Extended Lexical Units in Dutch

Michaela Poß and Ton van der Wouden

The Leiden University Centre for Linguistics (LUCL)

Abstract

The paper describes ongoing empirical research into a fundamental problem of linguistics, viz. the architecture of grammar, or the division of labor between lexicon and grammar. We try to find an answer to this question by investigating which part of the utterances in a recent corpus of spontaneous spoken Dutch consists of "Extended Lexical Units" (ELU's), hypothesized to be stored in the lexicon, rather than new syntactic constructs creatively generated from lexical atoms. We describe some problems involved in the identification of these ELUs and in the implementation of them in an NLP system. For the latter, we assess the usability of basic assumptions from frameworks such as Construction Grammar and Head-driven Phrase Structure Grammar.

1 Introduction: The importance of ELUs in Language Use

Recent developments in linguistic theory are starting to put into question the traditional picture of the language system consisting of an interesting grammar *vis à vis* a boring lexicon. Large parts of everyday spoken language are arguably constructed out of "extended lexical units" (ELUs), which we will use as a pretheoretical term to refer to all linguistic building blocks larger than words, be they compositional or not, that must be assumed to be stored in the lexicon (sometimes also known as "constructicon"), because they have idiosyncratic properties as regards their phonology, morphology, syntax, semantics, pragmatics, style level, etc. Note that lexical storage of these ELUs does not preclude the possibility that they possess various degrees of grammatical structure and/or grammatical freedom.

In any case, the very existence of these ELUs raises fundamental questions with respect to the architecture of the grammar faculty (Jackendoff 1997). From another perspective, these ELUs are a key problem for the development of large-scale, linguistically sound natural language processing technology (Sag, Baldwin, Bond, Copestake and Flickinger 2001).

The time pressure inherent to spontaneous speech situations leaves less time for the complex mental computations involved in language production than is the case in the scrupulous composition of written text. One of the strategies for speakers to overcome this problem is to employ ELUs to construct their utterances, instead of being completely original. Kuiper (1996) observes that in certain high-pressure situations, most of speakers' utterances consist of stored (or, to use another metaphor, pre-compiled) linguistic material, with very little syntactic computation going on. We may assume that this is one extreme of a cline, the other extreme being Chomsky's idealization of a creative language user with an infallible memory and infinite processing power, with enough time to verbalize new ideas in an original way.

Time pressure is not the only reason for extensive usage of ELUs, ritualization of speech situations, or of social interaction in general, being another (Wray 2002).

If a speaker of English wants to convince his or her conversation partner that he/she is following the partner's line of reasoning, he/she can do so by saying such a thing as *I see*, and if one wants to get the attention of a more or less formal gathering of people, one can raise one's voice and chant *ladies and gentlemen*. The last example immediately shows two aspects of the fixedness of this ELU: it (usually) works, even if the speaking person would never even think of individually addressing the members of audience with *lady* or *gentleman*, and the reverse order *gentlemen and ladies* is far less effective, to say the least.[1]

In the literature, estimates concerning the quantitative importance of ELUs differ greatly: Sprenger (2003) reports that some 10% of the content words in a corpus of Dutch newspaper text was part of some larger lexical entity whereas Altenberg (1998) writes: "A rough estimation indicates that over 80 per cent of the words in the corpus form part of a recurrent word-combination in one way or another." (p. 102). Bybee (2005) takes a middle position, citing Erman and Warren (2000) who "found that what they call prefabricated word combinations constitute about 55% of both spoken and written discourse"; Sag et al. (2001) offer comparable estimates.

As regards the number of ELUs used in the language community or belonging to the linguistic competence of native speakers of a language, Jackendoff (1997, 157) quotes Weinreich (1969) citing estimates of over 25,000 fixed expressions stored in the lexicon of an average speaker of English, whereas Mel'čuk reportedly claims the "phrasal lexicon" to be one order of magnitude larger than the word lexicon (Kuiper 2004). Anyway, the conclusion in Jackendoff (1997, 157) that "[t]here are too many idioms and other fixed expressions for us to simply disregard them as phenomena 'on the margin of language'" seems entirely warranted.

2 ELUs on a cline

We will not try to give a definition of ELUs here, other than the one already given above. In the literature one finds discussion of comparable concepts, such as fixed expressions ("vaste verbindingen" (Everaert 1993)) and Multi Word Expressions (Sag et al. 2001, Odijk 2003). A definition such as Everaert's, given below in (1) in the translation of Villada Moirón (2005, p. 2–3), is typical in at least three aspects: 1. it is complex; 2. it contains at least one disjunction; and 3. it refers to undefined or useless concepts such as compositionality (Zadrony 1994):[2]

(1) A combination of two or more words that must at least satisfy the (a) condition and perhaps, but not necessarily, condition (b) and/or (c):
(a) the word combination is fixed;

[1] Other arguments for the importance of ELUs in language use can be found in Wray (2002).

