A Sign Expansion Approach to Dynamic, Multi-Purpose Lexicons

Jon Atle Gulla
GMD - IPSI
Dolivostraße 15
D-64293 Darmstadt, Germany
gulla@gmd.de.

Sjur Nørstebø Moshagen
Computing Centre for the Humanities
Harald Hårfagres gt. 31,
N-5007 Bergen, Norway
sjur.moshagen@hd.uib.no

Abstract

Two problematic issues in most lexicon systems today are their size and restricted domain of use. In this paper, we introduce a new approach to lexical organization that leads to more compact and flexible lexicons. The lexical entries are conceptual/phonological frames rather than word entries, and a number of expansion rules are used to generate entries of actual words from these frames. A single frame supports not only all forms of a word, but also words of different categories that are derived from the same semantic basis. The whole theory is now being implemented in the TROLL lexicon project.

1 Introduction

Due to the complexity and wide coverage of lexical information, full-fledged lexicon systems easily grow undesirably big and must cope with intricate nets of dependencies among lexical items. For keeping the speed of access at a satisfactory level, lexical information is often repeated in different entries to reduce the number of consultations needed for a single user query. This simplifies and speeds up the access of lexical information, but also blows up the size of the lexicon and leads to huge maintenance problems. In many cases, it also clutters the lexicon structure, so that important lexical relationships and generalizations are lost.

Structuring the lexicon in inheritance hierarchies opens for more compact lexicon representations. So far, lexicons have been structured in syntactic inheritance hierarchies, in which more or less abstract syntactic classes form the upper nodes and actual words are associated with the leaf nodes (Flickinger and Nerbonne, 1992; Russell et al., 1992). However, the nature and number of these abstract syntactic classes are not very clear, and it seems difficult to come up with a sound method for how to decide on such classes. At the same time, there are also good reasons for assuming a similar hierarchy based on semantic properties (Hellan and Dimitrova-Vulchanova, 1994). Representing many competing hierarchies in the lexicon is a problem in itself and is here even more problematic as there are many complex relationships between semantic and syntactic properties (Gropen et al., 1992; Hellan and Dimitrova-Vulchanova, 1996).

Another problem is related to the notions and structures adopted in the lexicon systems. Most lexicons today are constructed within the framework of some syntactic theory. This theory guides the structuring of lexical information and also decides what information should be available to the user (Andry et al., 1992; Flickinger and Nerbonne, 1992; Mel'čuk and Polguère, 1987; Russell et al., 1992; Krieger and Nerbonne, 1991). Some lexicon systems try to be reasonably theory-independent, though they still have to adopt some basic syntactic notions that locate them into a family of theories (Goñi and González, 1995; Grimshaw and Jackendoff, 1985; Grishman et al., 1994).

The *Sign Expansion Approach* forms a basis for creating non-redundant lexicon systems that are structured along semantic lines. The stored lexical entries are sign frames rather than actual words, and a whole system of expansion rules and consistency rules are used to generate dynamic entries of words that contain all the necessary semantic, syntactic, and morphological information.

In Section 2, we give a brief introduction to a sign expansion theory called the Sign Model. Section 3 explains the use of lexical expansion rules, whereas some concluding remarks and directions for further work are found in Section 4.

$$PAINT \rightarrow \left\{ \begin{array}{l} paint_V \rightarrow \left\{ \begin{array}{l} paint\langle(\uparrow \textsc{subj})\rangle \\ paint\langle(\uparrow \textsc{subj})(\uparrow \textsc{obj})\rangle \rightarrow paint\langle(\uparrow \textsc{subj})(\uparrow \textsc{obj})(\uparrow \textsc{xcomp})\rangle \\ paint\langle(\uparrow \textsc{subj})(\uparrow \textsc{obl})\rangle \end{array} \right. \\ \\ paint_N \end{array} \right.$$

Figure 1: The stored frame *PAINT* is expanded into actual words with syntactic properties.

