information that this is a third declension ("3") non-neuter (uter) gender ("u") common noun ("nn") inflected like the noun *film*.

between implicit and explicit knowledge about language recognized in second-language acquisition research (ELLIS, 2008). With such rampant idiosyncrasy in language, does not Ockham's razor require us to assume exemplar storage rather than generalization as the basic mechanism of language knowledge, as, e.g., in usage or exemplar-based linguistic models (BYBEE, 2013), or the various lazy-learning paradigms (BAAYEN, 2003) such as memory-based learning (DAELEMANS; VAN DEN BOSCH, 2010)? In any case, we cannot afford the idealizations required for many existing linguistic generalizations in this area if we are to obtain satisfactory practical coverage.

In adding the morphological information to the macro-resource, we have used existing grammatical descriptions of Swedish inflectional morphology – above all the two Swedish reference grammars published by the Swedish Academy (TELEMAN; HELLBERG; ANDERSSON, 1999; HULTMAN, 2003) – as well as the inflectional information provided in existing Swedish reference dictionaries. For practical and sometimes theoretical reasons we have deviated from these descriptions (which notably are not always consistent among themselves), on at least three counts:

(1) Our inflectional patterns are quite generous as to which forms are supposed to exist for a lemma. We thus subscribe to the notion of "potential form" which is inherent in the concept of inflectional paradigm, the general principle being that there should be a clear(ly statable) grammatical, semantic or pragmatic reason for us to postulate the absence of some form or forms in the paradigm of a lexical item. In practice, this is often the case with number in nouns, comparison in adjectives and certain adverbs, and past (passive) participles in verbs.

(2) A lexicon for language technology must lend itself to the analysis of arbitrary free text, e.g., on the internet, where we will find many word forms which are not accepted by normative dictionaries of the written language, but still recognizable as possible variant inflected forms of some existing lemma. Hence, the morphological component recognizes many attested (but not normative) forms.[6]

(3) We recognize MWEs as full-fledged lexical entries with their own inflectional behavior, as discussed in section 2.3 below.

The macro-resource is thus not normative, but rather strives to be maximally descriptive. At the same time the notion of inflectional patterns (inflectional classes, paradigms) contains a kind of normativity, namely that which is an irreducible element of linguistics itself, i.e., the formulation of lawlike generalizations about our languages. It is also a recognition of the fact that, however large a corpus we collect, we will never see all the inflected forms of all the entries in our lexicon, not even in a morphologically challenged language like Swedish (FORSBERG; HAMMARSTRÖM; RANTA, 2006; KETTUNEN; AIRIO; JÄRVELIN, 2007).[7]

At the same time we know as language users that some forms of some words are not only not attested, but actually non-attestable, e.g. the past participle forms of some verbs, or comparative and superlative forms of (past participle-like) Swedish adjectives in *-ad* (e.g., *långfingrad* 'long-fingered'). The reasons for the lack of some forms in a paradigm can be various, semantic or formal (the latter seems to be the case for the adjectives in *-ad*), but paradigms can also have "holes" in them for completely idiosyncratic reasons (HETZRON,

[6] Since SALDO has no information about which words or forms are nonstandard – or about domain, style and formality level – it is not immediately usable as, e.g., a spelling checking dictionary.

[7] On the other hand, this no more to be expected than you would expect at some point to have seen "all the sentences of the language" as you collect more and more text.

1975). We take this into account to some extent, but we have preferred to err on the side of generosity in unclear cases, which means that our morphological description probably overgenerates. If the lexicon is used in language technology applications for analysis, this is not a problem, as long as a potential but impossible form does not coincide with an actual other form. The problem of dealing with this if the lexicon is to be used in natural language generation applications is left for future work.

2.2 Word senses

The initial plan was for all component resources to share the standardized form and sense descriptors of SALDO, which thus would serve as the formal interlinking mechanism of the macro-resource. Of course, for some of the historical lexica this was patently unrealistic already at the outset, since many of the words appearing, e.g., in Sweden's medieval law texts have no modern counterpart. However, even in the work on integrating the modern resources, this has turned out to be far from simple.