[2] Whether or not our ELUs can be equated to constructions, a classical concept that has gained popularity again in recent sub-branches of linguistics (cf. Fillmore, Kay and O'Connor (1988), Goldberg (1995), Croft (2001), etc.) remains to be seen. A very general definition such as the following probably covers most of our ELUs as well: "It is safe to say [...] that in essence a construction is a pattern in the formal properties of a language (i.e., in its form) that is associated with a particular function. While various theories may choose to interpret this definition broadly or more narrowly, the basic notion of a construction as a pattern of form and function remains the same" (Goldberg and Casenhiser n.d.).

(b) the combination as a whole has a non-compositional or partially compositional meaning;
(c) the syntactic/morphological behavior of the fixed expression and/or its parts is not to be expected given the syntactic/morphologic behavior of the individual words or the combination as a whole

The reason that a proper, exhaustive definition of ELUs is so problematic is that we are actually confronted with a large number of structurally different things. Expressions can be stored in the lexicon because of their noncompositional semantics, their unusual syntax, unexpected pragmatics, unexpected phonology, etc. It is therefore hard, if not impossible, to find a common basis on which ELUs can be categorized (cf. also Wray (2002)).

For the analysis of ELUs, we propose an approach that is based on Jackendoff's constructionist ideas of the organization of the lexicon (Jackendoff 1997). Jackendoff does not restrict the lexicon to word-sized elements, but counts lexically underspecified patterns – constructions – as lexical items as well. In accordance with basic Construction Grammar[3] (CxG) assumptions (cf. also Kay (2002), Goldberg (1995)), we see the lexicon as a hierarchically ordered inheritance network, with a cline of lexical fixedness.

We assume that, at least for computational applications, a categorization of ELUs on the basis of their lexical fixedness may be useful. Every (type of) ELU is given a particular place in the network of constructions that is assumed to be (a model of) our mental lexicon. A prominent advantage of this categorization is that ELUs are integrated into the lexicon in a very straightforward way, namely without giving them a special status compared to simple words or schematic phrases. This seems reasonable if one wants to account for the fact that the larger part of our daily language is made out of fixed expressions, prefabs, etc. Another advantage is the fact that thinking in terms of a cline doesn't lead to categorization problems when it comes to fuzzy boundaries. But let us take a look at the design of such a system first.

One end of our cline of lexically fixed constructions covers those lexical elements that are completely instantiated. This group includes all morphemes and single words that show no morphological variation, such as function words, but also elements that can be referred to as *words with spaces*. Their common characteristics is complete inflexibility, and we therefore have to assume that they are stored in the lexicon as such. Examples for Dutch words-with-spaces are named entities (*United Nations, Kofi Annan*) and prepositional phrases like *op grond van* (on ground of) 'on the basis of' and *ter ere van* (to-the honor of) 'to honor' (the latter being a frozen archaic expression containing a fossilized case of nominal inflection). These simple constructions are listed "as is", and the features that are listed with them include combinatorial, phonological and semantic information.

Walking further on our cline, we find lexical entries that show inflection but do not allow for other types of alternations. At this position, lexical words like nouns and verbs (simplices and compounds) are stored, but also ELUs that don't allow for

[3]Even if we use the "capital c capital g" notation, we refer to a variety of constructionist approaches rather than the one specific by Fillmore and Kay

syntactic alternation, although they may show inflection. These constructions carry additional information about the inflectional paradigm they partake in.

The next landmark on the cline are idiomatic expressions that still do not allow for lexical variation but that are syntactically flexible. An example is the expression *het loodje leggen*: its semantics is non-compositional (the literal meaning is 'to lay the little lead', the actual meaning is 'to come of badly, to die'), but its syntax is more or less like any old direct object construction:[4]

(2) Ponyclub dreigt loodje te leggen.
Pony-club threats lead-DIM to lay.
'Pony club appears to die.'

(3) Internetbedrijven leggen massaal het loodje.
Internet-companies lay massively the lead-DIM.
'Internet companies massively go bankrupt.'

Although idioms such as *het loodje leggen* are supposedly listed with their complete lexical content, they inherit their alternation patterns from more general constructions (e.g. the transitive construction) higher up the cline in order to allow for phenomena like verbal concatenation. The fact that processes such as passivization are not possible in this case is a particular feature of the more specific construction which has to be specified explicitly as well.

The first real lexical variation is found a bit further up the cline. Here we find constructions that are still rather fixed in their overall design, but allow for lexical alternation in particular slots. An example is the *Subj V er geen NP van*-construction (cf. e.g. Hoeksema (2001a)). It is a negative polarity construction used (in informal speech only) to express that one does not understand or believe something. The construction is syntactically flexible up to a certain degree (but it does not allow for passivization) and has three (not completely flexible) slots: The subject is an agent who has to be able to understand, the verb comes from the semantic paradigm of *understand*, *believe* and *being able to* and the slot in the Direct Object-NP is filled with a member of a rather narrow but semantically hard to define paradigm.[5] The following sentences give examples for this particular construction.

