The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 21–61

A Contrastive Lexical Description of Basic Verbs

Examples from Swedish and Czech

Silvie Cinková

Abstract

This paper aims at a lexical description of frequent uses of frequent lexical verbs in Swedish on the background of Czech, with some implications for the lexical description of such verb uses verbs in general. It results in a draft of a production lexicon of Swedish frequent verbs for advanced Czech learners of Swedish, with focus on their uses as light verbs.

The introductory sections (1 and 2) discuss semantic shifts in highly frequent lexical verbs, whose most literal or 'primary' uses express motion, location, or physical control; e.g. *stand, put, go, hold*. These verbs are called *basic verbs*, which is a term coined by Viberg (Viberg, 1990) that suggests that they typically denote events belonging to *basic level categories* described by Lakoff (Lakoff, 1987). The 'literalness' of verb uses is judged according to how much they are the ones speakers pick first to illustrate the meaning of that given verb (*cognitive salience*, a term coined by Hanks in (Hanks, forthcoming). Hanks pointed out an interesting discrepancy between the cognitive salience and the actual frequency of a given verb usage in large corpora. This discrepancy is extremely significant in basic verbs. Some of their uses exhibit such a low cognitive salience, that they are not even noticed by native speakers. This has consequences in second-language acquisition. Foreign learners, even the advanced ones, often lack competence in using the most frequent lexical verbs of the second language in their most frequent patterns.

Basic verbs often act as light verbs. Sections 3 to 7 are dedicated to light verbs and light verb constructions. Section 8 discusses the morphosyntactic variability in predicate nouns (i.e. the nominal components of light verb constructions) and their possible semantic impact on the entire light verb construction.

Different aspects of polysemy of basic verbs are dealt with by contrasting Swedish examples to Czech in Section 9. Special attention is paid to uses of basic verbs that denote relations between abstract entities. Section 10 focuses on grammaticalizing uses of lexical verbs. It gives a Swedish example *of context-induced reinterpretation* – an interesting semantic shift that often leads to grammaticalization.

Please cite this article as: Silvie Cinková, A Contrastive Lexical Description of Basic Verbs: Examples from Swedish and Czech. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 21–61.

All the aspects of basic verbs discussed in Sections 1–10 are integrated in a structure of a Swedish-Czech lexicon, which captures verbs and predicate nouns in two respective interlinked parts. Sections 11-14 give its detailed decription.

1. Introduction

Probably every human language operates with a set of very frequent lexical verbs that are primarily perceived as verbs denoting location, motion, or physical control over something; e.g. *stand, go, put, keep, get*, etc. Their literal meanings (meanings not derived by metaphorical transfers) are very *cognitively salient* (Hanks, forthcoming); i.e. they are intuitively first associated with the verb ("what we think words mean"). For instance, the most cognitively salient meaning of *go* would have to do with spatial motion. On the other hand, the research on large text corpora reveals an interesting fact that the cognitive salience of a word usage does not necessarily correspond to its *social salience* ("the actual meanings that we use"), such as the most socially salient meaning of *go* can possibly be the future tense *(to be going to)*, which we would hardly consider the "most typical" meaning of *go*.

The socially salient meanings arise regularly through metaphorical shifts and processes of semantic deployment or generalization, in which the given verb loses or generalizes some of its semantic features (cf. Bybee et al., 1994; Heine et al., 2001) to expand its collocability. When the general semantic feature (in other words actually cognitive category) is relevant for an entire class of lexemes (such as future is for verbs), the distribution of the given lexeme (here a verb or its particular morphosyntactic form) gradually ceases to be limited by the collocability of the primary meaning, and the lexeme step by step turns into a universally usable language element. At that stage, it is perceived as a part of the grammar system. This process is called *grammaticalization* (Hopper, 1987) or *grammaticization* by some (Bybee, 1985; Bybee et al., 1994). It is a gradual process that spreads from isolated words, collocations, and phrases. Grammar is to be understood as, in Hopper's terms, "a real-time, social phenomenon", which is "always in process but never arriving, and therefore emergent", and not "the only, or even the major, source of regularity, but instead grammar is what results when formulas are rearranged, or dismantled and re-assembled, in different ways."

Evidently, there is a transition area between phraseology and syntax. There are millions of idiomatic expressesions that arose as new collocations or phrases constituted by a cluster of collocates as new sematic units in their own right. When the tightness of the collocation lies mainly in the cooccurrence of two autosemantic lexemes (e.g. *be in one's shoes*), rather than in the cooccurrence of one of several lexemes with a given morphosyntactic constellation in the environment (*keep + -ing*), it is well in place to refer to it as to a *lexicalized* expression, which is a phraseological term. Theoretically, we could make a closed list of idiomatic expressions occurring in a given language.

On the other hand, there are a number of systematically occurring collocations of lexemes and structural elements that are not easily captured by the phraseological approach; e.g. the colloquial structure *don't go doing something*, which intensifies the negative connotation of the event in question (or indicates it when the verb itself is stylistically neutral):

(1) *…that poor man probably gets compared to that character all the time. Don't go bothering him.* [COCA]

(2) *"It's okay," she said. "I'm fine. Don't go bothering about me when you've got Georgy lying here in this state."* [COCA]

(3) *"You'll do fine," he told her confidently on one of their walks along the Danube. "Just don't go marrying an Australian. I must have my little girl back someday."* [COCA]

(4) *Is it too much to ask, Jack, honey, that just once after we make love you don't go rushing off like there's a three-alarm fire?* [COCA]

(5) *"In the defense world, you're notified of bids, you negotiate a long-term relationship and you don't go knocking on doors and say, Here's our brochure…"* [COCA]

The actual meaning of this particular structure, in which the semantically heaviest part, namely the *-ing* form of the verb governed by *go*, is freely variable, is very general. An idiomatic expression, on the contrary, typically bears a complex meaning of its own and can either be paraphrased (e.g. *push out the daisies = be dead*) or related to a particular situation (*damn it! – you say it when you are very irritated, cheers – you say it when making a toast, or when informally thanking somebody, or when parting somebody informally*). This is, however, not the case when the semantically heavy part is just the one in the construction that can be replaced.

The collocational interplay between lexical items and structural elements that we are used to perceiving as syntax, is obscuring the borderline between grammar and lexis. For this phenomenon, Hoey (Hoey, 1998) re-used the term *colligation*, originally coined by J.R. Firth for *collocation*. Hoey (Hoey, 1998, quotation taken from Hunston, 2001) defines *colligation* the following way:

- The grammatical company a word keeps (or avoids keeping) either within its own group or at a higher rank.
- The grammatical functions that the word's group prefers (or avoids).
- The place in a sequence that a word prefers (or avoids).

Such observations have been made earlier; cf. Hunston in Hunston (2001, p. 15): "If we take seriously Sinclair's assertion that there is no longer any sense in distinguishing between lexis and grammar […], then the distinction between collocation and colligation to a large extent disappears. On the other hand, the term *colligation* is helpful in drawing attention to the fact that the evidence in many instances of naturally-occurring language can be used to explain behaviour that is traditionally associated with grammar. Just as the discipline called 'lexis' has been assisted

by corpus-based approaches to collocation, so the discipline 'grammar' benefits from corpus-based approaches to colligation."

An attitude that no longer makes a distinction between grammar and phraseology can be very useful in describing semantically depleted uses of common lexical verbs that have not achieved universal collocability with a clearly defined word class (be it part of speech or even a generally known semantic criterion such as 'animate nouns', 'verbs denoting states', etc.). Such patterns of uses do not often penetrate grammar textbooks, but are on the other hand in a way too vague to be described as multiword units in lexicons. They are often ignored by native speakers as 'untypical' uses. Their cognitive salience can be extremely low, while their social salience can be high at the same time; and that is why they deserve special attention.

2. Basic Verbs

This paper aims at a lexical description of socially, but perhaps not enough cognitively salient uses of frequent lexical verbs in Swedish on the background of Czech, with some implications for the lexical description of basic verbs in general. It results in a draft of a production lexicon of Swedish basic verbs[1] for advanced Czech learners of Swedish, with focus on their uses as light verbs (see Section 3).

Verbs possess the ability to bring entities into relations and create propositions. An analysis of the most frequent lexemes in Swedish (Viberg, 1990) shows an interesting fact: there are far fewer verbs in the language than there are e.g. nouns. The Swedish frequency dictionary (Allén, 1972) contains 39 486 nouns but about 8,5 times fewer verbs (4 649).

We reflect the manifold features of different entities by a vast amount of nouns at our disposal, whereas we evidently need a significantly smaller amount of verbs to describe the relations these entities enter. Besides, there is an evident preference for just a selection of verbs. Viberg (Viberg, 1990) observed in Swedish, in full accordance with Zipf's law, that almost one half (45.5%) of the verb occurrences is represented by the 20 most frequent verbs. Almost every second verb used in the language is then one of the top-twenty.[2] This implies that some verbs have an extreme potential to fit into many different contexts.

For these verbs, Viberg coins the label *basic verbs*. According to Viberg, they are characterized by the following features:

1. They are simple stems rather than derivations or compound words.
2. They have a phonologically simple form.

[1] a term coined by Viberg in Viberg (1990)

[2] The 20 most frequent nouns cover only 8.1% of noun occurrences, the 20 most frequent adjectives cover 24.2% of adjective occurrences and the 20 most frequent adverbs have similar rate as verbs – 42.1% of adverbial occurrences.

3. Their conjugated forms are often irregular.[3]
4. They occur in the respective languages with high frequency.
5. Typologically, they have a broad distribution (their equivalents exist in many languages)
6. They have many "secondary" meanings[4].
7. They have a significant potential to become grammatical markers.
8. They act as syntactic prototypes (i.e. they allow for many valency patterns and occur in more compound words and derivations).
9. They are preferred at the early stages of first as well as of second language acquisition.

For Swedish, the top 20 verbs are the following:

1. *är/vara (to be)*
2. *ha (to have)*
3. *kunna (can)*
4. *ska (shall)*
5. *få (to get)*
6. *bli (to become)*
7. *komma (to come)*
8. *göra (to do, to make)*
9. *finnas* (existential *to be*, lit. *to be found*. Similar to the German *es gibt*.)
10. *ta (to take)*
11. *säga (to say)*
12. *gå (to go)*
13. *ge (to give)*
14. *se (to see)*
15. *måste (must)*
16. *vilja (to want)*
17. *stå (to stand)*
18. *visa (to show)*
19. *böra (ought)*
20. *gälla (to apply, to be valid)*

For the purpose of this paper, only a subset of what Viberg calls basic verbs is analyzed: verbs that are able to act as light verbs[5]. Copula verbs and modal verbs are ignored.

[3]This can indicate that the respective forms are acquired by rote learning and remain further unanalyzed by speakers (cf. Bybee, 1985).

[4]Many studies on this have been published in the Scandinavian area also by other authors. Among others Ekberg (1993), Fenyvesi-Jobbágy (2003), Hansen (1974), Jakobsson (1996), Jensen (2000), Malmgren (2002), Pihlström (1988), Reuter (1986).

[5]see Section 3

As a rule, the verbs in question rank among the top 50 most frequent verbs. It is mostly verbs of spatial motion, location, and physical control. The most typical members of this group are *stå (stand), ligga (lie), sitta (sit)* and their causative counterparts *ställa, lägga,* and *sätta*, *ge (give), ta (take), hålla (hold/keep), gå (go), komma (come), göra (do/make), falla (fall), fälla* (causative to *falla*), *bjuda (offer), visa (show/exhibit), möta (meet/face)*, and *få (get)*. A list of potential light verbs was obtained earlier by extracting verb-noun collocations from the 20-million morphosyntactically tagged Swedish corpus PAROLE (Cinková, 2004).