In LT work it has long been recognized that too fine-grained word sense inventories – such as the 59 senses of the verb *break* in Princeton WordNet[8] – are difficult to distinguish reliably to machines and people alike, with the possible exception of highly trained lexicographers (KILGARRIFF, 1997; HANKS, 2000). This implies that an optimal lexical resource for LT should be able to provide a higher level of abstraction in its word sense represesentation than PWN-type lexicons. In our work, this issue has cropped up mainly in the use of SALDO sense PIDs as lexical units (LU) in the emerging Swedish FrameNet (SweFN). For the first half of the project, SweFN has been compiled mainly manually, by an experienced lexicographer and a group of computational linguists, using the English Berkeley FrameNet (BFN) as the point of departure. In their work, the compilers use existing SALDO word senses as far as possible, but they also propose new word senses for those cases where they feel that this is required.

Initially, this seemed unproblematic, and simply a way of discovering and adding missing items to SALDO, which, like every lexical resource, is never complete. With time it became clear, however, that there was a fundamental division in the project group between "splitters" and "lumpers", and furthermore that the splitters were motivated by at least two reasons, namely *lexicographic tradition* and *cross-lingual transfer*.

Cross-lingual transfer is due to our using the BFN frameset and translating it into Swedish, and manifests itself as a SALDO word sense appearing as an LU candidate in several frames, revealing a sense distinction made in English but possibly not in Swedish. For this situation, there are in principle three solutions: (1) Linguistic tests will reveal that the distinction is valid for Swedish, too, but not expressed as a distinct word, so a word sense should be added to SALDO; (2) The frame structure is modified, most likely by postulating a new Swedish frame; (3) The restriction that an LU can appear in only one frame is lifted. However, adopting the last solution would turn the resulting structure into something formally different from a framenet – which requires distinct word senses to be postulated for a lemma appearing in more than one frame (RUPPENHOFER ET AL., 2010, ch. 1) – and is consequently an option with potentially far-reaching repercussions.

[8] PWN simply follows lexicographic tradition here; at <http://dictionary.reference.com/browse/break> (based on the *Random House Dictionary*) we find 68 senses for the verb *break*.

What may still speak in favor of this third alternative is the fact that most of these problems do not apply to the frame distinctions as such, but to individual words. For instance, there are several words that fit both the Suasion and the Attempt_Suasion frame, or both the Suitability and Compatibility frame. It is of course possible to make these distinctions in Swedish, but it is not always necessary; and, for some words, the difference is a matter of vagueness rather than ambiguity. The currently available alternative to assuming (possibly unmotivated) polysemy in these cases is to choose one of the frames over the other, thereby (over-) restricting the semantic characterization of the word. This kind of problem does not only concern cross-lingual transfer, but applies within a single language as well.

However, the more interesting and challenging case, at least from an LT perspective, is that of lexicographic, i.e., descriptive, tradition, since this is intimately tied up to how we conceptualize language and the linguistic knowledge involved in understanding and producing language. Traditional lexicography leans toward the "splitting" camp, which in turn seems to be predicated on a strong form of compositionality, in the extreme cases including "lexical items" defined through *idiosyncratic decomposition* of expressions which "are decomposable but coerce their parts into taking semantics unavailable outside the [multi-word expression]" (BALDWIN ET AL., 2003, p. 89). Even in the ordinary case, strong compositionality means, roughly, that there is no scope for rich general rules of inference in interpreting linguistic expressions; rather, words should carry as much as possible of their interpretation in each specific context with them, which potentially leads to as many meanings as there are distinct contexts. This precludes the positing of more general "meaning potentials" (HANKS, 2000, 2013) or even overlapping senses (ERK, 2010) for lexical units, which would have to rely on a sophisticated and information-rich interpretation procedure on the part of the language user, perhaps involving something like mutual constraint satisfaction. Empirical results from LT research point to the usefulness for NLP applications of both more coarse-grained and overlapping word senses (ERK, 2010).

Hence, an issue still to be resolved in the macro-resource is how to reconcile the conflicting requirements of traditionally organized lexical resources on the one hand and the practical needs of NLP applications on the other – e.g., the need for different degrees of word sense granularity – in a way which is both practically and theoretically satisfying. Note that this definitely concerns the SweFN in a more narrow sense, too, since the frameset in a particular framenet is defined by (word) meanings, and different conceptions of what constitutes a word meaning will lead to different framenet organizations.