(4) hij wist er geen bal van
he knew there no ball of
'he didn't know a thing about it'

(5) ze snapt er geen flikker van
she understands there no faggot of
'she doesn't understand a thing of it'

[4]Examples found via Google, URLs http://home-1.tiscali.nl/~kuifje/editie150101.htm and http://wijkcent.a2000.nl/wijkcent/krant/editie/2001/01033002/txt/050501.htm (01.12.2004).

[5]Sentential negation may also be "raised", as in *ik denk niet dat hij er iets van snapt* (I think not that he there anything of understands) (I don't think he understands a thing of it) or be incorporated in the subject, as in *niemand begreep er ene flikker van* (nobody understood any faggot of it) 'nobody understood a word of it' (Postma 2001); the last example shows another optional particularity discussed by Postma, viz., a negative polarity variant *ene* of the indefinite article whose unmarked form is *een*.

Even if the interesting slots in this construction are flexible throughout, we still do not consider it to be a schematic idiom (i.e., a idiomatic pattern with underspecified slots like the *way*-construction or the resultative construction), as the fillers of the slots are chosen from narrow paradigms and the possible candidates must be learned along with the idiom.[6]

On the other hand, we consider the so-called *way*-construction as a real schematic idiom; it is exemplified in the following sentences:[7]

(6) Braid virus baant zich een weg door email.
Braid virus *banen-3rdSG* itself a way through email.
'Braid virus makes its way through email.'

(7) Twee bussen boren zich een weg naar het hart van Istanbul.
Two busses drill themselves a way through the hart of Istanbul.
'Two buses make their way to the hart of Istanbul.'

(8) De flits baant zich een gloeiend heet pad door de lucht.
The lightning *banen-3rdSG* itself a red hot path through the air.
'The lightning makes his red hot path through the air.'

(9) een prachtige streek waarin zeven riviertjes zich een pad
a wonderful area where-in seven rivers-DIM themselves a path
kronkelen naar de zee
wind to the sea
'a wonderful area where seven rivers wind to the sea'[8]

This kind of construction differs from the last one by an even bigger degree of lexical flexibility. The *X er geen Y van*-construction fills its slots with elements from a narrow paradigm (either to be defined semantically, or stipulated purely lexically), whereas the *way*-construction allows for much more variation, especially as far as the verb slot is concerned.

Even if there is not a single lexically fixed slot in this construction, we would like to consider it an ELU, too, for at least two reasons. Firstly, there is a prototype that shows statistical significance, namely the combination of the verb *banen* and the nominal head *weg*.[9] Secondly, the semantics of the *way*-construction is highly idiosyncratic (see Verhagen (2003)).

What we expect from a categorization like this is a twofold achievement. On the one hand, we assume that a thorough insight into the structural fixedness of ELUs helps retrieving them from corpora, as the corpus search can be refined to a large degree if possible alternations are accounted for (see also Villada Moirón (2005)). On

[6]Moreover, there are strong collocational effects between the verb and the expressions of minimal quantity (Hoeksema 2001b).
[7]For a thorough description of the Dutch *way*-construction, see Verhagen (2003), for a description of the English counterpart, see Goldberg (1995).
[8]Found on http://www.freewebs.com/maisjo/infooverdestreek.htm (01.02.2005).
[9]Verhagen (2003) provides us with corpus evidence revealing that e.g. more than 50 per cent of the instantiations of this construction are built with *banen* (that occurs in this construction only), whereas the rest is spread over about 20 verbs with little significance.

the other hand, we see prominent advantages in implementing hierarchically ordered inheritance networks (cf. Poß (2005)). As we are investigating into the nature of inheritance and which features are inherited by which kind of ELU, we are in need of a categorization of some sort. The advantages of this kind of approach (no problems with fuzzy boundaries, no special theoretical status of ELUs) weigh heavier than the known disadvantages (like e.g. the still missing definition and the fact that we completely neglect semantics).

3 Retrieving ELUs from a corpus

3.1 The goal

We try and operationalize the question of the division of labor between the grammar and the lexicon by investigating, in a corpus of spontaneous spoken language, the amount of constructs that recur significantly more often than chance predicts. In this respect, we restrict ourselves, at least for the time being, to ELUs that can be defined in a statistical way.

3.2 The corpus

For our investigations, we use the Spoken Dutch Corpus (CGN), a collaborative effort of several Dutch and Flemish universities, funded by both the Dutch and the Belgian government, completed in 2003. The corpus contains almost 1000 hours of continuous speech, which amounts to a little less than 10 million words (Oostdijk 2000). The corpus was intended as a major resource both for linguistic research and speech technology. To serve this dual purpose, it contains text fragments recorded in a wide range of communicative settings: spontaneous face-to-face and telephone dialogues, interviews, debates, news broadcasts, etc. Two-thirds of the material is collected in the Netherlands, one third in the Dutch speaking part of Belgium. It is the largest and most diverse database of spoken Dutch collected so far.