2 The Sign Model

In the sign expansion approach, the lexicon is viewed as a dynamic rule system with lexical frames and various kinds of expansion rules. The *Sign Model (SM)* by Hellan and Dimitrova-Vulchanova (Hellan and Dimitrova-Vulchanova, 1994) is a semantically based sign expansion theory and is used as the lexical basis of our lexicon. It posits an abstract level of sign representation that is not associated with any word classes and establishes a framework, within which word relationships as well as relationships between different kinds of linguistic properties can be described in a systematic way. At the abstract level of representation, one defines conceptual/phonological frames that underly the actual words found in a language. The frames combine with lexical expansion rules to create dynamic entries of actual words with morphological and syntactic properties, as illustrated by the LFG representations in Figure 1. No particular syntactic terminology is assumed, since the theory is intended to fit into any syntactic theory.

2.1 Minimal Signs

The conceptual/phonological frame, which is referred to as a *minimal sign*, is made up of a semantic (conceptual) part and a realizational part. As we do not have very much to say about phonological representations here, we assume in the following that the realizational part is a simple graphemic representation. The semantic part is a conceptual structure of the sign, which is to capture all grammar-relevant aspects of its meaning. The meaning of a sign is analyzed as a situation involving a number of participants (also called arguments), and these participants as well as the situation as a whole are modeled in terms of aspectual values, semantic roles, criterial factors, and realizational and selectional properties.

Consider the minimal sign *PAINT* in Figure 2, which is the lexical entry underlying the related words $paint_V$, $paint_N$, $painting_N$, $paintable_A$, etc. The realizational part is the string "paint", whereas the semantic part denotes a situation with two arguments, indexed as 1 and 2. The aspectual value (– *punctual*) describes the situation as durative, whereas the selectional restriction DIM states that argument 2 is to serve as some two-dimensional surface. Argument 1, the painter, possesses the semantic roles SOURCE and CONTROLLER. SOURCE means that this argument is the source of energy for the force involved in a painting process, whereas CONTROLLER indicates that the argument is in control of the process. Correspondingly, argument 2 is the entity on which the force is used (LIMIT) and the entity being controlled by argument 1 (GOAL). Argument 2 is also given the MONOTONIC role, which means that it undergoes some monotonic change in the course of painting. The change, of course, is that the surface is gradually covered by some paint. Each semantic role is further characterized by means of a criterial factor that imposes certain role-related observational properties on the argument. Specifying SOURCE and LIMIT as *coloring* means that the painter's use of force involves some observable actions that identifies him as painting, and that the surface being painted is recognizable from the same force. The gradual covering of the surface with paint, which is modeled by MONOTONIC, is also of the *coloring* type, since we can verify the covering by looking at the surface. CONTROLLER's and GOAL's factor *noncriterial* means that no particular observable behavior is required for an argument to play these

$$\begin{array}{ll} \textbf{Real:} & \text{“}paint\text{”} \\ \textbf{Sem:} & \left[\begin{array}{l} -\ punctual \\ \left[\begin{array}{ll} \textsc{source} & coloring \\ \textsc{controller} & noncriterial \end{array} \right]_1 \\ \left[\begin{array}{ll} \textsc{dim} & \textit{2-dim} \\ \textsc{limit} & coloring \\ \textsc{goal} & noncriterial \\ \textsc{monotonic} & coloring \end{array} \right]_2 \end{array} \right] \end{array}$$

Figure 2: Stored entry for minimal sign *PAINT*.

$$\begin{array}{ll}\textbf{Real}: & \textit{“walk”} \\ \textbf{Sem}: & \left[\begin{array}{l} -\,\textit{punctual} \\ \left[\begin{array}{ll} \text{SOURCE} & \textit{coloring} \\ \text{CONTROLLER} & \textit{noncriterial} \\ \text{MONOTONIC} & \textit{1-dim} \end{array}\right]_{1} \end{array}\right]\end{array}$$

Figure 3: Stored entry for minimal sign *WALK*.

particular roles. In general, the criterial factors affect the implicitation of arguments in syntactic expressions (e.g. argument 2 in *Jon painted*) and the introduction of new ones (e.g. *red* in *Jon painted the house red*).

As shown by the lexical entry of *WALK* in Figure 3, naturally intransitive verbs are rooted in minimal signs with only one conceptual argument. The argument of *WALK* is a SOURCE and a CONTROLLER, and it undergoes a monotonic development with respect to some one-dimensional path. In a sentence like *Jon walked to the school*, the phrase *to the school* describes this monotonic development of argument 1. *Away* in *Jon walked away* is another optional constituent that can describe argument 1's movement along a one-dimensional path.