3. Light Verb Constructions

Light verbs and light verb constructions (henceforth LVC's) are an interesting instance of semantic shifts in lexical verbs. Their numerous definitions set by many different linguists agree that LVC's [6]) consist of a lexical verb and a noun phrase and that it is the noun that carries the semantic weight. The verb is deprived of its original meaning. It only delivers morphosyntactic categories and, possibly, some semantic features to the resulting event description. Not seldom, the valency behaviour of the verb changes when the verb acts as a light verb. For instance, *give a sigh* has nothing to do with *giving* but with *sighing*, which is supported by the fact that *give* in this case obviously opens no addressee slot.

Lexical verbs which lose their concrete meaning when combined with abstract nouns and nominalizations and which occur in such combinations very productively, appear to be very common in modern European languages, but also beyond Europe, as already noted by R. Jakobson (for reference see Jelínek, 2003, p. 50). They were even observed in South-Asian languages (Butt, 2003), which are linguistically as well as culturally very distant from the European languages.

Butt in (Butt, 2003) claims that although light verbs potentially are a universal linguistic phenomenon, they have different structural features in the respective lan-

[6]This paper uses a term coined by Jespersen in Jespersen (1954), but they are also known under many other names. In English linguistics it is e.g. *support verbs, support verb constructions, expanded predicates, verbo-nominal phrases, delexical verbs, stretched verbs*.

German linguistics has studied *Funktionsverbgefüge* and *Funktionsverben* (also under different terms) intensively since the term was coined by von Polenz (Polenz, 1963). Interest in this issue rose especially with the onset of generative and transformational grammar (among others in Rothkegel's studies on fixed syntagms Rothkegel, 1973). To be mentioned are also at least Persson's studies on causativity (Persson, 1975;Persson, 1992), as well as the research in German as a foreign language (Helbig and Buscha, 1996 and Günther and Pape, 1976).

The terms, especially the German terms as *Funktionsverben, Nominalisierungsverben, verblasste Verben, Streckformen*, etc., cannot be used interchangeably. Some authors using the respective variants were observing only the combinations of a verb and its direct object, others only the combinations of a verb and its prepositional object. For a summarizing comparison of the light-verb related terms in German and English see e.g. Hanks et al. (2006).

guages[7]. Hence all syntactic tests for defining light verbs and light verb constructions are language-specific (p. 24 in the web-released manuscript of Butt, 2003). E.g., in Germanic languages, the following criteria are commonly quoted:

- Light verb constructions with the predicate noun in the position of the direct object cannot be passivized.
- The predicate noun cannot be replaced with an anaphoric expression.
- There should be at least an option of the predicate noun to occur without a determiner (a criterion applied to Swedish, see Dura, 1997).

Few verbs are light under all circumstances: there belong those that combine only with nominalizations or event nouns, such as *perform, carry out*[8]. The syntactic behaviour of the word combination is an important clue for all verbs that can either act as lexical verbs **or** as light verbs according to their context. However, the syntactic criteria do not apply 100%.

Hanks et al. point out in (Hanks et al., 2006) that "lightness is a matter of degree", and that "some uses [of verbs that can act as light verbs, S.C.] are lighter than others" (p. 441). They emphasize the collocational and semantic criteria for deciding whether a verb use is light or not: "The problem lies in the expectation that necessary and sufficient conditions can be established for delicate grammar categories, as opposed to characterizations of typical features. Light verbs typically focus attention on an event or process, and events and processes are very often expressed in nouns that are nominalizations (i.e. cognates of verbs) – but the focus is still on the event, even when the direct object is a word that denotes a physical entity" (p. 443). They introduce the notion of *semantic lightness* in their analysis of the verb – direct object combinations, and there is no apparent reason not to relate this term also to verb – prepositional object combinations, which their paper does not address.

Butt (Butt, 2003) draws an interesting conclusion from diachronic English studies, which supports favouring semantic and collocation criteria over syntactic the syntactic – although in their function similar to auxiliary verbs, light verbs, unlike auxiliaries, do not underlie the grammaticalization process in the development of a given language: "Light verbs straddle the divide between the functional and lexical in that they are essentially lexical elements but do not predicate like main verbs" (p. 4 and 13 in the web-released manuscript of Butt, 2003).

[7] E.g. in Butt's example from Urdu, a light verb construction even requires a second lexical verb attached to the light verb in the verb-noun structure.

[8] In this context it is to be added that evaluative expressions that are neither nominalizations nor event nouns act as such in light verb constructions; e.g. *He committed something horrible.*

4. Light Verb Constructions as Collocations

LVCs can be regarded as a type of collocation. Malmgren (Malmgren, 2002, p. 12)[9] describes a number of candidate LVCs, calling them a kind of "prototypical collocations" that consist of a semantically impoverished verb and an abstract noun. The abstract noun keeps its meaning, hence it is considered to be the more stable member of the collocation – the collocational base (or *node*, see Sinclair, 1993). Its verbal collocate is generally unpredictable.

Inspired by Mel'čuk's Meaning-Text-Theory (Mel'čuk, 1996), Malmgren analyzes Swedish verbal collocates and associates them with nouns by means of the lexical function Oper. Fontenelle (Fontenelle, 1992, p. 142) also claims that "Support Verbs roughly correspond to the type of lexical relation that can be encoded through the Oper Lexical Function used by Mel'čuk".

The understanding of nouns as collocational bases in verb + abstract noun constructions is clearly shared by Čermák, (e.g. František Čermák, 1995): "Abstract nouns seem to follow a few general patterns in their behaviour, which seem to be more structured, allowing for much less freedom than concrete nouns. The patterns the abstract nouns enter are determined by their function and meaning".[10]

While Helbig and Buscha were seeking to identify a distinct class of "Funktionsverben", and Baron and Herslund (Baron and Herslund, 1998), Rothkegel (Rothkegel, 1973), and Persson (Persson, 1975, Persson, 1992) were trying to define light verb constructions by the semantic relation between the noun phrase and the verb, Fontenelle, Malmgren, and Čermák focused on the noun, in full accordance with the pregnantly formulated observation of Hanks (Hanks, forthcoming): "…it seems almost as if all the other parts of speech (verbs and function words) are little more than repetitive glue holding the names in place".

Even in the cross-linguistic perspective, it is usually the noun that is the common denominator for the equivalent light verb constructions: "The verb […], although often the only one that is correct and idiomatic, can seem totally arbitrary. In another language – mutatis mutandis – totally different verbs often occur which work as place holders; that is why prototypical collocations often cause translation problems" (Malmgren, 2002, p. 11, and cf. Schroten, 2002).[11] Malmgren further notes that "sometimes, though by far not always, one can anticipate a sort of metaphorics" in the choice of the verb. According to Malmgren, the eventual metaphors can be traced back and explained ex post facto, but they are definitely not predictable within any one given language, let alone cross-linguistically.

[9] Malmgren's starting point is the system-oriented understanding of collocations coined especially by German linguists as Hausmann and Heid (Heid, 1998, p. 302) rather than the original English contextualist approach to collocations.

[10] Though Čermák explicitly avoids the term 'collocation', using the expression 'stable combinations' instead, among which "some are undoubtedly more frequent than others".

[11] The quotations of Malmgren, Ekberg and Dura were translated from Swedish by S.C.

5. Semantic Aspects of LVCs

From the semantic point of view, the noun seems to be a part of a complex predicate rather than the object (or subject) of the verb, despite what the surface syntax suggests (cf. Schroten, 2002, p. 93, and Boje, 1995, pp. 53, 145). As already stated by many authors (e.g. Helbig and Buscha, 1996), light verbs are in fact lexical verbs that have to some extent lost their lexical meaning, in order to provide the predicate nouns with verbal morphological categories (which is the feature that makes them resemble a verb class according to Helbig and Buscha (1996) – *Funktionsverben*, and Jelínek (2003, p. 40) – *operational verbs* (*operační slovesa*.)

Many students of this topic have observed that verbs, when occurring in an LVC, start to carry more abstract semantic features. Rothkegel (1973) considers the semantic bleaching[12] of the verb to be the antipode of verbal polysemy. She shows that the meaning of a given lexical verb in LVCs neither matches any of its meanings outside LVCs, nor does it create new meanings when associated with the respective noun phrases, which implies that instead of just being deprived of a part of its original meaning, the lexical verb acquires an additional, more abstract meaning that is reserved for the verb's occurrence in LVCs.

Butt (2003, p. 18 of the web-released manuscript) proposes that light verbs are characterized precisely by the ability to express general features, as described by Rothkegel (1973). However, Butt is explicit in that she does not regard light verb uses as semantic derivations of the primary meanings of the verbs, but contrary to that, she assumes that "the lexical specification of a handful of verbs (somewhere between 5 and 20) cross-linguistically allows for a use as *either* a main verb *or* a light verb. Some common examples crosslinguistically are the verbs for *come, go, take, give, hit, throw, rise, fall*, and *do/make*. […] Their lexical semantic specifications are so general that they can be used in multitude of contexts, that is, they 'fit' many constellations."

6. LVCs and Event Structure

LVCs are often referred to as a means of modifying the event structure of a locution, especially in languages such as Swedish, which do not (regularly) indicate aspect by morphological means (i.e. by stem vowel alternations or affixes). In such languages the aspect remains underspecified, unless lexical markers (e.g. temporal adverbs) are employed in the utterance. A kind of event structure opposition is assumed between an LVC and its corresponding synthetic predicate (when there is one). Butt (2003, p. 18 of the web-released manuscript) in accordance with many other authors, emphasizes that "light verbs modulate or structure a given event predication and do so in a manner similar to that of modifiers with respect to semantic notions such as

[12]She quotes other authors' terms, such as 'das Verblassen der Merkmale bei den Verben", "Bedeutungsentleerung", "depletion of the designatum".

benefaction, suddenness, etc.[13] […] The light verbs also tend to add further information about the aktionsart of the complex predication. In particular, there is often a telic/boundedness or a causation component." In this respect they have a function similar to verbal prefixes or particles (Butt, 2003, p. 16).

LVCs are built as compositional events or constructions consisting of a 'verbal' and a 'nominal' subevent. Yet the 'verbal' event does actually never 'take place' due to the semantic depletion in light verbs (cf. Fillmore et al., 2003). The given light verb only passes some semantic features on to the 'nominal' event. Durative events are by definition atelic (e.g. *to have problems*), with the reservation that multiple telic 'nominal' events combined with a durative atelic light verb express iterativity, e.g. *to suffer from attacks*.

LVCs denoting transitions (i.e. changes of state) are generally regarded as telic (cf. Pustejovsky, 1991), no matter what telicity value the given light verb would have if used as a lexical verb outside the LVC. Bjerre (1999) puts it this way: "LVCs denoting transitions are invariably achievements[14], either inchoatives or causatives […], the SV [i.e. *support verb*, which is the term Bjerre preferres to *light verb*. S.C.] always denotes an underspecified subevent$_1$.[…] Not surprising *terminative* is the negative counterpart of *inchoative*."

Bjerre's examples make it more clear: "*Situationen kom ud af kontrol – [The situation came out of control]* denotes a situation in which the resultant state is the negative of that in *Situationen kom under kontrol [The situation came under control]*. […] This may be paraphrased: (subevent$_1$:) The situation was under control when something happened as a result of which (subevent$_2$:) the situation was out of (= not under) control". Bjerre notes that light verbs denoting transitions are either achievement verbs with inherently underspecified subevent$_1$ (*come, bring* etc.), or they are verbs of motion or location which lose their specific relation when used as light verbs.

