PAPERS FROM THE FIFTH SCANDINAVIAN CONFERENCE OF COMPUTATIONAL LINGUISTICS

Helsinki, December 11—12, 1985

Edited by Fred KARLSSON

1986

Lars Borin
Uppsala University
Center for Computational Linguistics
Box 513
S-75 1 20 UPPSALA
Sweden

WHAT IS A LEXICAL REPRESENTATION?

1. Introduction

In this paper, I will discuss one aspect of the lexicon, namely its morphological organisation.

For about two years I have been working with Koskenniemi's two-level model (Koskenniemi 1983), on a Polish two-level description. In this work, I have become more and more interested in the formalism itself, something that has tended to push work on the language description into the background. It seems to be the case that in most concrete two-level descriptions, the rule component is forced to carry too heavy a burden in comparison with the lexicon, perhaps because the lexicon is very simple as to its implementation. Like many other lexical systems in computer applications it is implemented as a tree, with a root node and leaves, from which the lexical entries are retrieved when the analysis routine has traversed the tree. The two-level lexicon is a bit more sophisticated than this, however, in that there is not only one, but several lexicon trees, the so called mini-lexicons. The user links the minilexicons into a whole, most often into a root lexicon and a number of suffix lexicons. In Hockett's terminology, we could speak of an Item-and-Arrangement (IA) model (Hockett 1958, pp 386ff). Elsewhere I have characterized this kind of lexicon system as most suited for describing suffixing languages with a comparatively high degree of agglutination (Borin 1985, p 35); This is mainly due to the fact that the system works according to what Blåberg (1984, p 61) aptly has termed the "forget-where-you-came-from"

[Publications No. 15. University of Helsinki, Department of General Linguistics. 1986. 25–34]

principle. It makes it very hard to describe discontinuous dependencies within word forms; one is forced either to duplicate a considerable amount of information in the lexicon or, at least in the two-level model, let the two-level rules take care of some of the morphotax description, which is not their primary purpose (cf. Karttunen 1984, pp 178-181).

In the minilexicons, stems and affixes are grouped into conceptually motivated collections and the links between them describe the morphotax of the language in question. It is not only morphotax, however, that is described by the lexicon linkages: alternations that comprise more than a single segment are often taken care of in separate minilexicons, and in this way one tends to mix linguistically motivated categories with categories that are introduced to ease the work of the lexicon writer; of. the "technical stems" in Hellberg's system for Swedish morphology (Hellberg 1978, p 13ff; Doherty et al 1986).

2. The problem

It seems that one would need a more sophisticated lexicon system to take care of a number of important morphological phenomena that, for different reasons, should not fall in the domain of morphophonological rules, be they of the generative kind or two-level rules. Here is a representative, even if not exhaustive, list of the kind of phenomena I have in mind:

- Discontinuous morphs, like in Sw. förstora 'enlarge'; Ger. gesagt 'said'; Po. najstarszy 'oldest'.
- Inflection Or certain compound types, like in Fi. kolmekymmentä-viisi 'thirty-five', Adessive kolmellakymmenelläviidellä; Ru. vagon-restoran 'restaurant car', Genitive vagona-restorana.
- Reduplication, like in Gr. lelpô 'leave', Perfect leloipa.

- Suprasegmental features, like accent, as far as they are reflected in the morphology, like in Ru. bol'šój 'big', ból'šij 'bigger'.

I am currently working on a lexicon system for the two-level model that should be capable of dealing with these phenomena, in a formalism that is intentionally in line with traditional linguistic taxonomy. In my view, one should not aim at a clear separation of the conceptual organization of the lexicon description from implementation details. The former is important from a theoretical linguistic point of view, while the status of the latter is at best uncertain.