3.3 Some problems

Standard computational techniques from collocation research (Manning and Schütze 1999, ch. 5) are useful to find certain types of ELUs. For example, a program such as Wordsmith Tools[10] has no problems in finding classical collocation types such as fixed prepositions with adjectives, such as *trots* ***op*** 'proud of'. The table below show the most frequent two word clusters with *trots* in the Dutch part of the corpus. The collocation we were looking for ranks first.

(10)

Wordsmith clusters with **trots**		
N	Cluster	Freq.
1	*trots op* 'proud of'	75
2	*heel trots* 'very proud'	15
3	*trots en* 'proud and'	7

[10] http://www.lexically.net/wordsmith/index.html

With these methods, one may also find relations between lexical items that are not in the books of reference. For example, employing the same methods, it has been demonstrated (van der Wouden 2002) that the Dutch complex focus particle *niet eens* 'not even' shows strong collocational effects with, among other things,

- other particles (i.e., high frequency (function) words), such as *nog* 'yet' and *meer* 'anymore':

  (11) dat is nog niet eens zo lang geleden
  that is yet not even so long ago
  'that isn't even that long ago'

  (12) de man luistert niet eens meer
  the man listens not even anymore
  'the man doesn't even listen anymore'

- (high frequency content) verbs such as *weten* 'to know'

  (13) ik weet niet eens wie Judith Bosch is
  I know not even who Judith Bosch is
  'I even don't know who Judith Bosch is'

  (14) die weten niet eens waar Nederland ligt
  they know not even where Netherlands lies
  'they don't even know where the Netherlands are'

- modal auxiliaries, especially *kunnen* 'can'[11]

  (15) dus ik kan niet eens werken als ik zou willen
  so I can not even work if I would want
  'so I can't work even if I wanted'

  (16) kunnen nog niet eens hun naam en adres schrijven
  can yet not even their name and address write
  '[they] don't even know [how] to write their names and addresses'

On the other hand, we can already be sure that certain types of ELUs will be missed by our techniques. Consider, for example, the Dutch verb *krijgen* 'to get'. Not unlike its English translation, it can be found in quite a number of idioms, phraseological idioms, light verb constructions etc. The dictionaries and grammars list tens or even hundreds of these ELUs. Some of these ELUs to be found there hardly sound familiar to native speakers of the language, which may mean that these combinations are obsolete or dialectal or something like that. Many others, however, are recognized immediately by native speakers. A case in point is the invective *krijg de klere* 'drop dead', which

[11] This collocational effect appears to be restricted to one reading/usage of *kunnen*, viz., the dynamic (i.e. non-epistemic, non-deontic) one: 'be able to'.

literally means ‘get the cholera’. The ELU also has a variant *je kunt de klere krijgen* ‘you can get the cholera’, which is just as rude as the imperative form. Although both variants are known to all native speakers we questioned, our corpus techniques will not find them, at least not in the corpus we have chosen to use, as *krijg de klere* occurs only once in it, and *je kunt de klere krijgen* not at all.

The sparse data problem, i.e. the fact that our corpus (and probably anyone’s corpus) is too small to offer statistical evidence for all ELUs, is a real one, just like it is a real problem in all other quantitative approaches to language. However, it remains to be seen yet whether the restriction to ELUs defined statistically will seriously flaw the answer to our fundamental question regarding the division of labor between grammar and lexicon, between computation and storage.

Apart from these false negatives, i.e. real ELUs not found by statistics, our quantitative methodology yields false positives as well. Consider *ja* ‘yes’, which is the most frequent word in the Dutch part of the corpus. According to Wordsmith Tools, the following are among the 10 most frequent multi word clusters involving the string *ja*:

(17)

Frequent clusters with *ja*		
cluster	**rank**	**#N**
ja ja	1	181347
ja ja ja	2	84696
ja ja ja ja	3	34430
oh ja	5	18512
ja dat	6	17721
ja maar	7	16343

Most clusters in this little table qualify as ELUs. The first one, e.g., *ja ja* usually functions as a discourse marker (cf. e.g. Schiffrin (1987)) and may express either consent or doubt, depending on the intonation with which it is pronounced (which we assume to be lexicalized with the ELU). Consider the following conversation fragment (edited slightly for expository purposes):

(18) ja die uh die tante Hennie zit een beetje te miepen hè.
yes that uh that aunt Hennie sits a bit to whine PART
‘that aunt Henny is whining a bit, isn’t she?’

(19) ja die wilde zich laten euthanaseren of niet?
yes she wanted self let euthanatize or not
‘yeah, she wanted to have herself euthanatized, didn’t she?’