7. Productivity vs. Lexicalization in LVCs

Whereas traditional views emphasize that it is mostly the lexicalized units that tend to show a specific syntactic behaviour and, therefore, LVCs are to be considered as more or less lexicalized phrases, Ekberg (1987) and Dura (1997), as well as Persson (1992), concentrate on the apparent productivity of LVCs and the regular production patterns they form. Ekberg notes that many lexicalized phrases "have an almost completely or at least partly predictable meaning and new ones can be formed according to productive rules within the grammar" (Ekberg, 1987, p. 32), while Dura goes even further, adding that "even the newly-formed phrases show the same syntactic restrictions as the lexicalized ones" and interpreting this phenomenon as evidence that

[13] Cf. also Schroten (2002).

[14] Transitions are further divided into two subtypes. In *achievements* the subevent$_1$ is underspecified, unlike in *accomplishments*, e.g. *Carl built a house* (accomplishment) × *The expedition reached the top of a mountain* (achievement). See Bjerre (1999).

"these restrictions indicate that something is meant as a lexicalization rather than that they are the result of lexicalization" (Dura, 1997, pp. 1–3). She considers article-less verb-noun combinations to be an evidence that there is "a kind of word combination that is not controlled by the regular syntax but aims at lexical composition" and that it is thus "possible to form new phrases which can act as lexical units. The ordinary syntax is oriented at combining lexical units with obligatory grammatical categories, but there even seems to be another syntax, a syntax which allows language users to build larger conceptual units without involving the grammatical categories". Dura and Ekberg approach the issue from the semantic side, though they seek to draw syntactic conclusions. The syntactic criteria are eventually more important for Dura and Ekberg than they are for Hanks and others.

8. Grammatical Interference in Lexicalized Collocations?

When the morphosyntactic behaviour of a multi-word cluster systematically deviates from the regular grammar rules, it is traditionally regarded as intensively lexicalized, i.e. several words are thought of as growing together into one single semantic unit. Moreover, Dura (1997, see above) suggests that the cause – consequence relation also works the other way round: collocations that **are meant** by the speakers to be perceived as a single semantic unit are deliberately taken out of the regular language system.

Many authors since the onset of corpus linguistics have observed that the regular language use to a significant extent consists of prefabricated blocks. Needless to say, this phenomenon goes far beyond idioms and terminology. For instance, Wray (2002) builds her hypotheses on formulaic sequences on the premise that "although we have tremendous capacity for grammatical processing, this is not our only, nor even our preferred, way of coping with language input and output. […] much of our entirely regular input and output is not processed analytically, even though it could be" (p. 10).

Light verb constructions appear to be such formulaic clusters. Collocations that sometimes behave according to grammar rules and sometimes do not, would normally be regarded as somewhere half-way to the ultimate lexicalization; i.e., they would be expected to exhibit only irregular behaviour in the future development of the given language.[15] However, morphosyntactic realizations of semantically transparent collocations in text may not just vary in the extent to which they comply with the rules of grammar in terms of 'right' versus 'wrong', but, on the contrary, different

[15]This is of course not the case of idiomatic expressions, whose idiomatic meaning is inseparable from their morphosyntactic realization; e.g. *abandon ship*. Example 1 implies that the ship is thought to be sinking, whereas Example 2 lacks this implicature:

(1) *Abandon ship!*

(2) *They abandoned the ship in a bay near Hong Kong.*

grammatical realizations of collocations can have different semantic/pragmatic implicatures in the particular context according to the speaker's preference. A default behaviour of lexicalized semantically transparent collocations may often be irregular (e.g. zero article, no modifiers allowed, etc.), but the corpus evidence suggests that there is not necessarily a clear ban on a step back to the regular grammar when the morphosyntactic features help reflect the communicational intentions of the speaker in a particular discourse situation.

In other words, the assumption is that regular morphosyntactic behaviour is reintroduced when the speakers explicitly want to add the semantic features triggered by regular morphosyntactic behaviour, but they are by no means obliged to do it. The presence or absence of semantic differences between two or more alternative morphosyntactic structures is very much context-dependent, and the semantic oppositions can be obscured by the fact that they happen to be irrelevant in a particular context. That implies that the alternative expression forms will not always be mutually exclusive, but that the speakers only have the option to select the non-default pattern when they feel a particular reason for doing that.

To mention a Swedish example, the light verb construction *sätta rekord (set a record)* is normally used without an article, even when *rekord* is modified by one or more adjectives (adjective modifiers usually require the use of an article in Swedish):

(6) *Mustafa Mohammed satte personligt rekord.*
Mustafa Mohammed set a personal record.

(7) *Stefan Holm klarade 2,37 i Globen och satte nytt personligt rekord.*
Stefan Holm made 2.37 in Globen and set a new personal record.

The collocation *sätta rekord (set a record)* appears to be a very lexicalized one, judging from the predominating zero article. The large Swedish corpus Konkordanser showed that the absolute majority of the occurrences of *sätta rekord* had no article preceding *rekord*. The Konkordanser subcorpora yielded 223 occurrences of the forms *sätta, sätter, satte* and *satt*, respectively, with *rekord* following within the same sentence[16]. The noun *rekord* occurred with the indefinite article only 17 times. The percentual rates were the following:

- 2 % in the infinitive
- 0 % in the present tense
- 11 % in the simple past tense
- 9 % in the perfect tense

The definite singular form *rekordet* and the definite plural form *rekorden* occurred 11 times and once in collocation with *sätta*, respectively.

[16]Unfortunately, in Konkordanser, modern Swedish texts (newspapers and fiction) are split into 14 subcorpora, and the interface does not allow multiple selection. None of the subcorpora in Konkordanser is either tagged or lemmatized, and the interface does not support CQL. Simple Boolean queries or wildcard searches can be performed, but they cannot be combined, which significantly limits the searching power.

The 29 hits with (any) article represented 12% of the total of 235 hits.

The most frequent case (indefinite article) does not seem to be affected by tense. A closer analysis of the broader contexts showed at least one situation in which the insertion of the indefinite article may be triggered by the context (approx. 1/3 of the hits with the indefinite article) – it is when the discipline in which the record was set is specified later in the text (selection):

(8) *Svensson har satt ett oslagbart svensk rekord som sportjournalist: under cirka 49 år hade han fast jobb på samma redaktion i samma tidning, Arbetet i Malmö.*
Svensson has set an unbeatable Swedish record as a sports journalist: for approximately 49 years he had had a regular job at the same publishing office, at the same newspaper, Arbetet i Malmö.

(9) *Förre RIK-aren Peter Gentzel har satt ett nytt rekord i tyska Bundsliga. Den svenske landslagsmålvakten har på 34 omgångar tagit hela 53 straffar för Nordhorn.*
Former RIK-player Peter Gentzel has set a new record in the German Bundesliga. The goalkeeper of the Swedish national team has got 53 yellow and red cards for Nordhorn in 34 rounds.

(10) *Massorna, som köade i en halvmil för att slutligen komma till Hyde Park, satte ett nytt rekord i levande opinionsbildning.*
The crowds that were queuing for a half mile in order to finally get into Hyde Park set a new record in live opinion making.

(11) *Anette var andra halvlekens gigant och satte då ett personligt rekord. – Har aldrig gjort åtta mål i en och samma halvlek i elitserien.*
Anette was the giant of the second half and it was then that she set a personal record. – I have never shot eight goals in a single half in the elite series.

In other two cases (one with an indefinite pronoun) the sentence describes an unreal or non-specific condition:

(12) *Han säger att visst, landslaget skulle väl vara kul och visst sätta ett svenskt rekord skulle väl också vara kul, men det är saker han inte går och tänker på.*
He says that yes, the national team would obviously be cool and obviously it would also be cool to set a Swedish record, but that is stuff he doesn't go thinking about.

(13) *Om jag sätter något rekord så kommer det snart någon och slår det.*
Even if I set a record, someone else will soon come and break it.

Also setting two entities in contrast normally requires an article, as can be seen in Example 14:

(14) *Hägerstenskillen […] satte ett personligt rekord och tangerade ett: Han presterade 60 kilo i stöt (tangerat pers.) och 47,5 kilo i ryck (personligt med 2,5 kilo).*
The guy from Hägersten […] set a personal record and attacked another one: He lifted 60 kg ……

In addition, the discipline in Example 14 was specified later.

Example 15 originates from a context where records were expected in several different disciplines. A certain swimming discipline was the first discipline in the entire competition where it happened: a European record was set. In this particular context, the European record, which is a unique uncountable entity in the context of one single discipline, is regarded as countable and a member of a set.

(15) *Engelsmannen Adrian Moorhouse blev den första att sätta ett Europarekord i Strasbourg.*
The Englishman Adrian Moorhouse was the first one to set a European record in Strasbourg.

In all the other 10 hits except one, the noun *rekord* with the indefinite article was modified by one or two adjectives. All of the adjectives denoted restrictive attributes. The use of a restrictive attribute implies that that particular record was one of a set, which is normally a good reason for employing an article. Nevertheless, the zero-article is strongly preferred in this context and with the modifiers *svensk (Swedish), personlig (personal), ny (new)*, even when they concatenate. No differences in the broader context were observed that would explain why the article was used. Only a sample is presented here.

(16) *Även om serien inte var perfekt satte han ett nytt prydligt personligt och svenskt rekord med 387,60 poäng.*
Even though the series was not perfect he set a new nice personal and Swedish record by 387,60 points.

(17) *Orbit Air vann både försök och final i fjol och satte ett nytt svenskt rekord.*
Orbit Air won both the trial and the final last year and set a new Swedish record.

The definite article (found 12 times) was consequently used when referring back to one particular record mentioned before – either to the same entity (the same discipline, the same year, the same person), or to a contrasting entity. Only a selection is presented.

(18) *Hennes svenska rekord på 1.500 meter på 4.09,0 är internationellt gångbart och den tiden är ingen yttersta gräns för Gunilla. Det finns mer att ge. – När jag satte det rekordet var jag inte ens trött efter loppet. Det kändes som att dansa fram.*
Her Swedish record in the 1 500 meters at 4.09,0 is internationally accepted and this time is not the ultimate limit for Gunilla. There is more to give. – When I set that record I was not tired at all after the run. It felt like dancing.

(19) *När Bartova satte det kortlivade rekordet i Prag snodde hon det från just Flosadottir som tog sig över 4,42 …*
When Bartova set the short-lived record in Prague, she had just stolen it from Flossadottir, who got over 4,42..

(20) *Det svenska skattesystemet sätter det ena otroliga rekordet efter det andra.*
The Swedish tax system sets one incredible record after another.

It is interesting to investigate to which extent the regular grammar continues to affect multi-word clusters that already have reached the stage of lexicalization, which in principle allows them to ignore grammar. This kind of research suggests the cases in which speakers may deliberately decide **to exploit** grammar in pursuit of a particular communicative goal, since they are not forced to respect grammar for its own sake. Investigating grammar in positions where the default is not to use it at all can reveal a lot about the semantic potential of our traditional grammar categories in general.

9. Polysemy

9.1. Relations among Concrete Entities

The previous sections discussed light verb constructions and the light verbs. The majority of verbs that can be used as light verbs is also polysemous in other ways. A contrastive, corpus-based comparison of the use of basic verbs reveals that in many different contexts where Swedish employs a basic verb, the Czech equivalent is stylistically marked or more specific with respect to the given context. Quite naturally, this difference lies partly in the Czech aspectual dichotomy, which can be realized morphologically – i.e. by a stem change – as well as by derivation. Even so, however, Czech employs many more verb lemmas with mutually unrelated stems than Swedish. This implies that the Swedish basic verbs have a far higher collocation potential and a more intricate polysemy than the corresponding Czech basic verbs. In other words, a Swedish learner of Czech must learn many different verbs with a relatively low collocation potential to produce idiomatic text, while a Czech learner of Swedish must acquire very elaborate cognitive maps of collocations appropriate for a few verbs, respectively.