A general solution that we have discussed is to lift the requirement that the linking relation between SALDO and other lexical resources always be identity. Instead, we could add. e.g., *supersense* (broader concept) and *subsense* (narrower concept) relations where needed. This would allow us to keep the formal requirement in FrameNet that different frames must have disjoint sets of lexical units, while still keeping the resulting polysemy local to SweFN.

2.3 Multi-word expressions in SweFN++: no pain, no gain

SALDO, the hub of the macro-resource, at present contains almost 6,000 multi-word expressions (MWEs), making up just under 5% of the entries. However, out of new entries being added to SALDO, the share of MWEs is growing steadily. We feel that any serious large-scale lexical resource must have a principled and practical way of dealing with MWEs,

even if some lexicographers and linguists feel that they are "the 'black hole' of semantics" (BARANOV; DOBROVOL'SKIJ, 2008, p. 567).

On the one hand, MWEs have been characterized as a "pain in the neck for NLP" (SAG et al., 2001). On the other hand, the ambiguity that has always plagued automatic syntactic analysis, where even relatively short sentences may have tenths or even hundreds of analyses, can be greatly reduced if we are able to automatically identify MWEs in authentic text and treat them on a par with single words. Kokkinakis (2008) has demonstrated that the identification of complex terminology and named entities simplifies a following syntactic analysis considerably, and, e.g., Attardi and Dell'Orletta (2008) and Gadde et al. (2010) have shown how pre-identification of different types of local continuous syntactic units may improve a subsequent global dependency analysis.

Consequently, a good deal of thought has gone into accommodating MWEs in SALDO in a way that is both practical and linguistically satisfactory. At the moment, we distinguish 3 different kinds of MWEs. These types are practical to distinguish for (written) Swedish, and no particular claim is made here as to their universality:[9]

(1) **Continuous MWEs**; these correspond to the "fixed expressions"' and "semi-fixed expressions"' of Sag et al. (2001). They may exhibit any combination of internal and external inflection, but the order of the component words is fixed and other sentence material (other words) never intervenes between the parts of the MWE. For example, the MWE *enarmad bandit* 'slot machine', literally 'one-armed bandit' has the indefinite nominative plural *enarmad**e** bandit**er***.

(2) **Discontinuous MWEs**; these are, by and large, the "syntactically-flexible expressions" of Sag et al. (2001). In these, other sentence material may intervene. The prototypical examples of these MWEs are particle (or phrasal) verbs, and support-verb (or light-verb) constructions.

(3) **Constructions**; these are the kinds of phenomena that are studied under the heading of *construction grammar* (e.g., FILLMORE; KAY; O'CONNOR, 1988; GOLDBERG, 1995; HOFFMANN; TROUSDALE, 2013), especially partially schematic constructions, i.e., syntactic fragments (or templates) with one or more slots for items specified as to, e.g., part of speech (in a dependency framework) or phrase type (in a constituency framework), and semantic type.

Constructions and the linguistic and computational issues connected with their description are the focus of the new SweCxn project (see section 3 below), but the first two kinds of MWEs are already fully integrated descriptively in the SALDO morphology, and partly integrated w.r.t. morphological processing. For these MWEs, we simply assume "word-like" or lexical semantics. The trivial observation that a single orthographic word in one language often corresponds to a conventionalized MWE in some other language supports this assumption. The fact that such MWEs sometimes have compositional, non-MWE readings in addition to the conventionalized/lexicalized one is in principle no more of a *theoretical* problem than when a lexicalized compound also has a compositional reading in a language like German or Swedish (but it may of course present a very concrete *practical* problem for

[9] It seems to us that – at least in the computational linguistics literature – "multi-word expression" is a pre-theoretical, essentially negative characterization, which to boot is dependent on the vagaries of individual orthographies. Discussions of MWEs in computational linguistics rarely refer to the vast linguistic literature on the problems connected with defining the entity *word* in a cross-linguistically reasonable way (see, e.g., ANDERSON, 1985; AIKHENVALD 2007), and we are not aware of any typological studies to establish which kinds of MWEs there are cross-linguistically or how frequent they are across the world's languages.