(20) ja ja die wil euthanasie tegen die tijd.
yes yes that wants euthanasia against that time
‘sure, in due time she wants euthanasia’

The combination *ja dat* ‘yes that’, however, does not seem to be an ELU. The high frequency of its occurrence is due to the interplay of a number of factors concerning the grammar of Dutch and the organization of Dutch conversations: one can use a

discourse marker such as *ja* to express consent and to take the turn at the same time (Mönnink 1988), and it is good practice (Onrust, Verhagen and Doeve 1993) to start one's turn by referring to a topic salient in the conversation by means of a deictic element such as the demonstrative pronoun *dat* 'that'.

(21) ja dit moet nog gedaan worden
yes this must yet done become
'this has yet to be done'

(22) ja dat klopt
yes that knocks
'yes that's correct'

3.4 Some solutions

For the problem of false negatives, there is no principled solution – increasing the size of your corpus may help you to find statistical evidence for certain combinations, but new hapax combinations will turn up; moreover, the number of false positives increases with the size of the corpus. Various solutions have been proposed to overcome the problem of the false positives involving high frequency function words (Manning and Schütze 1999). One approach is to pass the candidate phrases through a part of speech filter which only lets through those patterns that are likely to be interesting combinations (Justeson and Katz 1995), another one excludes certain words (e.g., high frequency function words) from participating in candidate combinations (Smadja and McKeown 1990), a third one is using more sophisticated statistics (e.g. Krenn and Evert (2001)). A common feature of these approaches is that they all work some of the time, but none of them works 100% all of the time, at least not for all types of ELUs (van der Wouden 2001, Villada Moirón 2005).

We will not propose the ultimate solution here, as we didn't find it. Moreover, we assume that there is no such thing as a unique ultimate solution for the problem of identifying all and only ELUs from a corpus, for the simple reason that ELUs are too heterogeneous for that. Ultimately, they are the surface manifestation of a number of complex phenomena, the result of a variety of interactions of atoms, mechanisms, rituals and habits from grammar, lexicon and the extra-linguistic real world.

This conclusion implies that we will have to combine the existing techniques and heuristics that are on the market, and think of developing and validating new ones. Various types come to mind: the corpus is enriched with Part Of Speech information, and part of it is annotated syntactically: these two annotations layers open new horizons to searching for types of ELUs that cannot be found by simple (or not so simple) string matching and statistics on the raw text.

4 The implementation of ELUs

Another interesting challenge lies in the implementation of ELUs. Idiosyncratic combinations are still a hurdle for parsing and generating that is particularly hard to take. Depending on the type of expression, many implementations vary from inelegant to

impossible. E.g., how does a system deal with expressions that do not conform with general grammatical rules (like *by and large*)? And how does the system get hold of the non-compositional semantics of idioms (like *to spill the beans*)? We will give a short overview of an approach of the analysis of ELUs by Sag et al. (2001) that offers solutions for various kinds of ELUs. However, we will raise the question whether a more integrating approach could be adopted if there is a deeper understanding of what the building blocks of constructions are and how they can be used in a computational system.

4.1 ELUs in HPSG

For an analysis in Head-Driven Phrase Structure Grammar (HPSG), a method of dealing with various kinds of ELUs has been offered by Sag et al. (2001). This approach accounts for various types of non-compositional multiword expressions, at least as long as they have at least one stable lexical item. Sag et al. (2001) come to the conclusion that different kinds of multiword expressions should be analyzed in different ways, depending on their nature. Three different categories are established, each category of expressions is dealt with in its own way.

The so-called *fixed expressions* are those that are completely immutable like *by and large*. They are treated like words-with-spaces, therefore, in an implementation this leads to string-type listing. The *semi-fixed expressions* show a still low degree of flexibility, as they only allow for inflection, variety in the choice of determiners, different reflexive pronouns, etc., but not for variation with regard to new lexical items. This group includes decomposable (*spill the beans*) and non-decomposable idioms (*kick the bucket*), compound nominals (*part of speech*), and proper names. As semi-fixed expressions behave like single parts of speech, they are represented in the lexicon as single items, with pointers to the element(s) that can undergo inflection resp. can be replaced. As opposed to the semi-fixed expressions, the group of syntactically flexible expressions allows for a much higher degree of structural variability. Light-verb constructions (*make a mistake*) and verb-particle constructions (*look up*) belong into this category, as well as structurally rather free decomposable idioms (*let the cat out of the bag*). For this heterogeneous group, the different analyzing techniques range from subcategorization (e.g. the verb *hand* subcategorizes for *out*) to a so-called idiomatic-construction analysis.[12]

With these techniques, Sag et al. (2001) can cover a big part of the range of fixed and semi-fixed expressions. What they cannot deal with is the kind of phenomena that we referred to as schematic idioms. As soon as there is no direct lexical trigger in the expression, the method does not work anymore.