Fig. 1 shows one instance of this equivalent discrepancy: to express that X caused Y to sit in prison, underspecifying whether condemned to or literally escorted, Swedish uses predominantly the verb *sätta*, which is stylistically neutral. Alternatively (with far lower frequency) it uses *kasta (throw)*, which is expressive. In Czech, a number of verbs is used in place of these two Swedish ones, with the frequency counts decreasing continuously, with no abrupt drops. The counts were obtained from the corpora PAROLE (SW) and SYN2005 (CZ) (Hajič, 2004 and Spoustová et al., 2007).

The discrepancy between the collocation potential of Czech and the Swedish basic verbs grows even more evident in cases like Fig. 2 when the collocate of the given Swedish basic verb is not a single noun but a set of non-synonymous nouns that all have the same semantic relation to the basic verb. Here Czech operates with a vast amount of not mutually interchangeable verbs, which are chosen in accordance with the semantic features of the respective noun collocates. There is, unlike in Swedish,

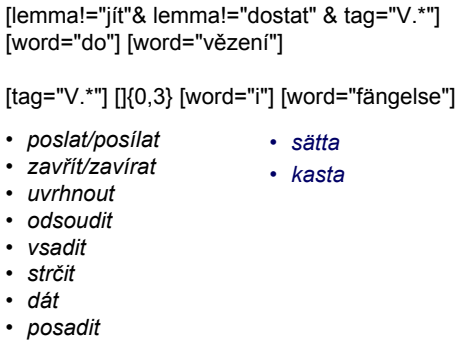

Figure 1. X puts Y into prison: verbs for put in Swedish vs. in Czech

no superior verb that would be universally used with all these nouns. The counts were obtained from the same corpora as those in Fig. 1

9.2. Swedish Spatial Conceptualization

On the one hand, Swedish seems to operate with fewer verbs than Czech. On the other hand, there is a conceptual area where Swedish systematically requires a higher degree of lexical specification than Czech. Swedish does not have any direct equivalent to the Czech dát *(give)* in the sense *put (place something somewhere)*. The speakers of Swedish must learn to choose the right verb from the set *sätta, ställa*, and *lägga*, depending on the spatial orientation of the object being moved, on the character of the target location, or even on whether the object is being attached to its target destination (e.g. with glue) or whether it keeps its new position by itself. Needless to say, a conventionalized world knowledge specific to the Swedish language community comes into the play.

To name a few examples that a Czech speaker would never resolve correctly unless he has explicitly learned them: Something that can be regarded as attached or stuck is mostly regarded as "sitting" and, accordingly, "being put into a sitting position". Thus a football can "sit" in a broken window pane, and what a post-it pad usually does on a door is also "sitting". Hence also the motivation for the example illustrated

Figure 2. X inserts a sharp object into Y: verbs for insert in Swedish vs. in Czech

by Fig. 2. Besides, you can also *sätta* a plate on the table, as well as you can *ställa* it (*put* vertically or something vertical), while the plate, once placed on the table, stands there (*stå*). Jakobsson (1996) claims that a plate can also *ligga (lie)*, but only when it is positioned upside down or when it is broken. (No cooccurrence of *ligga* and *tallrik* (*lie* and *plate*) was found in PAROLE to prove it, though.) The motivation is that the functional part of the plate points up, which gives the concept of the entire object a vertical flavour, although it is actually flat and horizontal.

In Czech, opposite to that, *sedět/stát sitta/sätta* is out of place with *plate*, nor is it usually used to express location/placement at all. On the other hand, *ležet (lie)* and *stát (stand)*, along with the corresponding causatives, are in this respect both synonymous and roughly equally frequent, as the large Czech corpus SYN2005 reveals.

9.3. Polysemy in Relations with Abstract Entities

Basic verbs belong to lexemes that "encode major orientation points in human experience" (Bybee et al., 1994, p. 10) and as such they have a mighty potential of metaphorical shifts. This paper focuses mainly on those semantic shifts that can be regarded as grammaticalization. However, the lexical description would be incomplete if it ignored those semantic shifts that have little potential to expand their collocation potential to become a universally distributed auxiliary; the more so that the shifts are language-dependent. A contrastive view is therefore absolutely necessary here. This section is dedicated to metaphorical uses of basic verbs, in the sense of "figurative" rather than what we intuitively perceive under "grammaticalized" (although, as noted above, there is no clear boundary between these two groups, and we would better perceive them as two ends of a scale rather than two sets).

Metaphorical uses that do not expand their collocation potential often arise through what Heine et al. (2001) call *Metaphorical extension from one semantic domain to another*[17]. Metaphorical abstraction relates concepts across semantic domains.[18]
Metaphorical abstraction is the way humans conceptualize the non-concrete aspects of the world. It is the naive picture of the world, in which it does not matter what the world actually is like, but what humans believe it is like. The naive view of the world is anthropocentric. Thus the closest and most discrete objects are parts of the human body and objects that can be physically manipulated. They help to 'manipulate' the less distinct entities in discourse by acting as metaphorical vehicles (Lakoff, 1987).

Heine et al. assume that the semantic domains make a hierarchy of metaphorical abstraction, through which source structures develop into target structures:

PERSON-OBJECT-ACTIVITY-SPACE-TIME-QUALITY[19]

Just to illustrate one pair, the SPACE-to-QUALITY transfer means that structures suggesting that an object is located at a place or aims in a direction regularly express that the object finds itself in a certain state or a certain situation:

(21) *The country is **sliding into** a depression.*

(22) *Belinda fell completely in love with her daughter: 'I felt **high** for about four days, not thinking about anything but caring for her.'*

Metaphorical shifts can be explained ex-post, but they cannot be predicted. This implies that there is no general principle, according to which metaphorical uses of the source language could be universally transformed into the corresponding target language. The only solution seems to be sufficient exemplification with respect to the

[17] Nevertheless, Heine et al. also delievr many examples of established function words that have arisen through this semantic shift.

[18] though metaphorical transfers also occur within a single semantic domain

[19] Heine et al. (2001, p. 48).

learner's language background; i.e. make sure to provide cases, in which Swedish would use a verb in a way unpredictable for a Czech speaker. For instance, Czech speakers *are* in a divorce when divorcing *(být v rozvodovém řízení)*, while Swedish speakers *lie* in a divorce *(ligga i skillsmässa)*. There are hundreds of such examples, and all must be consciously learned.

10. Grammaticalization through Context-Induced Reinterpretation

10.1. Context-induced Reinterpretation

As the most frequently used terms suggest, many authorities regard generalization, which lies behind or accompanies grammaticalization, as a loss of certain semantic components, compared to the core meaning of the original lexeme: *semantic bleaching* (coined by Givón) and *weakening of semantic content* (Bybee, Perkins and Pagliuca). Yet Heine, Claudi, and Hünnemeyer (Heine et al., 2001) argue that generalization is not always a reduction of meaning (p. 40f.). They present examples of negation of the core meaning and examples of addition of further semantic components not present in the core meaning. Generalization typically occurs in the following types of semantic changes:[20]

- *Metaphorical extension from one semantic domain to another*[21]
- *Context-induced reinterpretation*[22].

Heine et al. (2001, p. 70) note that a metaphorical transfer appears rather discrete. However, they propose that the transitions from one semantic domain into another create a continuum of linguistic expressions and call this continuous grammaticalizing process *context-induced reinterpretation*. They explain it on the verb *to go* in the following sentences:

(23) *Henry is going to town.*

(24) *Are you going to the library?*

(25) *No, I am going to eat.*

(26) *I am going do to the very best to make you happy.*

(27) *The rain is going to come.*[23]

Examples 23, 24, and 25 illustrate a SPACE-TIME metaphorical transfer. In Example 23, the verb *to go* has a clearly spatial meaning, whereas in 26 and 27 it has a clearly

[20] according to Heine et al. (2001) and partly Bybee et al. (1994)

[21] see previous section and cf. Heine et al. (2001).

[22] *inference* or *conventionalization of implicature* in Bybee et al. (1994)

[23] Quoted from Heine et al. (2001, p. 70). According to an English native speaker's view, 27 sounds unidiomatic and should be rephrased as *It is going to rain*.

temporal meaning. Yet Sentences 24 and 25 are ambiguous, depending very much on the context. The sentences can be interpreted in the following way:

(28) *Henry is going to town.* SPACE

(29) *Are you going to the library?* SPACE

(30) *No, I am going to eat.* (as answer to 24) INTENTION (+ relics of spatial meaning are still present)

(31) *I am going do to the very best to make you happy.* INTENTION

(32) *The rain is going to come.* PREDICTION

Both 26 and 27 have temporal meaning, but they differ in the desire of the respective subjects to pursue the event, since *rain*, let alone the empty *it*, cannot have a will or desire, while a human can.

To explain the semantic continuum, Heine et al. (2001) introduce three idealized stages of semantic shifts:

Stage I: A lingustic form F acquires a side-meaning B in addition to its core meaning A when employed in a certain context. At this stage, the utterance can be ambiguous as long as the context (both intra- and extralinguistic) does not eliminate the ambiguity, and it can be misunderstood by the recipient. (This would apply for 25.)

Stage II: The form F can be used in contexts where only the meaning B can be employed. (This would apply for 25.)

Stage III: The meaning B becomes conventionalized and cognitively salient enough to be conceived as a second meaning of the form F, which becomes polysemous. (This applies for 26 and 27). However, the meanings A and B are conceptually linked as the transition was continuous (p. 72).

Heine et al. (2001) later revised their A-meaning-to-B-meaning model, introducing the terms *focal sense* and *non-focal sense*. In this revised model, A and B at Stage III would be focal senses. At Stage I, B would only be a non-focal sense. It would be only an exploitation of the meaning A. The meaning A is supposed to have a set of conversational implicatures in addition to its core, partial pragmatic meanings which are triggered by various contexts. When a non-focal meaning B becomes highlighted as particularly suitable for expressing a given communicational purpose, it becomes more frequent and gradually gains its own set of conversational implicatures. Then it develops into to a new focal meaning B. B, undergoing grammaticalization, then generalizes even to contexts where formerly only A was accepted.

The revised model of *context-induced reinterpretation* implies the following: when determining the meaning of a grammatical entity, not only the focal meanings have to be observed, but also the conceptually prior non-focal meanings and recurring 'later' meanings likely to develop into new focal meanings must be recorded. Sentences 26 and 27 show the completed development from a volitional to a predictional future. When the structure *be going to* is used with an agentive subject, it typically has the

meaning of INTENTION: *I am going to draw this …so that he can have a full picture.*[24] As a result of the PERSON-OBJECT metaphorical transfer[25], the volitional future construction has been exploited in order to create a new convention, which implies future in events with non-human and non-agentive subjects. The evident conversational implicature is that non-human and non-agentive subjects do not activate the *will* feature in the future since they cannot pursue any will on their own: *It is going to be hot today* (PREDICTION). However, due to the generalization of the new interpretation, the PREDICTION-meaning is extendable back to sentences with agentive and human subjects: *We are going to have a new mum.* Here the structure *to be going to* is ambiguous since without the context or knowledge of the situation it is impossible to tell whether the speakers (potentially volitional) are planning to have a new mum or whether they are rather assuming that this will happen, no matter their will.

The context-induced reinterpretation appears to be the most interesting semantic change for a lexicographer seeking out "regularities which promise interest as incipient sub-systems" (Hopper, 1987). It has also been described in other words by Hanks (*exploitations of norms* in Hanks, forthcoming) as the result of a long-termed lexicographical work with authentic language data. The next section gives an example of a context-induced reinterpretation of a Swedish construction that normally expresses the progressive tense.