2.2 Lexical Rules

The general format of the expansion rules is as follows:

(1) *X* IF *Y* COMPOSITION *S*

X contains the information to be added and *Y* the requirement for using the rule. *S* concerns the structure on which the rule is used and specifies which parts of this structure should be considered by the rule. Interpretationally, the rule in (1) can be applied on a structure *Z* if *Y* is a substructure of *Z* and *X* unifies with the selection of *Z* specified in *S*. The result of the operation is exactly this unified structure, and the operation itself is referred to as a derivation. If the whole lexical entry is to be addressed by the rule, the COMPOSITION part is omitted in the rule specification. Similarly, if the IF *Y* part is not present, it means that there is no requirement for using the rule. The expansion rules fall into five categories, depending on what kind of information they insert into the lexical representations: (1) Morpho-syntactic augmentations, (2) inflections, (3) conceptual expansions, (4) syntactic mappings, and (5) compositions.

Morpho-syntactic augmentation rules add a word category and an inflectional paradigm to a minimal sign. The morpho-syntactic augmentation rule shown in Figure 4(a), for example, derives the basic entry for the verb $\textit{paint}_V$ from the minimal sign *PAINT*.

Assuming that the lexical entry has already been given a word class and a paradigm, the *inflectional rule* expands the graphemic representation into a particular inflected word form. The rule in Figure 4(b) expands the basic entry for $\textit{paint}_V$ into the more specialized entry for the past form $\textit{painted}_V$. The inflectional rules are grouped together into paradigms that are associated with the appropriate words (e.g. *v1* is linked to $\textit{paint}_V$).

Conceptual expansion rules are rules that extend the semantic part of the signs without combining them with other sign structures. These rules are semantically conditioned and typically explain how a particular sign can support a variety of subcategorization frames. The rule in Figure 4(c) shows how a resultative construction like *Jon painted the wall red* is supported by a minimal sign like *PAINT*. If the conceptual structure contains an argument that undergoes some monotonic development, the conceptual structure can be expanded with a new argument that serves as the medium for this development and has a dimension matching the criterial property of the MONOTONIC role. When an argument is a medium for some other argument, it means that its monotonic development is manifested or materialized through this other argument. Hence, as argument 2 of *PAINT* has a MONOTONIC role, the rule is able to add an argument that describes the resulting monotonic change of the surface being painted. The realization of this argument as an adjective (like *red*) comes from the fact that the new argument is of dimension *coloring*. For a minimal sign like *WALK* (see Figure 3), which contains an argument (the walker) that monotonically moves along some one-dimensional path, the rule adds a new argument of dimensionality *1-dim*. The medium must then describe a one-dimensional path, as for example *to the school* in *Jon walked to the school*.

Syntactic mapping rules are rules that derive syntactic properties from conceptual structures. Since no special syntactic notions are assumed, we must here decide on an existing syntactic theory before the mapping rules can be defined. The rule shown in Figure 4(d) is based on Gulla's rules (Gulla, 1994) for mapping from SM conceptual structures to LFG grammatical functions (Kaplan and Bresnan, 1982). It states that if a verb is used

$$\begin{array}{ll} \textbf{Cat:} & \text{V} \\ \textbf{Infl:} & \left[\text{paradigm:} \quad \text{v1} \right] \end{array}$$

(a)

$$\begin{array}{ll} \textbf{Infl:} & \left[\text{form:} \quad \text{past} \right] \\ \textbf{Real:} & \text{insert } \textit{“ed”} \text{ at end} \end{array}$$

(b)

$$\begin{array}{ll} \textbf{Sem:} & \left[\begin{array}{l} \left[\text{MONOTONIC}_i \quad \alpha \right] \\ \left[\begin{array}{ll} \text{DIM} & \alpha \\ \text{MEDIUM} & \dashrightarrow \text{i} \end{array} \right]_j \end{array} \right] \\ \text{IF} & \\ \textbf{Sem:} & \left[\left[\text{MONOTONIC} \quad \alpha \right] \right] \end{array}$$