NLP). Cf. the Swedish compound *husbil* 'camper, trailer', but also compositionally 'house car' (e.g., it could be used to refer to a builder's van with a drawing of a house on the side), or the (already mentioned) Swedish MWE *enarmad bandit* 'slot machine', but also compositionally 'one-armed bandit'. How often is a lexicalized MWE used with the alternative compositional reading? We don't know.[10] In fact, we don't know this about compounds either (at least we are not aware of any linguistic studies investigating this), only that the compositional reading is always possible if all the component parts of the compound are also living words in the language. The facts that a compositional reading of a conventionalized compound normally has to be forced, and is generally used for humorous effect, indicate that this is not the normal state of affairs.

MWEs are consequently not characterized as any kind of phrases, i.e., with an internal syntactic structure. This is completely analogous to how we treat structurally complex single-word items, such as compounds or derived words. We do not let the compound *husbil* inherit its formal characteristics from its last member – even though we could do this – but rather provide it with its own inflectional information, as if it were a non-derived word. This is not to deny the value of such a description, which is what we expect to find in linguistic works on word-formation. We simply follow the usual practice of lexicography, where the formal structure of complex words – compounds or derivations – is not normally made explicit in the lexicon. With the possible exception of constructions with variable constituents (cf. section 3 below), we see no compelling reason to treat MWEs differently.

Consequently, we treat the first kind of MWEs in the list above formally as "words with spaces", and subject to general morphology-like inflectional processes. We have yet to encounter some formal mechanism in such MWEs, which we would not also expect a general (inflectional) morphological processor to handle ('internal' inflection, discontinuous dependencies among word components, multiple discontinuous exponence, coreference to word-internal components, etc.; see, e.g., NIDA, 1949; JENSEN, 1990).

With the second kind of MWEs, things become a bit more complex. The components of MWE verbs (and sporadically MWEs from other parts of speech) can appear discontinuously in clauses. In theory, the intervening items can be arbitrarily long, but in practice they tend to be short, typically one to two words, as in the following example with the MWE *rycka upp sig* ('pull oneself together'):

(1) *Då **ryckte** hon verkligen **upp sig**.*
then **pulled** she really **up herself**.
'Then she really pulled herself together.'

However, we still consider these kinds of MWE verbs to fall on the lexical side of the fence. The description is in terms of word semantics, and the formal treatment is one of "sequences with holes". The 'morphological analysis' component will have to pull a heavier load and interact more closely with the syntactic analysis, which may be desirable anyway, for independent reasons (see the next section).

[10] Although it has been claimed in the literature that "corpus studies suggest that [...] the institutionalization of an idiomatic meaning is typically associated with non-use of possible literal meanings" (MOON, 2000, p. 102).

3. Constructions and the constructicon

Given a distinction between grammar (or more precisely: syntax) and lexicon, the first two types of MWEs belong to the lexical domain, and may be accounted for by the descriptive apparatus of lexicography. Constructions, however, incorporate both lexical and grammatical properties and therefore require other methods. The constructions we are mostly concerned with here are the semi-general, partially schematic kind of patterns that are typically neglected in linguistic and language technology resources both. Such resources are either grammars or lexica, designed to account for either grammatical or lexical phenomena. Therefore, neither is well equipped to deal with patterns involving a combination of lexical, syntactic, morphological, semantic, and pragmatic features (in spoken language, also prosody).

For instance, consider the Swedish Aux Aux construction (cf. LINELL; NORÈN, 2009), as illustrated in the following dialogue:

(2) *–Kan du göra det? –Tja, kan kan jag säkert; frågan är om jag vill.*
–Can you do it? –Well, can can I surely; the-question is if I want.
'–Can you do it? –Well, I surely CAN; the question is whether I want to'

This construction is characterized by two identical auxiliaries, picking up the same auxiliary from the preceding utterance and questioning its relevance. Thus, the Aux Aux construction is reactive in the sense that it depends on specific properties of the preceding utterance. To account for such a construction, one has to both represent its internal structure (two identical auxiliaries) and appeal to its context.