10.2. A Swedish Example: *hålla på*

The verb *hålla* enriched with the particle *på* is known to have grammaticalized uses. It has three valency patterns, in which the lexical verb is represented by the hypothetical verb *verba (to verb)*:

1. *X håller på med Y (Y = noun)* (lit. *X holds on with Y*)
2. *X håller på (med) att verb-a* (lit. *X holds on (with) to verb*)
3. *X håller på och verb-ar* (lit. *X holds on and verb-s*)

The progressive use approximately corresponds to the English gerund *to be verb-ing*. It is used for backgrounding events in the discourse and to indicate ongoing processes. Unlike the English gerund it is unacceptable with verbs denoting states and with verbs denoting transitions (see above). The progressive meaning is only activated in combination with atelic verbs. The combination with telic verbs yields the tendential meaning (see below). It can be used together with verbs in passive.

The progressive meaning can be rendered by *X håller på att verb-a* as well as by the coordinated construction *X håller på och verb-ar*. Pihlström observed speakers' preference for the coordinated construction, even though it had not yet been accepted as standard in the 80's. SAG does not comment on the respective variants' stylistic val-

[24] Heine et al. (2001, p. 171ff).

[25] The trasformation of volition into prediction can be seen as the transformation of *X wants* into *X wants to happen = X will happen*.

ues but adds the same observation. According to SAG, some speakers even make a sharp semantic distinction between the two variants in that they exclusively associate *X håller på att verb-a* with tendentiality and *X håller på och verb-ar* with progressivity. However, SAG mentions another tendency that goes against this semantic distinction: the coordinated construction is strongly preferred with animate agentive subjects although it is still considered odd with inanimate non-agentive subjects:

(33) *Klimatet* ***håller på att*** *bli varmare.*
?Klimatet ***håller på och*** *blir varmare.*
The climate is becoming warmer.

PAROLE contains only 118 instances of **X håller på och verb-ar**, out of which indeed only in one the subject is inanimate (a computer) but it is agentive:

(34) *och att en dator nu* ***höll på och*** *smälte svaren.*
and that then a computer was digesting (i.e. processing) the answers.

Progressivity marking is typical in telic verbal clauses in the past tense when the context indicates that the described event was prevented from reaching the expected terminal point (Teleman et al., 1999, p. 340):

(35) *Karin* ***höll på att tvätta håret*** *men blev avbruten.*
?Karin ***tvättade håret*** *men blev avbruten.*
Karin was washing her hair but was interrupted.

Interestingly, the construction *hålla på och* as well as *hålla på att* (though less frequently) appears to acquire the meaning *constantly* (which is a sort of an opposite to progressivity).

The parallel Czech-Swedish corpus has yielded one spectacular example. It is the Swedish translation of a text originally written by B. Hrabal in very colloquial Czech:

(36) *proto se taky náš farář musel* ***v jednom tahu*** *modlit, aby nebyl tak zlej…*
därför måste också vår präst ***hålla på och*** *be* ***stup i ett****, så att han inte skulle vara så elak …*
and that's why our priest had to be praying ***all the time*** *in order not to be so evil…*

The Swedish idiom *stup i ett* is perfectly equivalent to the Czech *v jednom tahu*. However, the translator added the *hålla på* construction partly to emphasize that the priest had been praying constantly or very often, but also as an indication of colloquial register[26].

More sentences containing a combination of *hålla på* and atelic verbs were sought in PAROLE, which might be bearing the semantic component of constancy. No unambiguous declarative sentence in the past tense has been found that would be acceptable without a disambiguating adverbial. Most hits (approx. 30) were propositions

[26]This assumption was confirmed by the translator in personal communication (Larsson, 2006).

with low facticity, i.e. negative sentences, sometimes with the imperative *ska (should, ought to)*, questions, and infinitives.

All the instances from PAROLE seem to be quotations of direct speech or free indirect speech[27], which suggests that this use of *hålla på* is still confined to spoken language. Exceptions will be discussed below.

Here are a few sentences from PAROLE in which *hålla på* could be substituted with *hela tiden (all the time):*

(37) *I princip tyckte hon det verkade botten att **hålla på och** knega mellan nio och fem .*
Basically she meant that it appeared miserable to keep working from nine to five.

(38) *Ni ska inte **hålla på och** larva er sådär, för jag har ingenting att skämmas för.*
You are not supposed to keep acting like this because I have nothing to be ashamed of.

(39) *Men i längden så kan vi ju inte **hålla på att** bara försvara oss.*
But for a longer time we can't just keep defending ourselves.

(40) *Är det slut? — Det vet jag inte heller. Varför ska du **hålla på och** fråga så där?*
Is that the end? – I don't know, either. Why do you keep interrogating me like this?

(41) *Men jag tyckte det var lika bra att vara kvar och inte **hålla på att** bråka.*
But I meant the best thing to do was to stay there and not to keep fighting.

PAROLE yields just one instance of a positive declarative sentence in the present tense, and, in this particular case, *hålla på* was disambiguated by temporal adverbials in the close context (cf. Example 36):

(42) *"Det var **alltid** bara som du inbillade dig." "Du förnekar det **fortfarande**. Det är otroligt." "Det är otroligt att du **fortfarande håller på och** ältar det. Jag gillade henne aldrig."*
*"You had just been fancying it". "You **still keep denying** it. It's incredible". "It's incredible that you **still keep agonizing** over that. I **never** liked her."*

How is it that a progressive construction has acquired just the opposite meaning? The progressive *hålla på* is the default interpretation of *hålla på* with atelic verbs. It appears in positive as well as in negative declarative clauses, questions etc., in all tenses. On the other hand, the 'constancy' *hålla på* seems to almost exclusively appear in negations, questions and infinitives (this is at least what the corpus evidence says). It is negation that gives a clue for the semantic change. In a negated progressive sentence, it is not just a single moment of the given event that is negated, but it is the

[27] Wikipedia (http://en.wikipedia.org): **"Free indirect speech** (or **free indirect discourse** or **free indirect style**) is a style of third person narration which has some of the characteristics of direct speech. Passages written using free indirect speech are often ambiguous as to whether they convey the views of the narrator or of the character the narrator is describing. Free indirect speech is contrasted with direct speech and indirect speech."

entire event. For instance, the sentence *De håller på att bråka (They are fighting)* focuses just on one moment in the ongoing action. The same goes for the progressive aspect as a discourse backgrounder: *De höll på att bråka när jag kom (They were just fighting when I arrived)*. However, the negation of the sentence predicate says that **the entire event** does not take place (at the moment of reference), not that **a single moment of the event** does not take place at the moment of reference. This is best perceived in the imperative; for instance, by saying: *Don't be doing*, the speaker necessarily means: "**Stop** doing that you have been doing just long enough to annoy me". Implicitly, the event really must have been taking place.

Nevertheless, the relation between progressivity and facticity also works the other way round: when the 'constancy' *hålla på* is employed in a negative imperative with an event, it suggests that the given event is actually taking place and should be stopped [28]. In written language, the reader naturally has no way to decide whether the given event is just taking place or not. By employing the 'constancy' *hålla på* the speaker adds some kind of asserting modality:

(43) *Vi **ska inte hålla på** och keynesianskt försöka mota konjunkturer. Vi ska bygga en robust arbetsmarknad*
och en stabil privat konsumtion som bägge ska klara att anpassa sig till chocker. [29]
We are not supposed to be reaching for conjunctures in a Keynesian fashion. We are supposed to build a robust labour market….

(44) *De latinamerikanska, asiatiska och afrikanska staterna **ska inte hålla på och**blanda sig i USA´s och Europas affärer hela tiden!*[30] *The Southamerican, Asian and African states should not permanently get involved in the USA's and Europe's affairs!*

In these examples, the speakers virtually underspecify the actual event(s). What they do instead is label them with expressions that are evaluative, with clear (here negative) connotations: *reach for conjuncturalisms instead of building a solid labour market*, *get involved in someone else's affairs without being invited*. The transition of the pro-

[28] According to Larsson (2006), the authentic sentence

(1) *Jag har sett som min uppgift att övertyga mitt eget folk om att vi inte kan hålla på och förtrycka ett annat folk.*
I have considered it to be my task to convince my own nation that we cannot keep suppressing another nation.

really assumes that the suppressing is taking place. It would be unacceptable to say

(2) **…att vi inte kan hålla på och förtrycka ett annat folk genom att börja bygga järnvägar på deras mark.*
**…that we cannot keep suppressing another nation by starting to build railways in their territory.*

[29] Quoted from Google, 2006-09-19, URL<http://forum.svt.se/jive/svt/report.jspa?messageID=84154>.

[30] Quoted from Google, 2006-09-19, URL<debatt.passagen.se/show.fcgi?category=3500000000000014&conference…>.

gressive *hålla på* into a 'constant' *hålla på* is a good example of a not yet completed context-induced reinterpretation. The focal sense A is clearly progressivity. The conversational implicature associated with A is 'X is happening at the time of reference'. Constancy is the non-focal sense B. The conversational implicature associated with B is 'X has been happening to the time of reference'. The corpus evidence suggests that the sense B is still bound to contexts where ambiguity is not likely to arise (negative statements, infinitives, questions and declarative clauses with disambiguating adverbials).

11. Organizing a Lexicon of Basic Verbs

Polysemy, various stages of grammaticalization, and the morphosyntactic variability of nouns in light verb constructions – this is what any lexical description of basic verbs must be especially sensitive to. A twin-lexicon is being proposed that seeks to cover these problem areas. It consists of a valency lexicon of verbs (SweVallex) and a lexicon of predicate nouns (Predicate Noun Lexicon). Methodologically, the verb lexicon draws on the Czech valency lexicon Vallex (Lopatková et al., 2007), which is based on the valency theory of the Functional Generative Description (Panevová, 1974; Panevová, 1975), but combined with Hanks's Corpus Pattern Analysis (Hanks and Pustejovsky, 2005) and enriched with Czech equivalents in context. Unlike Vallex, whose valency frames are defined syntactically, the SweVallex frame is defined by the Corpus Pattern Analysis (CPA) criteria, which take into account both the syntactic and the semantic description of the collocates. However, each complementation has been assigned a functor (semantic label used in the Functional Generative Description) and has been classified as obligatory or optional according to the Dialogue Test used in the valency theory of FGD. Each frame (here called *pattern* in terms of Corpus Pattern Analysis) consists of a *proposition* – the given verb in conjugated form, supplied with its valency complementations. Each Swedish proposition is accompanied by one or several equivalent Czech propositions.

The Predicate Noun Lexicon captures nouns acting as predicate nouns (nominal components of light verb constructions). It captures the light verb collocates of the respective predicate nouns and sorts them according to the Mel'čukian Lexical Functions (Mel'čuk, 1996). It describes the valency of the given predicate noun with each of the light verbs, respectively, and it provides information about its morphosyntactic preferences, such as determiner/modifier insertion options.

The structure of the proposed lexicon was motivated by the needs of an advanced Czech student of Swedish. There are numerous good monolingual Swedish lexicons (in the first place Svenskt Språkbruk (Clausén et al., 2003), which do not only explain the meaning of lexemes, but also describe their behaviour in context and partly their morphosyntactic restrictions (e.g. *used only with negation*). However, even Svenskt Språkbruk pays little attention to the morphosyntactic variation and to the modifying options in phrasemes and light verb constructions.

In addition, no monolingual dictionary can anticipate all contrastive issues that arise for learners with different native-language backgrounds. A nice example is the Swedish triple *sätta-lägga-ställa* versus the English *put* (something somewhere), where the Czech equivalent *dát (give)* has the same problem as English, namely being too unspecific in comparison to Swedish. It is extremely difficult to create lexicon definitions of these three respective Swedish verbs that would teach the non-native speaker to consistently choose the proper variant: the choice is based on the Swedish native conception of items as predominantly vertical vs. predominantly horizontal, or 'axis irrelevant', in connection with other aspects (whether the item must be fixed or whether it keeps its position by itself, etc. See Section 9.2, above).