(c)

$$\begin{array}{ll} \textbf{Syn:} & \left[\text{XCOMP}_i \quad [\,] \right] \\ \text{IF} & \\ \textbf{Sem:} & \left[\begin{array}{l} + \textit{completed} \\ \left[\begin{array}{ll} \text{DIM} & \textit{coloring} \\ & \quad \text{OR} \\ & \textit{existence} \\ \text{MEDIUM} & \dashrightarrow \text{j} \end{array} \right]_i \end{array} \right] \end{array}$$

(d)

$$\begin{array}{ll} \textbf{Sem:} & [\,]_k \\ \text{IF} & \\ \textbf{Sem:} & \left[\begin{array}{l} - \textit{punctual} \\ \left[\text{CONTROLLER} \quad - \right]_i \\ \left[\text{ROLE:} \quad \alpha \right]_j \end{array} \right] \end{array}$$

COMPOSITION main Suffix
where $\alpha \neq$ *noncriterial*

(e)

Figure 4: (a) Morpho-syntactic augmentation. (b) Inflectional rule. (c) Conceptual expansion. (d) Mapping rule. (e) Compositional rule.

in a completed sense[1], MEDIUM arguments of dimensionality *coloring* or *existence* can be mapped onto the XCOMP function. Used together with rule 4(c) on *PAINT*, it introduces an XCOMP element that describes the resulting state of the surface being painted. A similar approach to the assignment of syntactic functions in LFG can be found in (Alsina, 1993).

The *compositional rules* combine two sign structures and create a new compound structure that includes parts of both of them. The rule in Figure 4(c) uses a suffix to create a noun that refers to some controlled, durative activity. Except for the control and duration requirement, the conceptual structure must also contain a *criterially anchored* argument, i.e. an argument that includes at least one semantic role that is not *noncriterial*. The COMPOSITION part says that there are two structures involved, a *main* structure and a *suffix* structure, whereas the expansion part turns the whole conceptual structure into an argument *k*. On the basis of the minimal signs *PAINT* and *WALK*, the rule can create the nouns $painting_N$ and $walking_N$.

3 The Expanding Lexicon

In a sign expansion lexicon system, we must distinguish between *stored* lexical entries and *generated* lexical entries. The stored entries are all minimal signs, and they are usually not very interesting to the lexicon user. The generated entries are produced by combining stored entries with one or more expansion rules, and these entries are more or less elaborate specifications of actual words. A simple generated entry is the result of combining the minimal sign *PAINT* in Figure 2 with the morpho-syntactic augmentation rule in Figure 4(a). This yields the basic verb entry $paint_V$, which does not contain any information about syntactic realization. More elaborate entries are then generated by expanding the $paint_V$ entry with the different subcategorization frames that are possible for $paint_V$. For a user requesting information from the lexicon, the stored entries may be completely hidden and only the elaborate generated ones may be made available.

Consider the rather elaborate entry in Figure 5, which represents the past form *painted* used in the following resultative construction:

[1] Following the ideas of telicity in (Depraetere, 1995), we define a clause to be *completed* if it reaches a natural or intended endpoint. A non-repetitive resultative construction is always completed, whereas constructions like *Jon is painting* and *Jon paints every day* are incompleted.

$$
\begin{array}{ll}
\textbf{Cat}: & V \\
\textbf{Infl}: & \begin{bmatrix} \text{paradigm:} & \text{v1} \\ \text{form:} & \text{past} \end{bmatrix} \\
\textbf{Real}: & \textit{``painted''} \\
\textbf{Sem}: & \begin{bmatrix}
\textit{- punctual} \\
\textit{+ completed} \\
\begin{bmatrix} \textsc{source} & \textit{coloring} \\ \textsc{controller} & \textit{noncriterial} \end{bmatrix}_1 \\
\begin{bmatrix} \textsc{dim} & \textit{2-dim} \\ \textsc{limit} & \textit{coloring} \\ \textsc{goal} & \textit{noncriterial} \\ \textsc{monotonic}_4 & \textit{coloring} \end{bmatrix}_2 \\
\begin{bmatrix} \textsc{dim} & \textit{coloring} \\ \textsc{medium} & \rightarrow 4 \end{bmatrix}_3
\end{bmatrix} \\
\textbf{Syn}: & \begin{bmatrix} \textsc{subj}_1: & [\,] \\ \textsc{obj}_2: & [\,] \\ \textsc{xcomp}_3: & [\,] \end{bmatrix}
\end{array}
$$

Figure 5: Generated entry for resultative use of *painted*$_V$.