From the viewpoint of the macro-resource, constructions are discontinuous MWEs with variables. In addition to the possibility of intervening items, parts of the constructions themselves may vary – within certain restrictions, such as belonging to a particular linguistic category. In the case of Aux Aux, it may be instantiated by any auxiliary as long as the other constraints of the construction are met. Still, to any system capable of identifying auxiliaries, these are single units, in principle equivalent to words.

However, constructions including constituents merely specified as nominal or adjectival, for example, are a different matter. Such constituents may consist of simple nouns or adjectives, respectively, but also of larger phrases, themselves including various types of modifiers. To handle constructions of this kind, it is necessary to also be able to identify and analyze noun phrases, adjective phrases etc. This goes beyond lexical resources and requires some kind of syntactic analysis tools. In linguistics, this is common practice, but for language technology it is a nontrivial problem to mix lexical and syntactic processing in this way.

One option is to develop a larger constructicon, including general phrase types and other basic syntactic constructions. Another is to combine the constructicon with a syntactic parser. In effect, both options may amount to pretty much the same thing. In an LT application, parsers tend to generate massive ambiguity, and so do many construction descriptions. It is therefore highly desirable to combine the tools in a way in which they reduce their respective ambiguities instead of multiplying them. How this can be achieved is one of the challenges faced by a project such as this.

It should be noted that constructicons are a new kind of resource. Linguists have of course studied grammatical constructions for a long time, but typically in the form of detailed case studies. Existing large-scale resources are either grammars or lexica, and construction-oriented resources with comprehensive coverage are still lacking. Hence, development of a constructicon requires new methodology. To our knowledge, all constructicon endeavors initiated so far, including the Swedish one, are designed as additions to the FrameNet of the language in question. Hence, a brief comparison is in order:

Frames are essentially semantic units, defined by their meaning, whereas constructions are defined by their meaning *and* their formal structure. Accordingly, frame elements are typically semantic roles, whereas construction elements are also syntactic constituents. Hence, constructions and construction elements require not only a definition but also a structural representation. Again, the crucial difference between constructional and lexical resources is that construction entries cannot ignore the internal formal structure.

Conclusions

In this paper, we have described our work on a large-scale integrated lexical macro-resource for Swedish language technology, to which we are now adding a Swedish constructicon. In this presentation we have focused on some areas where close encounters between the requirements of LT applications – 'full accountability' and relentless formalization – and those of linguistic descriptions – formulation of valid generalizations and acknowledgement of a rich methodology and a long descriptive tradition – raise interesting and difficult questions about the relationship among attested language, linguistic intuitions and linguistic description. We are convinced that our attempts to address these questions will enrich both fields in a longer perspective.

Acknowledgements

The research presented here was supported by the Swedish Research Council (grant agreement 2010-6013), by the Bank of Sweden Tercentenary Foundation (grant agreement P12-0076:1), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken, by the European Commission through its support of the META-NORD project (under the ICT PSP Programme, grant agreement no 270899), and by a Swedish Academy Fellowship for Benjamin Lyngfelt, sponsored by the Knut and Alice Wallenberg Foundation.

References

ADESAM, Y.; AHLBERG, M.; BOUMA, G. bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa ... Towards lexical link-up for a corpus of Old Swedish. In: *Proceedings of KONVENS 2012*, Vienna, 2012. p. 365–369.

AHLBERG, M.; BOUMA, G. A best-first anagram hashing filter for approximate string matching with generalized edit distance. In: *Proceedings of COLING 2012*. Mumbai: ACL,

2012.

AIKHENVALD, A.Y. Typological distinctions in word-formation. In: SHOPEN, T. (Ed.). *Language typology and syntactic description (2nd ed.). Volume III: Grammatical categories and the lexicon*. Cambridge: Cambridge University Press, 2007. p. 1–65.

ANDERSON, S.R. Inflectional morphology. In: SHOPEN, T. (Ed.). *Language typology and syntactic description (1st ed.). Volume III: Grammatical categories and the lexicon*. Cambridge: Cambridge University Press, 1985. p. 150–201.