The lavish exemplification of the *put*-like reading of *sätta* makes SweVallex resemble the clue page of a textbook exercise rather than a dictionary entry. The examples are simply chosen from a number of random concordances (in case of *sätta* some 2 000 of the total 9 000 concordances). Such concordances are preferred that appear surprising to the Czech speaker (e.g. *sätta en pil i*, since Czech requires a more specific verb than the equivalent of *put* (approximately *sting*), and so for a Czech speaker *put* is absolutely unpredictable in this context).

The lexicon is bilingual, with Czech being the target language. The Czech part includes just a minimal description of the Czech equivalents. This feature makes the lexicon more or less useless to a Swedish-speaking student of Czech. Creating a Swedish-Czech lexicon as a production-focused lexicon for Czechs can also seem as missing the point; apparently, the most straightforward way for the non-native Swedish text production would be using a reliable Czech-Swedish dictionary. However, production dictionaries 'atomize' the description of the source-language units according to their equivalents in the target language, such that the picture of the uses of one single Swedish word gets lost. This is also why advanced language learners prefer using monolingual dictionaries of the source language instead of bilingual dictionaries: a good monolingual dictionary seems to help draw a cognitive map' of the given lexeme. This map is a blending of semantic features and collocation options.

What production-oriented bilingual as well as monolingual dictionaries can easily miss is a target-language-specific forewarning for collocational as well as cognitive mismatches within the given language pair. There is a need for a description system that would capture the language traps explicitly – at least those based on morphosyntax and on collocability. Such a system is tested by a Czech-related description of cognitively and collocationally difficult Swedish verbs (basic verbs), which are so frequent that nobody can avoid them, and yet they are not fully explained in the teaching materials.

SweVallex-PNL is machine-readable, and its structuring allows for an automatic extraction of a Czech-Swedish glossary. The Czech glossary obtained by the extraction of the Czech equivalents of Swedish verb uses has the advantage of being fully Swedish-centered. If the lexicon was primarily designed as a Czech-Swedish dictionary, it would be Czech-centered: the cognitive map of each word would remain

Czech, and the Swedish equivalents would be chosen in a way that would disambiguate the respective Czech-centered readings of the given Czech word *('how do I say X in Swedish?')*.

As a result, among all the potential Swedish equivalents such Swedish equivalents would be intuitively selected, whose collocational preferences are not much wider than those of the Czech source word, and the commonest verbs (which are the vaguest) would be in danger of being omitted.

On the other hand, creating an ex-post Czech glossary from a Swedish-Czech lexicon allows the learner to avoid what John Sinclair (Sinclair, 1993) noticed long ago: learning rare words instead of using the less cognitively salient uses of the commonest words. A Swedish-Czech lexicon with a Czech glossary preserves the 'cognitive maps' of the Swedish words and can be used for learning more about one particular difficult (i.e. vaguely polysemous) verb, as well as it encourages the user to use these verbs in an idiomatic, native-speaker-like way.

The issue of sense disambiguation in bilingual dictionaries is very interesting, and the approach chosen varies from dictionary to dictionary. In each described word, there is a dilemma of whether the reading split is to be based primarily on differences in the collocational preferences in the source language, or rather on differences of the equivalents in the target language. SweVallex attempts at avoiding this dilemma by defining the respective readings by corpus patterns, enhanced with functors and the information on their obligatoriness. The internal structure of the entries is described in Sections 12 and 13. As a result, the Czech equivalents of one Swedish reading are not necessarily synonymous, as Fig. 3 illustrates.

[[Human, Device--]]ACT-obl ***sätter*** [[Physical Object --]]PAT-obl [[Location, Physical Object--]]DIR3-obl
(where it is meant to come, and the entity to be placed is not perceived as primarily vertical or primarily horizontal)
[[Human, Device--]]ACT-obl ***dá umístí usadí strčí zastrčí připevní přibije přilepí přišpendlí přišije přitiskne nasadí vloží přiloží zasune*** [[Physical Object--]]PAT-obl [[Location, Physical Object--]]DIR3-obl

Figure 3. Non-synonymous Czech equivalents

In sum, Swevallex-PNL was designed with respect to the following points:

1. to describe and explain a given Swedish lexeme in detail like a monolingual dictionary,
2. to provide the morphosyntactic and collocational preferences for each reading in form of a corpus pattern,
3. to determine the underlying valency frame of each Swedish corpus pattern,
4. to provide Czech equivalents and their patterns with valency frames,
5. to list phrasemes and indicate their variability options,

6. to pay special attention to light verb constructions and their morphosyntactic preferences with respect to the definiteness of predicate nouns,
7. to inform about the options of modifier insertion in light verb constructions, and
8. to provide enough examples from the corpus.

SweVallex as well as PNL are xml files with their respective document type definitions (DTD's) and CCS templates. The data was edited in the XMLMind editor (XMLmind). The CCS templates, although they may resemble dictionary entries, have no greater ambition but to facilitate the navigation through the data during the editing, and thus, this is to be emphasized, **they are not meant as the final layout for the users**. Creating the final layout, e.g. for a CD or web release, has never been the purpose of this study, which is a purely linguistic one.

Sections 12 and 13 analyze and explain the structures of both the lexicon parts, respectively.

12. SweVallex

12.1. Macrostructure

SweVallex is the lexicon of verbs. Its structure is to the greatest extent possible derived from the structure of Vallex 2.5 (Lopatková et al., 2007), the Czech verb valency lexicon. The major deviations from the Vallex 2.5 DTD are motivated by the adaptation to Swedish and by including a second language and the Corpus Pattern Analysis.

The lexicon Swevallex consists of elements `lexeme_cluster` nested in the root element `swevallex_verbs`. Lexeme clusters bring together verbs (elements `lexeme`) that are related by word formation, e.g. *sätta, sätta sig, värdesätta, sätta på*.

```
<?xml version='1.0' encoding='UTF-8'?>
<!ELEMENT swevallex_verbs (lexeme_cluster+)>

<!ELEMENT lexeme_cluster (lexeme+)>
<!ATTLIST lexeme_cluster
              cluster_id ID #IMPLIED
>
```

Each element `lexeme` has its unique ID. Each element `lexeme` contains the elements `lexical_forms` and `patterns`. Here Swevallex starts to differ from Vallex 2.5. *Patterns* is an element of the same level as `lu_cluster` in Vallex 2.5, but its function is different. Swevallex has **patterns** (like *Corpus Patterns*, instead of LU's (lexical units) introduced in Vallex 2.5. The element `lexeme` contains the actual lexicon entry.

```
<!ATTLIST lexeme
          lexeme_id ID #IMPLIED
>
<!ELEMENT lexeme (lexical_forms, patterns)>
```

12.2. Lemma

The element `lexical_forms` consists of a lemma (element `mlemma`), or a set of lemma variants (`mlemma_variants`). If the lemma is a homograph, it gets its homograph index. The past forms are listed for each lemma separately. Reflexive pronouns as well as particles are captured in the element `admorpheme`, which is optional and can be repeated. The element `admorpheme` has an obligatory attribute, which indicates its type. The values indicate whether the morpheme is a reflexive pronoun, or a particle. This solution was adopted due to the semantically relevant variability in their order – cf.: *ställa in sig* vs. *ställa sig in*.

```
<!ELEMENT lexical_forms ((mlemma| mlemma_variants),
          admorpheme*, constraints?) >
<!ELEMENT constraints (#PCDATA)>

<!ELEMENT mlemma (#PCDATA)>
<!ATTLIST mlemma
          homograph CDATA #IMPLIED
          preteritum CDATA #REQUIRED
          supinum CDATA #REQUIRED
>
<!ELEMENT mlemma_variants (mlemma+)>
<!ELEMENT admorpheme (#PCDATA)>
<!ATTLIST admorpheme
          type  (reflex| particle) #REQUIRED
>
```

12.3. Patterns

The element `patterns` consists of at least one element `pattern`. Apart from its unique ID, each element `pattern` carries the following information in the form of attribute values: is it an idiom or not? Is the form of the verb constrained for this particular pattern in any way (e.g. does it only occur in imperative)?

```
<!ELEMENT patterns (pattern+)>
<!ATTLIST pattern
          idiom (0| 1) #IMPLIED
          verb_form_constraints CDATA #IMPLIED
          pattern_id ID #IMPLIED
>
```

Each pattern consists of the following elements: `proposition`, `czech`, and `example`.

```
<!ELEMENT pattern (proposition, czech*, example*)>
```

The proposition is the Swedish corpus pattern. It has the form of a Swedish declarative sentence in the present tense (when possible), whose predicate is the lemma

verb. Its inner participants and free modifications are rendered by slots (element `slot`), integrated in the proposition (element `pattern_text`). Each piece of `pattern_-_text` has an attribute value according to whether it is the lemma verb or not. The proposition can finish with a (usually English) explaining gloss, which is called implicature (element `implicature`).

```
<!ELEMENT proposition (pattern_text| slot| implicature)*>
<!ELEMENT pattern_text (#PCDATA)>
<!ATTLIST pattern_text
          verb (1| 0) "0"
>
<!ELEMENT implicature (#PCDATA)>
```

Fig. 4 shows the proposition *sätta fart på något* in the sense of starting a motor. Note that the word *fart*, which is regarded as a predicate noun, is not explicitly present in the data, but it is referred to via a reference to PNL. The CCS template (in the picture) visualizes only the ID of the given predicate noun. For more details on the description of predicate nouns see Section 13.

Idiom:0
[[Human--driver]]ACT-obl ***sätter*** [[--]]CPHR-obl ***fart*** pnl_ref:fart-saetta-5 [[Car, Motorbike, Truck, Boat, Device--]]PAT-obl **på**

Figure 4. Swedish pattern (proposition)

The Czech equivalents are also presented in form of corpus patterns with slots, pattern text, and implications. When all the equivalents presented have the same corpus pattern, they are all placed in a row of the `pattern_text` elements with the attribute value `verb=1`. When an equivalent requires a different pattern, a new Czech pattern is created. Each Czech corpus pattern is classified according to whether it is an idiom or not and whether it really is an equivalent, or just a gloss (used in case there is a lexical gap in Czech).

```
<!ELEMENT czech (pattern_text| slot| implicature)*>
<!ATTLIST czech
          match (equivalent| gloss) #REQUIRED
          idiom (1| 0) "0"
>
```

Each pattern is accompanied by examples taken from PAROLE or (extremely rarely) from Konkordanser or Google. Examples are elements with free text. Sometimes, examples are shortened, but not consequently. In light verb constructions it is often the case that the examples even include some context.

```
<!ELEMENT example (#PCDATA)>
```

12.4. Slot

A lot of linguistic information is hidden in the complex internal structure of the slots. The slots have attributes and a nested element called `occupation`, which is present at least once per slot.

```
<!ELEMENT slot (occupation+)>
```

12.5. Surface Form

The element `occupation` carries the information about the surface form of the given slot; i.e., about prepositions, lemma, number, definiteness and other restrictions (this is important with very lexicalized collocations). Occupation can also be represented by a deliberate number of references to PNL (the optional and repetitive empty element `pnl_ref` with the obligatory attribute `ref`). The elements `slot` as well as `occupation` are common for both the Swedish and the Czech patterns. Some of the internal elements of `occupation` are therefore Swedish-specific, while others are Czech-specific, and some are common.

```
<!ELEMENT occupation ((surface_form| cz_surface)*, lexical?, pnl_ref*)>
<!ELEMENT pnl_ref EMPTY>
<!ATTLIST pnl_ref ref IDREF #IMPLIED>
<!ELEMENT surface_form EMPTY>
<!ATTLIST surface_form
          form (på| om| i| till| efter| från| framför| ifrån| för| av| med
  | utan| över| genom| att| vid) #IMPLIED
          case (basic| genitive) "basic"
>

<!ELEMENT cz_surface EMPTY>
<!ATTLIST cz_surface
          cz_form (bez| do| k| kolem| na| o| od| po| pro| před
  | s| u| v| vedle| z| za) #IMPLIED
          cz_case (1| 2| 3| 4| 6| 7) #REQUIRED


>
<!ELEMENT lexical (#PCDATA)* >
<!--text: word forms. Everything else should be in the attributes-->
<!ATTLIST lexical
          lemma CDATA #IMPLIED
          number CDATA #IMPLIED
          article CDATA #IMPLIED
          other_constraint CDATA #IMPLIED
>
```