A solution within HPSG (Pollard and Sag 1994) could be found in defining a lexical rule that transfers certain intransitive verbs into verbs triggering the *way*-

[12]The term idiomatic construction does not completely cover what it usually refers to in Construction Grammars. The difference (within HPSG) is an analysis where listemes do not get an idiomatic meaning assigned and then, in turn, subcategorize for other listemes with an idiomatic meaning. Instead, the elements are allowed to combine regularly, and the complex expression carries the idiosyncratic meaning. Condition for this is that every listeme in the idiom is known as such.

construction. The problem that we see, at least for the time being, lies in the fact that little is known about the further restrictions of the verbs.[13] Riehemann and Bender (2001) argue in favor of a construction-based rather than a strictly lexicalist approach when it comes to idiosyncratic patterns without a single stable lexical item. We want to go further and investigate into an approach that is entirely based on constructions, no matter if there is too little lexical information or not.

4.2 A Construction Grammar-based approach

Starting point of our approach are the assumptions that underlie the various streams of Construction Grammar (as e.g. Goldberg (1995), Croft (2001), Fillmore et al. (1988)). Langacker (2003) gives an overview of the shared notions within the various traditions. They boil down to non-derivationality, monostratality, unity of grammar and lexicon, a cline from specific to schematic constructions that are all stored in the so-called constructicon (Goldberg (2003)), the linking of constructions in an inheritance network, and unification as the motor that drives composition. The fact that lexical as well as phrasal entries are assumed to be stored in the mental lexicon offers an interesting alternative design of lexicon and grammar rules, compared to the lexicalist HPSG approaches.[14] Thus, the innovative point a system based on constructions may offer is a lexicon where apart from lexically specified entries, underspecified patterns are stored. Moreover, in the Construction Grammar framework, ELUs are treated exactly the same way and receive the same status as words (on the one end of the cline) or abstract schemas (on the other end of the cline). But how can these Construction Grammar tenets help us with the implementation of ELUs? In order to illustrate our approach, we give an analysis of the *way*-construction below.

Verhagen (2003, 34) presents a schematic description of the Dutch *way*-construction:

(23)

$$\left[\begin{array}{lccccc} \text{Sem:} & \text{creator} & \text{create-move,} & \text{for-self} & \text{created-way,} & \text{path} \\ & | & | & | & | & | \\ \text{Syn:} & [\text{SUBJ}_i & [\text{V} & [\text{REFL}_i & [\text{DO}] & \text{OBL}]]] \end{array}\right]$$

This formalization is in accordance with the basic CxG tenet that formal and semantic features of a construction go hand in hand. What the figure shows is the following: the (abstract) construction is built up from meaning components that must occur in all particular constructs. The syntactic structure is obligatory as well: a ditransitive pattern with an oblique argument (i.e., a PP, usually). The two sides of the description have no distinctive power in themselves. The semantic structure could also be

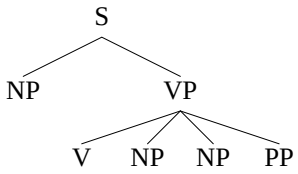
[13] This, of course, is not a proper argument, as every construction has to be analyzed carefully, anyway. Nevertheless, this part of the project aims at designing an experience model rather than a wide coverage grammar of spoken Dutch. The underlying question is not a technical, but rather a cognitive one: Which parts are built from smaller items and which are just bigger building blocks glued together. And for the latter: Which elements can be altered, and which alternations are taken care of by more global mechanisms.

[14] At least the strict lexicalist ones, following the standard approach of Pollard and Sag (1994). In more recent literature, the notion of construction and phrasal patterns rises, see e.g. Sag, Wasow and Bender (2003).

represented by an utterance that does not make use of the way-construction, and a ditransitive sentence with an oblique argument is nothing particularly special, either. What makes it unique is the linking of those two layers (i.e., the lines between the semantic and syntactic structure, showing which semantic component is expressed by which constituent or vice versa).

If we take Verhagen's analysis and translate it into information that can be recognized by a computational system, at least the following reliable information is available: The parser must find a parse of the following structure:

(24)



If the parser recognizes grammatical functions as well, the first NP must be the subject, the second NP must be the indirect object, and the third NP must be the direct object. This is the syntagmatic information. But there is paradigmatic (or even lexical) information available, too. The indirect object NP must be instantiated by a member of the paradigm of reflexive pronouns and cannot be modified, furthermore the pronoun must show agreement with the number and gender features of the subject NP. The direct object NP carries specific lexical information, namely that the determiner must be indefinite. Additional semantic information is available as well, namely that the NP must have PATH-semantics.[15] In the process of choosing the right construction, this information narrows down the list of possible candidates to a rather small set. The PP adjunct must be a directional adverbial, and directionality is caused by the semantics of the preposition. If we store all reliable information in a schema, we get the following picture:

(25)

$$\left[\begin{array}{llllll} \text{Sem:} & & & \text{refl} & \text{way,} & \text{dir} \\ & & & | & | & | \\ \text{Syn:} & [\text{SUBJ}_{\text{i}} & [\text{V} & [\text{IO}_{\text{i}} & [\text{DO}_{-\text{def}}] & \text{PP}]]] \end{array}\right]$$

Note at this point that this new schema is not a replacement for the one given in (23) above, it only mirrors the information that can be retrieved by a parser using conventional methods and that enables the system to distinguish this pattern from all other possible patterns. If the parser finds these features in the input sentence, it may categorize the sentence as an instantiation of the *way*-construction. Once a construction is recognized, all semantic and combinatorial features that its lexicon entry is enriched with must be applied.