3. The Lexicon Formalism

Fig. 1 shows a small sample lexicon in the format I propose to use<1>. As can be seen, morphotax is explicitly specified, in the form of regular expressions. The first morphotax specification in fig. 1 states that the category Adj (for adjective) has the constituents Stem, Comp (comparative suffix) and Final (gender/number/case port-manteau morphs), in the order they are given. The constituent Comp is optional, which is signalled by the parentheses around it. The category Adj has two alternative constituent structures, which are separated by a comma in the specification. The other possible structure given in fig. 1 is: Sup (superlative prefix), Stem, Comp and Final, where none of the parts are optional. The constituents in the morphotax specifications correspond to (groups of) minilexicons, where the minilexicons have names of the form 'category.constituent', e.g. things that can fill the Stem position in category Adj are found in

---

<1> Since this particular lexicon is used only to illustrate the lexicon formalism, I have taken the liberty to mix Polish (adjectives) and Russian (noun-noun compounds) in it.

```
MORPHOTAX
     Adj = Stem (Comp:(Degree:Pos)):(Degree:Comp) Final,
           Sup Stem Comp Final;
     Noun = Stem (Case:(Case:Nom)),
            Stem (Case:(Case:Nom)) Hyphen Stem (Case:(Case:Nom));
END

LEXICON Adj.Stem :(Cat:AdJ) =
     ENTRIES
        star (Type:Qual),
        now (Type:Qual),
        rad (Type:Short);
LEXICON Adj.Sup :(Degree:Sup) =
     ENTRIES
        naj+ ;
LEXICON Adj.Comp :(Type:Qual ) =
     ENTRIES
        +sz ;
LEXICON Adj.Final :(Type:0ual) =
     ENTRIES
        +y (Numb:Sg Gend:Masc Case:Nom),
        +a (Numb:Sg Gend:Fem Case:Nom),
        +e (Numb:Sg Gend:Neut Case:Nom);
LEXICON Adj.Final :(Type:Short) =
     ENTRIES
        O (Numb:Sg Gend:Masc Case:Nom),
        +o (Numb:Sg Gend:Neut Case:Nom),
        +a (Numb:Sg Gend:Fem Case:Nom);
LEXICON Noun.Stem :(Cat:Noun) =
     SUBLEX Dim :(Form:Dim) =
        Ek ;
     ENTRIES
        vagon ,
        restoran ,
        zegar Dim,
        ogon Dim,
        dub ,
        velikan ;
LEXICON Noun.Case =
     ENTRIES
        +a (Case:Gen ),
        +u (Case:Dat),
        +e (Case:Loc);
LEXICON Noun.Hyphen =
     ENTRIES
        - ;
END
```

Figure 1.

the lexicon(s) Adj.Stem<2>. Thus, the minilexicons are intended to group morph(eme)s of similar morphological categories. In addition, there is the concept of sublexicons. These were introduced for the sake of space and work economy. In an inflectional morphology, the sublexicons can be used to collect e.g. word endings (like derivational suffixes) with special inflections, like in the lexicon Noun.Stem in fig. 1, where the sublexicon Dim contains a diminutive suffix.

To account for phenomena like discontinuous morphs and agreement phenomena within words there is another mechanlam apart from the morphotax specifications, namely feature-value graphs (directed acyclic graphs, or DAG:s), that are checked for mutual consistency and added on to as the analysis routine moves through the lexicon. The DAG:s can appear at various points in the lexicon:

1) On a constituent in the morphotax specifications.

2) In a minilexicon or sublexicon as a whole.

3) In an individual lexical entry.

The adding-on procedure is a kind of unification<3>, as described in e.g. Karttunen 1984. This ensures that word forms like Ru. *vagon-restorana will not get any analysis, since (Case:Nom) in the first part will not unify with (Case:Gen) in the second part. The final DAG of a successful analysis is produced as part of the output.

The morphotax speolfications and the DAG:s together succeed very nicely in capturing the traditional linguistic concept of markedness:

<2> I.e., there may be arbitrarily many lexicons with the label Adj.Stem, grouped according to e.g. declensions.

<3> Actually, it is unification split up into two separate steps: a compatibility check and, if this succeeds, unification with copying.

with optional constituents one may specify a DAG that is added to the structure being built only if the constituent is not included in the analysis. This mechanism is used for aggignlug the positive degree to adjectives in the lexicon in fig. 1; if no material is taken from the lexicon Adj.Comp during analysis, the analysed adjective gets the positive degree by default.

Like the variuos versions we have of the two-level system, this extension to its lexicon is written in PASCAL.