12.6. FGD-Information

The element `slot` has two obligatory attribute values: functor and its obligatoriness according to the valency theory of the Functional Generative Description.

```
<!ATTLIST slot
          functor    (ACT| PAT| ADDR| EFF| ORIG| ACMP| ADVS| AIM| APP| APPS|
ATT| BEN| CAUS| CPHR| CNCS| COMPL| COND| CONJ|
CONFR| CPR| CRIT| CSQ| CTERF| DENOM| DES| DIFF|
                   DIR1| DIR2| DIR3| DISJ| DPHR| ETHD| EXT| FPHR| GRAD|
HER| ID| INTF| INTT| LOC| MANN| MAT| MEANS| MOD|
NA| NORM| PAR| PARTL| PN| PREC| PRED| REAS|
                   REG| RESL| RESTR| RHEM| RSTR| SUBS| TFHL| TFRWH|
THL| THO| TOWH| TPAR| TSIN| TTILL| TWHEN| VOC|
VOCAT| SENT| DIR| OBST| RCMP)   #REQUIRED

          obligatoriness (obl| opt| typ) #REQUIRED
>
```

12.7. CPA-Information

The information related to the Corpus Pattern Analysis is also contained in the slot. These attribute values are implied as the CPA is less formalized at this stage of the lexicon editing than the FGD-related part.

The attribute `sem_type` contains one or more instances from the current version of the ontology used in the Corpus Pattern Dictionary, which is being built by Hanks (Hanks and Pustejovsky, 2005).

```
        sem_type CDATA #IMPLIED
```

The attribute `lex_set` contains the lexical sets.

```
lex_set CDATA #IMPLIED
```

13. Predicate Noun Lexicon

13.1. Macrostructure

The Predicate Noun Lexicon (PNL) contains entries of nouns that occur as nominal components of light verb constructions. They are typically, but not necessarily, event nouns. Besides pure predicate nouns the lexicon also contains parts of phrasemes that exhibit morphosyntactic variability. These can be nominal components of phrasemes governed by a verb, as well as dependent parts of verbless phrasemes (e.g. *pris **på någons huvud***). Dependent parts of phrasemes governed by a noun have a simplified entry.

The root element of PNL is the element `predicate_noun_lexicon`, which consists of at least one element `pred_noun_entry` or at least one `phraseme_entry` in deliberate order.

```
<!ELEMENT predicate_noun_lexicon (pred_noun_entry+| phraseme_entry+)* >
```

13.2. Predicate Noun Lemma

The element `pred_noun_entry` displays the lemma, its possible homograph index, and the basic information about its gender and declension. As with the verb entries in SweVallex, variant lemmas (e.g. orthographic variants) are allowed.

```
<!ELEMENT lemma_variants (lemma)+>

<!ELEMENT lemma (#PCDATA)>
<!ATTLIST lemma
          lemma_id ID #IMPLIED
          homonym_index CDATA #IMPLIED
          genus (utrum| neutrum| NA| neutrum_utrum) #REQUIRED
          plural CDATA #REQUIRED
>
```

The introductory part of the entry is followed by up to three lists of typical adjectival and prepositional-group collocates of the given lemma, regardless the other context (elements `adjectives` and `pps`), and the most frequent compounds that occur with the given noun as the base (element `compounds`). Each item of the lists of collocates is surrounded with the nested element `collocate`.

```
<!ELEMENT adjectives (collocate+)>
<!ELEMENT compounds (collocate+)>
<!ELEMENT pps (collocate+)>
<!ELEMENT collocate (#PCDATA)>
```

13.3. Light Verb Unit

Like the verb entries were divided into patterns, the predicate noun entries are divided according to the combinations of the given predicate noun with a particular light verb (element `light_verb`).

```
<!ELEMENT pred_noun_entry ((lemma| lemma_variants),

adjectives?, compounds?, pps?, light_verb+)>
```

The light-verb unit consists of the optional element `czech`, which can have an unlimited number of instances, along with two optional elements that cannot be repeated: `definiteness` and `pred_noun_slots`.

```
<!ELEMENT light_verb (czech*, definiteness?, pred_noun_slots?)>
```

The element `light_verb` contains a lot of information in form of attribute values.

The lemma of the light verb occurring in the light verb construction described is to be filled in as the first attribute value.

```
<!ATTLIST light_verb
          lemma CDATA #REQUIRED
```

Each light verb construction in PNL has its unique ID:

```
 id_for_verbslot ID #REQUIRED
```

and it is classified by means of the Lexical Functions.

```
 basic_LF (Oper1| Oper2| Copul| Func| Labor1_2| Labor2_1| NA) #REQUIRED
          phasal_LF (Incep| Cont| Fin) #IMPLIED
          causative_LF (Caus| Perm| Liqu) #IMPLIED
          anti_LF (Anti) #IMPLIED
          prox_LF (Prox) #IMPLIED
```

In addition, three properties of the verb in its light-verb use are observed: telicity, punctuality, and volitionality:

```
 telicity (telic| atelic| NA) #IMPLIED
          punctuality (punctual| durative| NA) #IMPLIED
          volitionality (volitional| non-volitional| NA) #IMPLIED
>
```

The `NA` values stand for *non-applicable*, and they are selected when they depend on the context. The attribute `volitionality` describes whether or not the event denoted by the verb normally is a volitional action (regardless of the animacy and agentivity of the agent). The simplified entry for a dependent part of a phraseme does not contain the light-verb unit:

```
<!ELEMENT phraseme_entry ((lemma| lemma_variants),slot*)>
```

When the Czech equivalent is not given in the form of a corpus pattern within the verb entry in SweVallex, it is stated here. The Czech equivalents are obtained by a combination of introspection and searches in the Czech corpus SYN2005. They are nevertheless preferably captured in SweVallex. This element is much of an auxiliary element for editing noun entries that do not have their complements in SweVallex yet. As soon as they get a corresponding entry in SweVallex, the Czech equivalent gets the form of the corpus pattern and moves there.

```
<!ELEMENT czech (#PCDATA)>
```

13.4. Noun Definiteness, Modifier Insertion

Several parameters of noun definiteness are observed in the analysis of concordances of each light verb construction:

- noun with no determiner (element `bare_noun`)
- noun with the indefinite article (element `indef_art`)
- noun with the postpositive definite article (element `def_art_post`)
- noun with both the prepositive and the postpositive definite article (element `def_art_prepost`)

- noun determined by a genitive or by a possessive pronoun (element `posgen_determiner`)
- noun determined by other non-article determiner (element `other_determiner`)

When an option is clearly predominant or, conversely, extremely rare, it is indicated by a note. When some option does not occur at all in the concordances (or there are just few concordances and they are dubious), the entire element is omitted. Each option is documented by examples. The number of the examples is not necessarily proportional to the ratio of the given option in the concordances. On the contrary, more attention is paid to the less represented options: the examples tend to be longer in context in order to make it possible for the user to find out more about its motivation (e.g. markedness in the information structure, coreferential reasons, etc.). Hypotheses about the motivation of a rare pattern, when any, are formulated in the element `note`. The examples also contain implicit information about the option of the insertion of adjectival and prepositional modifiers.

```
<!ELEMENT definiteness
(bare_noun?, indef_art?, def_art_post?, def_art_prepost?,
          posgen_determiner?, other_determiner?)>
<!ELEMENT example (#PCDATA)>
<!ELEMENT note (#PCDATA)>

<!ELEMENT bare_noun (example| note)*>
<!ELEMENT indef_art (example| note)*>
<!ELEMENT def_art_post (example| note)*>
<!ELEMENT def_art_prepost (example| note)*>
<!ELEMENT posgen_determiner (example| note)*>
<!ELEMENT other_determiner (example| note)*>
```

13.5. Slot

The last unit in the PNL entry is the slot. It has a similar structure as in SweVallex: the attributes `functor` and `obligatoriness` and the element `occupation`. Unlike in SweVallex, obligatoriness is not an obligatory attribute in PNL, as the complementations are regarded as optional by default. The attribute obligatoriness is primarily used to mark surface obligatoriness of modifiers in multi-word phrasemes; e.g. *på **rätt/fel** spår, pris på **någons** huvud*.

```
<!ELEMENT pred_noun_slots (slot*)>

<!ELEMENT slot (occupation*)>

<!ATTLIST slot
functor   (ACT| PAT| ADDR| EFF| ORIG| ACMP| ADVS| AIM| APP| APPS|
ATT| BEN| CAUS| CPHR| CNCS| COMPL| COND| CONJ|
```

```
CONFR| CPR| CRIT| CSQ| CTERF| DENOM| DES| DIFF|
            DIR1| DIR2| DIR3| DISJ| DPHR| ETHD| EXT| FPHR| GRAD|
HER| ID| INTF| INTT| LOC| MANN| MAT| MEANS| MOD|
NA| NORM| PAR| PARTL| PN| PREC| PRED| REAS|
            REG| RESL| RESTR| RHEM| RSTR| SUBS| TFHL| TFRWH|
THL| THO| TOWH| TPAR| TSIN| TTILL| TWHEN| VOC|
VOCAT| SENT| DIR| OBST| RCMP)   #REQUIRED
obligatoriness (obl| opt| typ) #IMPLIED
>


<!ELEMENT occupation (surface_form, lexical, cpa, example*, ref*)>

<!ELEMENT lexical (#PCDATA)>
<!ATTLIST lexical
          lemma CDATA #IMPLIED
          number CDATA #IMPLIED
          article CDATA #IMPLIED
          other_constraint CDATA #IMPLIED

>

<!ELEMENT ref EMPTY>
<!ATTLIST ref ref IDREF #IMPLIED>

<!ELEMENT cpa EMPTY>
<!ATTLIST cpa
          sem_type CDATA #IMPLIED
          lex_set CDATA #IMPLIED
          implicature CDATA #IMPLIED
>


<!ELEMENT surface_form EMPTY>
<!ATTLIST surface_form
          form (possgen| hos| på| om| i| till| från| för| av|
          med| utan| över| genom| att| vid) #IMPLIED>
```

14. Linking

The SweVallex-PNL lexicon comprises two parts: SweVallex, which captures verbs and their patterns, and nouns and the valency frames they have in connection with

the respective light verbs with which they combine. Apart from that, PNL captures all multi-word idioms, whose structure is too complex to be described by the SweVallex pattern system. References go currently from SweVallex to PNL (Fig. 5), or from one PNL light-verb frame to another PNL light-verb frame. Lemmas and patterns/light verb frames have their ID's in both lexicons, such that more relations among and within the entries can be displayed in the future.