[15] In order to be able to process this information, there are two possible ways. Either the lexical items are organized in fixed sets of semantic groups, according to their relevant features, or a semantic ontology must be included into the system. For the time being, we stick to the first solution (with pleasing results), although the latter seems to be attractive as well.

4.3 Putting the puzzle together

The kind of approach we propose differs from lexicalist approaches in the sense that it assumes special phrasal entries rather than special lexical entries. According to constructionist approaches, the (idiosyncratic) semantics of any utterance (and therefore of any ELU)[16] is the contribution of the particular construction, in the first place. Structural information is used for recognition, but additional features are needed to identify a given construct (i.e. the instantiation of a construction).

The features in question range from strictly lexical to abstract semantic. For items like the verb *banen*, for instance, it is economical as well as elegant to adopt a pointer to the *way*-construction in the feature structure of the lexical entry, as *banen* is a hapax in the sense that it cannot occur outside the *way*-construction. But as there are instantiations without the specific lexical items, another way to trigger the construction is proposed in analyzing a bundle of features. In the case of the *way*-construction, the features are basically phrasal and semantic.[17] The in-depth analysis of more constructions will provide us with a deeper insight of the nature of the features that establish constructions, and de design of a system that deals with ELUs on a constructional basis. On a more theoretical level, we expect to find interesting insights regarding the design of the Construction Grammar theory by formalizing it to a degree that makes it implementable.

5 Summary

In this paper, we presented first results of the ongoing research project *Dutch as a Construction Language*. We described how we try to investigate the purely theoretical question of the division of labor between lexicon and grammar using computational methods, namely extraction on the one hand, and implementation on the other. When it comes to extraction, we found that there is not one sing2le statistical method that can cover the whole range of different phenomena we consider an ELU. Different from that, for implementation, we hope to find a useful technique that is inspired by Construction Grammar assumptions and that is able to handle the whole range of constructions using one single mechanism, namely the analysis of feature bundles as phrasal patterns.

References

Altenberg, B.(1998), On the phraseology of spoken English: The evidence of recurrent word-combinations, *in* A. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*, Clarendon Press, Oxford, pp. 101–22.

Bybee, J.(2005), The impact of use on representation: grammar is usage and usage is grammar, Presedential address LSA, January 8, 2005.

Croft, W.(2001), *Radical Construction Grammar. Syntactic theory in typological perspective*, University Press, Oxford.

[16] Not every construction is an ELU, but probably every ELU is a construction

[17] Semantic in terms of the lexical items found in the construction. Therefore, it does not get in conflict with the categorization of ELUs proposed above, as it still does not make use of the overall semantics of an ELU.

Erman, B. and Warren, B.(2000), The idiom principle and the open choice principle, *Text, an interdisciplinary journal for the study of discourse* **20**(1), 29–62.

Everaert, M.(1993), Vaste verbindingen (in woordenboeken), *Spektator* **22**(1), 3–27.

Fillmore, C. J., Kay, P. and O'Connor, M. C.(1988), Regularity and idiomaticity in grammatical constructions: the case of *let alone*, *Language* **64**(3), 501–39.

Goldberg, A.(2003), Constructions: a new theoretical approach to language, *Trends in Cognitive Sciences* **7**(5), 219–223.

Goldberg, A. E.(1995), *Constructions. A Construction Grammar approach to argument structure*, University Press of Chicago, Chicago.

Goldberg, A. E. and Casenhiser, D.(n.d.), English constructions, Ms. Princeton University, to appear in *Handbook of English Linguistics*. Blackwell Publishers.

Hoeksema, J.(2001a), Partitivity, degrees and polarity, *Verbum* **XXV**(1), 81–96.

Hoeksema, J.(2001b), Rapid change among expletive polarity items, *in* L. J. Brinton (ed.), *Historical Linguistics 1999. Selected Papers from the 14th International Conference on Historical linguistics, Vancouver, 9–13 August 1999*, John Benjamins, Amsterdam/Philadelphia, pp. 175–186.

Jackendoff, R.(1997), *The Architecture of the Language Faculty*, The MIT Press, Cambridge, Mass.

Justeson, J. S. and Katz, S. M.(1995), Technical terminology: some linguistic properties and an algorithm for indentification in text, *Natural Language Engineering* **1**, 9–27.

Kay, P.(2002), An informal sketch of a formal architecture for construction grammar, *Grammar* **5**, 1–19.