(3) *Jon painted the house red.*

The entry specifies a particular word form, contains a conceptual structure with three arguments, and lists the syntactic functions realizing these arguments. Indexing SUBJ with 1 means that argument 1 of the conceptual structure is to be realized as the subject. The whole entry is generated by a series of derivations, where each derivation adds a piece of information to the final lexical entry. Starting with the minimal sign *PAINT*, we use the rules in Figure 4(a) and 4(b) to generate a simple entry for *painted*$_V$. Then we expand the conceptual structure into a completed description (*+ completed*) using a rule called *Completed* and apply the rule in Figure 4(c) to add a third argument. The syntactic functions are added by the rule in Figure 4(d) plus two rules that we here can call *Subj1* and *Obj1*. Subj1 assigns the SUBJ function to arguments that contain SOURCE or CONTROLLER roles, whereas Obj1 requires a *+ completed* description and assigns the OBJ function to arguments that have a MONOTONIC role. The generation of the lexical entry in Figure 5, thus, can be written as the following derivational sequence:

(4) *PAINT* ++ 4(a) ++ 4(b) ++ Completed ++ 4(c) ++ Subj1 ++ Obj1 ++ 4(d)

When the system is to create a derivational sequence like that, we first have to indicate which morpho-syntactic rule to use. The system then chooses the correct inflectional paradigm, and it can start trying out the different expansion rules to generate complete lexical entries. The search space for this is restricted, since the rules are semantically conditioned and monotonic, and well-formedness conditions decide when to stop expanding the structure.

$$
\begin{array}{ll}
\textbf{Cat:} & \text{N} \\
\textbf{Infl:} & [\ \text{paradigm:} \quad \text{n1}\ ] \\
\textbf{Real:} & \textit{``paint(ing)''} \\
\textbf{Sem:} & \begin{bmatrix}
\textit{- punctual} \\
\begin{bmatrix} \textsc{source} & \textit{coloring} \\ \textsc{controller} & \textit{noncriterial} \end{bmatrix}_1 \\
\begin{bmatrix} \textsc{dim} & \textit{2-dim} \\ \textsc{limit} & \textit{coloring} \\ \textsc{goal} & \textit{noncriterial} \\ \textsc{monotonic} & \textit{coloring} \end{bmatrix}_2
\end{bmatrix}_3
\end{array}
$$

Figure 6: Lexical entry for suffix *ing*$_N$ and generated entry for *painting*$_N$.

In a similar vein, the noun *painting*$_N$ (referring to a painting process) is derived from the minimal sign *PAINT* and the suffix *ing*$_N$. The compositional rule from Figure 4(e) combines these two structures and produces the lexical entry shown in Figure 6. **Category** and **Inflection** stem from *ing*$_N$, **Realization** is a combination of the values in *PAINT* and *ing*$_N$, and **Semantics** is the minimal sign's conceptual structure expanded into a complex argument indexed as 3. Instead of storing two entries for *paint*$_V$ and *painting*$_N$ — that partly contain the same information — we derive the entries dynamically from a single *PAINT* entry.

4 Conclusions

The Sign Model (SM) gives a theoretical foundation for structuring lexical information along semantic lines. It prescribes a strong semantic basis and suggests various kinds of expansion rules for generating complete word entries. The sign expansion approach is now used as a basis for the TROLL lexicon project in Trondheim. In this project, a formalism for lexical representation as well as mechanisms for executing lexical rules are implemented in LPA Prolog (Gulla and Moshagen, 1995). A lexicon of Norwegian verbs is under construction, and SM-based analyses of English, German, and Bulgarian have been used in the design of the lexicon (Hellan and Dimitrova-Vulchanova, 1996; Pitz, 1994). Due to speed concerns, the stored entries and the expansion rules are in the TROLL lexicon supplemented with indexes that refer to well-defined derivational sequences for complete word entries. The work in the TROLL project is now concentrated on the construction of a complete lexicon for Norwegian, and this work is also to serve as an evaluation of both the lexicon structures and the Sign Model. The theory is still at a development stage when it comes to psychological and perceptional matters, even though some suggestions have been made (Gulla, 1994). The future work also includes establishing proper interfaces to various syntactic theories, so that the system can be integrated with existing parsers and generators.