ATTARDI, G.; DELL'ORLETTA, F. Chunking and dependency parsing. In: *LREC Workshop on Partial Parsing*. Marrakech: ELRA, 2008. p. 27–32.

BAAYEN, R.H. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers, 2001.

BAAYEN, R.H. Probabilistic approaches to morphology. In: BOD, R.; HAY, J.; JANNEDY, S. (Eds.). Probabilistic linguistics, Cambridge, Mass.: MIT Press, 2003. p. 229–287.

BALDWIN, T.; BANNARD, C.; TANAKA, T.; WIDDOWS, D. An empirical model of multiword expression decomposability. In: *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo: ACL, 2003. p. 89–96.

BARANOV, A.N.; DOBROVOL'SKIJ, D.O. *Aspekty teorii frazeologii*. [Aspects of the theory of phraseology]. Moscow: Znak. 2008.

BORIN, L. Med Zipf mot framtiden – en integrerad lexikonresurs för svensk språkteknologi. [With Zipf into the future – an integrated lexical resource for Swedish language technology]. *LexicoNordica*, vol. 17, p. 35–54, 2010.

BORIN, L.; DANNÉLLS, D.; FORSBERG, M.; GRONOSTAJ, M.T.; KOKKINAKIS, D. The past meets the present in Swedish FrameNet++. In: *14th EURALEX International Congress*. Leeuwarden: EURALEX, 2010. p. 269–281.

BORIN, L.; FORSBERG, M. Something old, something new: A computational morphological description of Old Swedish. In: *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakech: ELRA, 2008. p. 9–16.

BORIN, L.; FORSBERG, M. A diachronic computational lexical resource for 800 years of Swedish. In: SPORLEDER, C.; VAN DEN BOSCH, A.; ZERVANOU, K. (Eds.), *Language technology for cultural heritage*, Berlin: Springer, 2011. p. 41–61.

BORIN, L.; FORSBERG, M.; AHLBERGER, C. Semantic search in literature as an e-humanities research tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature. In: *NODALIDA 2011 Conference Proceedings*, Riga, 2011. p. 58–65.

BORIN, L.; FORSBERG, M.; KOKKINAKIS, D. Diabase: Towards a diachronic BLARK in support of historical studies. In: *Proceedings of LREC 2010*, Valletta: LREC, 2010. p. 35–42.

BORIN, L.; FORSBERG, M.; LÖNNGREN, L. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In: *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, Uppsala: Uppsala University, 2008. p. 21–32.

BORIN, L.; FORSBERG, M.; LÖNNGREN, L. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, forthcoming. DOI 10.1007/s10579-013-9233-4
. Published online 31 May 2013.

BYBEE, J.L. Usage-based theory and exemplar representations of constructions. In: HOFFMAN, T.; TROUSDALE, G. (Eds.), *The Oxford handbook of construction grammar*, Oxford & New York: Oxford University Press, 2013. p. 49–69.

DAELEMANS, W.; VAN DEN BOSCH, A. Memory-based learning. In: CLARK, A.; FOX, C.; LAPPIN, S. (Eds.), *Handbook of computational linguistics and natural language processing*, Oxford: Wiley-Blackwell, 2010. p. 154–179.

ELLIS, N. Implicit and explicit knowledge about language. In: CENOZ, J.; HORNBERGER, N.H. (Eds.), *Encyclopedia of language and education. Volume 6: Knowledge about language*, Berlin: Springer. p. 1–13.

ERK, K. What is word meaning, really? (And how can distributional models help us describe it?). In: *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, Uppsala: ACL, 2010. p. 17–26.

FELLBAUM, C. (Ed.) *WordNet: An electronic lexical database*. Cambridge, Mass.: MIT Press, 1998.

FILLMORE, C.; KAY, P.; O'CONNOR M. Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, vol. 64, p. 501–538, 1988.

FORSBERG, M.; HAMMARSTRÖM, H.; RANTA, A. Morphological lexicon extraction from raw text data. In: *FinTAL 2006*, Berlin: Springer, 2006. p. 488–499.