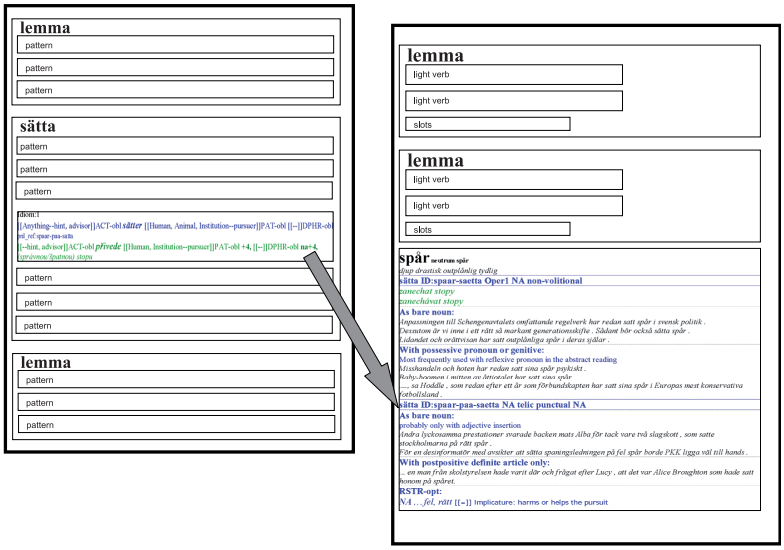

Figure 5. Reference from a pattern of sätta in SweVallex (left) to the relevant light-verb frame of spår in PNL (right)

15. Conclusion

A close corpus-based observation of basic verbs has resulted in a sketch of a learner's dictionary that would systematically and comprehensibly capture the trickiest issues of the basic verb use in Swedish. A number of linguistic theories as well as formal language description methods were critically examined, and their best features have been combined. The resulting structure is XML-based and enhanced with a simple CCS template to facilitate editing. It was tested and continuously adjusted on real corpus data.

Acknowledgements

This work was funded in part by the Companions project[31] sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434, by FP7-ICT-2007-3-231720 (EuroMatrix Plus), and by the grant of the Czech Ministry of Education No. 0021620823.

Corpora Used

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium[32].

Czech National Corpus[33] (Český národní korpus) – SYN2005. Ústav Českého národního korpusu FF UK, Praha 2005.

InterCorp[34]. Ústav Českého národního korpusu FF UK, Praha 2009.

Språkbanken vid Göteborgs universitet[35].

Corpus of Contemporary American English (COCA)[36]. Brigham Young University, Provo, Utah. Mark Davies

Bibliography

Allén, Sture. *Tiotusen i topp*. Almqvist & Wiksell, Stockholm, 1972.

Baron, Irène and Michael Herslund. Support Verb Constructions as Predicate Formation. In *The Structure of the Lexicon in Functional Grammar*. John Benjamins, Amsterdam/Philadelphia, 1998.

Bjerre, Tavs. Event Structure and Support Verb Constructions. In *Proceedings of the ESSLLI Student Session 1999*, 1999.

Boje, Frede. Hvor finder man *finde anvendelse*? In Ásta Svavarsdóttir, Guðrún Kvaran, and Jón Hilmar Jónsson, editors, *Nordiske Studier i Leksikografi Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995*, volume 3 of *Skrifter utgitt av Nordiske forening for leksikografi*, pages 51–68, Reykjavík, 1995.

Butt, Miriam. The Light Verb Jungle. *Harvard Working Papers in Linguistics*, 9(Papers from the Harvard/Dudley House Light Verb Workshop), 2003. URL `http://ling.uni-konstanz.de/pages/home/butt`, `quoted2007-01-19`.

Bybee, Joan. *Morphology: a study of the relation between meaning and form*, volume 9 of *Typological Studies in Language*. John Benjamins, Amsterdam/Philadelphia, 1985.

[31] `http://www.companions-project.org`

[32] `http://www.natcorp.ox.ac.uk/`

[33] `http://www.korpus.cz`

[34] `http://www.korpus.cz`

[35] `http://spraakbanken.gu.se`

[36] `http://www.americancorpus.org/`

Bybee, Joan, Revere Perkins, and William Pagliuca. *The Evolution of Grammar. Tense, aspect, and modality in the languages of the world*. The University of Chicago Press, Chicago & London, 1994.

František Čermák. Abstract Nouns Collocations: Their Nature in a Parallel English-Czech Corpus. In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Barnbrook, Danielsson & Mahlberg, Birmingham, 1995.

Cinková, Silvie. Extraction of Swedish Verb-Noun Collocations from a Large Msd-Annotated Corpus. *The Prague Bulletin of Mathematical Linguistics 82*, pages 99–102, 2004.

Clausén, Ulla et al. *Svenskt Språkbruk – ordbok över konstruktioner och fraser*. Norstedts Ordbok and Svenska Språknämnden, 2003.

Dura, Ela. *Substantiv och stödverb*, volume 18 of *Meddelanden från Institutionen för Svenska Språket*. Göteborgs universitet, 1997.

Ekberg, Lena. *Gå till anfall och falla i sömn. En strukturell och funktionell beskrivning av abstrakta övergångsfaser*, volume A 43 of *Lundastudier i nordisk språkvetenskap*. Lund University Press, Lund, 1987.

Ekberg, Lena. Verbet *ta* i metaforisk och grammatikaliserad användning. *Språk och Stil*, 3: 105–139, 1993.

Fenyvesi-Jobbágy, Katalin. Non-literal and non-metaphorical uses of Danish komme 'come': A case study. *Jezikoslovlje*, 2003.

Fillmore, Charles J., Christopher R. Johnson, and M. L. R. Petruck. Background to FrameNet. *FrameNet and Frame Semantics. International Journal of Lexicography – Special Issue*, 16:235–250, 2003.

Fontenelle, Thierry. Co-occurence Knowledge, Support verbs and Machine Readable Dictionaries. In *Papers in Computational Lexicography, COMPLEX'92, Budapest*, 1992.

Günther, Heide and Sabine Pape. Funktionsverbgefüge als Problem der Beschreibung komplexer Verben in der Valenztheorie. In Schumacher, Helmut, editor, *Untersuchungen zur Verbvalenz: eine Dokumentation über die Arbeit an einem deutschen Valenzlexikon*, Forschungsberichte/Institut für deutsche Sprache Mannheim, pages 92–128. Narr, Tübingen, 1976.

Hajič, Jan. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, 2004. ISBN 80-246-0282-2.

Hanks, Patrick. *Norms and Exploitations: Corpus, Computing, and Cognition in Lexical Analysis*. MIT Press, forthcoming. Manuscript, obtained 2003 from the author.

Hanks, Patrick and James Pustejovsky. A Pattern Dictionary for Natural Language Processing. *Revue Francaise de linguistique appliquée*, 10(2), 2005.

Hanks, Patrick, Anne Urbschat, and Elke Gehweiler. German Light Verb Constructions in Corpora and Dictionaries. *International Journal of Lexicography – Special Issue: Corpus-Based Studies of German Idioms and Light Verbs*, 19(4):439–458, 2006.

Hansen, Erik. *Stå, sidde, ligge*. *Mål & Mæle*, 1(2):26–32, 1974.

Heid, Ulrich. Towards a Corpus-based Dictionary of German Noun-verb Collocations. In Fontenelle, Thierry, Philippe Hilligsmann, Archibald Michiels, AndréMoulin, and Siegfried Theissen, editors, *Actes EURALEX'98 Proceedings*, volume 1, pages 301–312, Liège, 1998. Universitéde Liège, Départements d'anglais et de néerlandai.

Heine, Bernd, Ulrike Claudi, and Friederike Hünnemeyer. *Grammaticalization. A Conceptual Framework*. University of Chicago Press, 2001.

Helbig, Gerhard and Joachim Buscha. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Verlag Enzyklopädie, Leipzig, 1996.

Hoey, Michael. Introducing applied linguistics: 25 years on. In *31st BAAL Annual Meeting: Languages and Literacies*. University of Manchester, 1998.

Hopper, Paul. Emergent Grammar. *Berkeley Linguistics Conference (BLS)*, 13:139–157, 1987. URL `http://eserver.org/home/hopper/emergence.html`.

Hunston, Susan. *Patterns of Text: In Honour of Michael Hoey*, chapter Colligation, Lexis, Pattern, and Text. John Benjamins, 2001.

Jakobsson, Ulrike. Familjelika betydelser hos STÅ, SITTA och LIGGA. En analys ur den kognitiva semantikens perspektiv. Technical report, Lunds universitet. Institutionen för nordiska språk, Lund, 1996.

Jelínek, M. O verbonominálních spojeních ve spisovné češtině. In *Přednášky a besedy z XXXVI. běhu LŠSS*, pages 37–51. MU Brno, 2003.

Jensen, Torben Juel. Kan man 'ligge' i et mentalt rum? In *Nydanske studier & almen kommunikationsteori. Artikler om partikler.* , pages 73–100. Københavns Universitet. Institut for Nordisk Filologi, København, 2000.

Jespersen, Otto. *A Modern English Grammar on Historical Principles*, volume 6. London: George Allen & Unwin & Copenhagen: Ejnar Munksgaard., 1954.

Lakoff, George. *Women, Fire and Dangerous Things. What categories reveal about the mind*. Chicago University Press, Chicago, 1987.

Larsson, Mats. translation of B. Hrabal's text "Taneční hodiny pro starší a pokročilé". personal communication, September 18 2006.

Lopatková, Markéta, Zdeněk Žabokrtský, Václava Kettnerová, Karolina Skwarska, Eduard Bejček, Klára Hrstková, Michaela Nová, and Miroslav Tichý. VALLEX 2.5 – Valency Lexicon of Czech Verbs, version 2.5. Software prototype, 2007.

Malmgren, Sven-Göran. *Begå eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet 1800-2000*. Rapporter från ORDAT. Göteborgs universitet. Institutionen för svenska språket., Göteborg, 2002.

Mel'čuk, Igor A. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In Wanner, Leo, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–105. John Benjamins, Amsterdam/Philadelphia, 1996.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description II. *Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975.

Persson, Ingemar. *Das System der kausativen Funktionsverbgefüge. Eine semantisch-syntaktische Analyse einiger verwandter Konstruktionen*. Liber, Malmö, 1975. PhD thesis.

Persson, Ingemar. Das kausative Funktionsverbgefüge (FVG) und dessen Darstellung in der Grammatik und im Wörterbuch. *Deutsche Sprache*, 20:153–171, 1992.

Pihlström, Sven. *Hålla på* och *hålla på och*. *Språkvård*, 2:8–10, 1988.

Polenz, Peter von. Funktionsverben im heutigen Deutsch. *Sprache in der rationalisierten Welt. Wirkendes Wort*, Beiheft 5, 1963.

Pustejovsky, James. The Syntax of Event Structure. *Cognition*, 41:47–81, 1991.

Reuter, Mikael. Lägg ribban högt. *Reuters Ruta*, Forskningscentralen för de inhemska språken 1986. URL `http://www.kotus.fi/svenska/reuter/Kotimaistenkieltentutkimuskeskus.quoted2003-04-16`.

Rothkegel, Annely. *Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse*. Linguistische Arbeiten. Niemeyer, Tübingen, 1973.

Schroten, Jan. Light Verb Constructions in Bilingual Dictionaries. In *From Lexicology to Lexicography*, pages 83–94. University Utrecht. Utrecht Institute of Linguistics OTS., Utrecht, 2002.

Sinclair, John. *Corpus, Concordance, Collocation*. Oxford University Press, 1993. 3rd edition, 1st edition 1991.

Spoustová, Drahomíra, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*, pages 67–74, Praha, Czechia, 2007. Association for Computational Linguistics. ISBN 978-1-932432-88-6.

Teleman, U., S. Hellberg, and E. Andersson. *Svenska Akademiens grammatik*. Svenska Akademien/Norstedts, Stockholm, 1999.

Viberg, Åke. Svenskans lexikala profil. *Svenskans beskrivning*, 17:391–408, 1990.

Wray, Alison. *Formulaic Language and the Lexicon*. Cambridge University Press, 2002.

XMLmind. XMLmind XML Editor Personal Edition 3.7.0. Free software version, 2000–2007. URL `www.xmlmind.com/xmleditor/`.