References

Alsina, A. (1993). *Predicate Composition: A Theory of Syntactic Function Alternations.* Ph. D. thesis, Stanford University, San Fransisco.

Andry, F., N. M. Fraser, S. McGlashan, S. Thornton, and N. J. Youd (1992). Making DATR Work for Speech: Lexicon Compilation in SUNDIAL. *Computational Linguistics 18*(3), 245-268.

Coopmans, Everaert, and Grimshaw (Eds.) (1996). *Lexical Specification and Insertion.* Lawrence Erlbaum Ass., Inc.

Depraetere, I. (1995). (Un)boundedness and (A)telicity. *Linguistics and Philosophy 18*, 1-19.

Flickinger, D. and J. Nerbonne (1992). Inheritance and Complementation: A Case Study of Easy Adjectives and Related Nouns. *Computational Linguistics 18*(3), 269-310.

Goñi, J. M. and J. C. González (1995). A framework for lexical representation. In *AI95: 15th International Conference. Language Engineering 95*, Montpellier, pp. 243-252.

Grimshaw, J. and R. Jackendoff (1985). Report to the NSF on grant IST-81-20403. Technical report, Waltham, Department of Linguistics, Brandeis University.

Grishman, R., C. Macleod, and A. Meyers (1994). Comlex Syntax: Building a Computational Lexicon. In *Proceedings of the International Conference on Computational Linguistics (COLING-94)*, Kyoto.

Gropen, J., S. Pinker, M. Hollander, and R. Goldberg (1992). Affectedness and Direct Objects: The role of lexical semantics in the acquisition of verb argument structure. In B. Levin and S. Pinker (Eds.), *Lexical & Conceptual Semantics*, Cognition Special Issues, Chapter 6, pp. 153-196. Elsevier Science Publishers.

Gulla, J. A. (1994). A Proposal for Linking LFG F-structures to a Conceptual Semantics. Master's thesis, Department of Linguistics, The University of Trondheim, Trondheim.

Gulla, J. A. and S. N. Moshagen (1995, January). Representations and Derivations in the TROLL Lexicon. In H. Lødrup, I. Moen, and H. G. Simonsen (Eds.), *Proceedings of The XVth Scandinavian Conference of Linguistics*, Oslo.

Hellan, L. and M. Dimitrova-Vulchanova (1994, July). Preliminary Notes on a Framework for 'Lexically Dependent Grammar'. Lecture series at International Summer Institute in Syntax, Central Institutue of English and Foreign Languages, Hyderabad, India.

Hellan, L. and M. Dimitrova-Vulchanova (1996). Criteriality and Grammatical Realization. To appear in (Coopmans et al., 1996).

Kaplan, R. M. and J. Bresnan (1982). Lexical-Functional Grammar: A Formal System for Grammatical Representation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*, Chapter 4, pp. 173-281. MIT Press.

Krieger, H. U. and J. Nerbonne (1991). Feature-Based Inheritence Networks for Computational Lexicons. Technical Report DFKI-RR-91-31, German Research Center for Artificial Intelligence (DFKI), Saarbrucken.

Mel'čuk, I. and A. Polguère (1987). A Formal Lexicon in Meaning-Text Theory (Or How to Do Lexica with Words. *Computational Linguistics 13*(3-4), 261-275.

Pitz, A. (1994). *Nominal Signs in German.* Ph. D. thesis, Department of Linguistics, University of Trondheim, Trondheim.

Russell, G., A. Ballim, J. Carroll, and S. Warwick-Armstrong (1992). A Practical Approach to Multiple Default Inheritance for Unification-Based Lexicons. *Computational Linguistics 18*(3), 311-337.