GADDE, P.; JINDAL, K.; HUSAIN, S.; SHARMA, D.M.; SANGAL, R. Improving data driven dependency parsing using clausal information. In: *HLT: The 2010 Conference of the NAACL*, Los Angeles: ACL, 2010. p. 657–660.

GOLDBERG, A.E. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press, 1995.

HANKS, P. Do word meanings exist? *Computers and the Humanities*, vol. 34 n. 1–2, p. 205–215, 2000.

HANKS, P. *Lexical analysis: Norms and exploitations*. Cambridge, Massachusetts: MIT Press, 2013.

HETZRON, R. Where the grammar fails. *Language*, vol. 51, p. 859–872, 1975.

HOFFMANN, T.; TROUSDALE, G. (Eds.). *The Oxford handbook of construction grammar*. Oxford & New York: Oxford University Press, 2013.

HULTMAN, T. *Svenska Akademiens språklära*. [The Swedish Academy short reference grammar]. Stockholm: Norstedts ordbok, 2003.

JÄGER, G. Power laws and other heavy-Tailed distributions in linguistic typology. *Advances in Complex Systems*, vol. 15 n. 3–4, 2012.

JENSEN, J.T. *Morphology: Word structure in generative grammar.* Amsterdam: John Benjamins, 1990.

KETTUNEN, K.; AIRIO, E.; JÄRVELIN, K. Restricted inflectional form generation in management of morphological keyword variation. *Information Retrieval*, vol. 10 n. 4–5, p. 415–444, 2007.

KILGARRIFF, A. I don't believe in word senses. *Computers and the Humanities*, vol. 31 n. 2, p. 91–113, 1997.

KOKKINAKIS, D. Semantic pre-processing for complexity reduction in parsing medical texts. In: *Proceedings of the 21th Conference on the European Federation for Medical Informatics (MIE 2008)*, 2008.

KRAUWER, S. The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. In: *Proceedings of SPECOM 2003*, Moscow. 2003.

LINELL, P.; NORÉN, K. "Vågar vågar ni väl men…" – en reaktiv konstruktion i svenskan. ["Dare dare you, I guess, but... " – a reactive construction in Swedish]. *Språk och stil NF*, vol. 19, p. 72–104, 2009.

LYNGFELT, B.; BORIN, L.; FORSBERG, M.; PRENTICE, J.; RYDSTEDT, R.; SKÖLDBERG, E.; TINGSELL, S. Adding a constructicon to the Swedish resource network of Språkbanken. In: *Proceedings of KONVENS 2012*, Vienna, 2012. p. 452–461.

MOON, R. Lexicography and disambiguation: The size of the problem. *Computers and the Humanities*, vol. 34 n. 1–2, p. 99–102, 2000.

NIDA, E.A. *Morphology: The descriptive analysis of words*. Ann Arbor: University of Michigan Press, 1949.

REITER, E. The shrinking horizons of computational linguistics. *Computational Linguistics*, vol. 33 n. 2, p. 283–287, 2007.

RUNDGREN, F. The computational paradigm. *Fenno-Ugrica Suecana*, vol. 5, p. 235–248, 1982.

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M.R.L.; JOHNSON, C.R.; SCHEFFCZYK, J. FrameNet II: Extended theory and practice. (Printed September 14, 2010.) <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>. 2010.

SAG, I.A.; BALDWIN, T.; BOND, F.; COPESTAKE, A.; FLICKINGER, D. Multiword expressions: A pain in the neck for NLP. In: *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Berlin: Springer, 2001. p. 1–15.

SCHÜTZE, C.T. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

TELEMAN, U.; HELLBERG, S.; ANDERSSON, E. *Svenska Akademiens grammatik, 1–4*. [The Swedish Academy grammar, 1–4]. Stockholm: Norstedts ordbok, 1999.

WINTNER, S. What science underlies natural language engineering? *Computational Linguistics*, vol. 35 n. 4, p. 641–644, 2009.

WURZEL, W.U. *Inflectional morphology and naturalness*. Dordrecht: Kluwer, 1989.

ZIPF, G.K. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley, 1949.

Received: 12/02/2013
Accepted: 13/05/2013
Published: 31/10/2013