Krenn, B. and Evert, S.(2001), Can we do better than frequency? A case study on extracting PP-verb collocations, *in* B. Daille and G. Williams (eds), *COLLOCATION: Computational Extraction, Analysis and Exploitation. Proceedings of a Workshop during the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter, Toulouse, France, July 7th*, CNRS – Institut de Recherche en Informatique de Toulouse, and Université de Sciences Sociales, Toulouse, France, pp. 39–46.

Kuiper, K.(1996), *Smooth talkers: the linguistic performance of auctioneers and sportscasters*, Lawrence Erlbaum Associates, Mahwah, NJ.

Kuiper, K.(2004), [review of] A. Wray: Formulaic language and the lexicon, *Language* **80**(4), 868–872.

Langacker, R. W.(2003), Construction grammars. cognitive, radical and less so, Plenary Paper ICLC 8, Logroño, Spain.

Manning, C. D. and Schütze, H.(1999), *Foundations of statistical natural language processing*, The MIT Press, Cambridge, Mass.

Mönnink, J.(1988), *De organisatie van gesprekken: een pragmatische studie van minimale interaktieve taalvormen*, PhD thesis, Nijmegen.

Odijk, J.(2003), Towards a standard for multi-word expressions. ISLE Project Report, February 2003, http://lingue.ilc.cnr.it/EAGLES96/isle/clwg_doc/ISLE_D6.1.zip.

Onrust, M., Verhagen, A. and Doeve, R.(1993), *Formuleren*, Bohn Stafleu Van Loghum, Houten.

Oostdijk, N.(2000), The Spoken Dutch Corpus. Overview and first evaluation, *in* M. Gavralidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (eds), *Proceedings of the second International Conference on Language Resources and Evaluation*, ELRA, Paris, pp. 887–893.

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

Poß, M.(2005), Towards an implementation of constructions, *Proceedings of the Sixth Annual High Desert Linguistics Society Conference*. to appear.

Postma, G.(2001), Negative polarity and the syntax of taboo, *in* J. Hoeksema, H. Rullmann, V. Sánchez Valencia and T. van der Wouden (eds), *Perspectives on Negation*, John Benjamins, Amsterdam, pp. 283–330.

Riehemann, S. Z. and Bender, E.(2001), Absolute constructions: On the distribution of predicative idioms, *in* S. Bird, A. Carnie, J. D. Haugen and P. Norquest (eds), *WCCFL 18 Proceedings*, Cascadilla Press, Somerville, pp. 476–89.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D.(2001), Multiword expressions: A pain in the neck for NLP, LinGO Working Paper No. 2001-03 (CSLI Linguistic Grammars Online (LinGO) Lab, Stanford University); also in Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico City, Mexico, pp. 1-15.

Sag, I. A., Wasow, T. and Bender, E.(2003), *Syntactic Theory: A formal introduction*, 2 edn, CSLI, Stanford.

Schiffrin, D.(1987), *Discourse markers*, Cambridge University Press, Cambridge.

Smadja, F. A. and McKeown, K. R.(1990), Automatically extracting and representing collocations for language generation, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 252–259.

Sprenger, S. A.(2003), *Fixed expressions and the production of idioms*, PhD thesis, Nijmegen.

van der Wouden, T.(2001), Collocational behaviour in non content words, *in* B. Daille and G. Williams (eds), *COLLOCATION: Computational Extraction, Analysis and Exploitation. Proceedings of a Workshop during the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter, Toulouse, France, July 7th*, CNRS – Institut de Recherche en Informatique de Toulouse, and Université de Sciences Sociales, Toulouse, France, pp. 16–23.

van der Wouden, T.(2002), Particle research meets corpus linguistics: on the collocational behavior of particles, *in* T. van der Wouden, A. Foolen and P. Van de Craen (eds), *Particles. Belgian Journal of Linguistics 16*, John Benjamins, Amsterdam, pp. 151–174.

Verhagen, A.(2003), The Dutch *way*, *in* A. Verhagen and J. van de Weijer (eds), *Usage-Based Approaches to Dutch. Lexicon, grammar and discourse*, Lot, Utrecht, pp. 27–57.

Villada Moirón, B.(2005), *Data-driven Identification of Fixed Expressions and their Modifiability*, PhD thesis, Groningen.

Weinreich, U.(1969), Problems in the analysis of ididioms, *in* J. Puhvel (ed.), *Sub-*

stance and structure of language, University of California Press, Berkeley and Los Angeles, pp. 23–81. (reprinted in *On Semantics*, 1980).

Wray, A.(2002), *Formulaic Language and the Lexicon*, Cambridge University Press, Cambridge.

Zadrony, W.(1994), From compositional to systematic semantics, *Linguistics and Philosophy* **17**, 329–342.

acknowledgement

The research reported on here is carried out within the framework of the VIDI-project *Dutch as a construction language*, financed by NWO, the Dutch Organization for Scientific Research (project number 276-70-003), and Leiden University.