4. Future Plans

Tne preceding section gave a brief overview over the current status of the lexicon system. Among the things that have not yet been implemented, but are due for inclusion in the system in the near future are:

- Negative value specifications in the DAG:s, e.g. (Case:-Dat).
- Disjunction of values in the DAG:s, e.g. (Case: (Ack OR Gen)).
- Full regular expression capability in the morphotax specifications.

For handling the last two problems in the 11st in section 2, I am currently exploring the possibility of using a formalism that is reminiscent of and largely inspired by the ones used in autosegmental phonology (Goldsmith 1976; McCarthy 1981; 1982), or Aronoff's (1976) word formation rules (see fig. 2). The main idea in both these approaches (even though the details differ) could be interpreted as a kind of constraint equation for lexical entries. Reduplication would be handled in something like the following manner: With the minilexicon for stems that reduplicate would follow a template specifying the kind of reduplication and the conditions under which it applies:


```
TEMPLATE :(Tense:Perfect) =
    C V <C V *>
    1 2  1 2
```

The notation is a cross between McCarthy's and Aronoff's. The part within angle brackets is a condition on the stem: it should begin with a CV sequence. The expression as a whole states that this initial CV sequence may be reduplicated if the word form can be interpreted as perfect tense. These templates would act as filters, or constraint mechanisms, between the lexical and surface representations, quite independent of and in parallel to any morphophonological rules in the description. What the exact form and power of these templates would be is not entirely clear at the moment, but material on linguistic universals, like the data on the possible forms of reduplicative constructions collected by Moravcsik (1978), obviously has an important role to play here.

+--+

(46) Adjective Reduplication Rule (Aronoff 1976:77)
C V C V X§
1 2 3 4 5 -> 1 2 3 4 1 3 4 5

(17) a. CVCCVC (kattab) b. μ (McCarthy 1981:388)

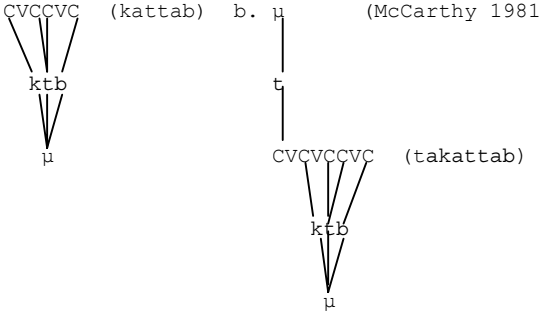

Figure 2.

+--+

The original motivation behind the development of autosegmental phonology was to account for phenomena like tone and accent, which are

relevant for the problem of accentually conditioned morphological alternations of the type cited last in the list in section 2.

5. New Problems

While the approach with feature-value graphs that are unified together solves some of the original problems in the list in section 2, they unfortunately introduce some new ones that, furthermore, are principally interesting.

One problem concerns word formation. Compounds of the type represented by the Finnish numerals can be handled very nicely with the proposed method, since the case agreement is enforced automatically in the lexicon. Some other kinds of word formation become very hard to handle in this framework, however. Category shifting affixes are prohibited at present, singe a stem with e.g. (Cat:Noun) will not unify with a suffix that tries to assign (Cat:Verb) to the derived word form. If one subscribes to the view of e.g. Aronoff, that word formation is a process that operates on words to form new words, where the category of the word in question may or may not change in the process, there must be some further mechanism to account for this.

Another problem is more speoifically connected with the framework this lexicon system was developed in, viz. two-level morphology. It is the problem of how the two-level rules should interact with the feature-value graphs. It is a characteristic trait of the two-level model that morphological features are segmentalized in the lexical representations (Nyman 1984, p 473), because this is the only way in which the two-level rules can access features that trigger some alternation, e.g. Tense:Perfect in the template in section 4 above. One could imagine that the morphological features, or rather, the DAG:s that contain them, be treated as lexical segments by the rules

i.e. be accessible only at specific points in the lexical representation. On the other hand, one could try to restrict the domain of the rules, something that is needed on independent grounds: some alternation types comprise only specific word categories, e.g. Swedish apophony, which concerns only verbs.

6. Conclusion

In conclusion, I would like to stress that even if the things mentioned above are an important part of a lexical representation, Any are still only a part of what is needed in a lexicon system for full-fledged natural language processing. In addition, the lexicon should reflect our knowledge as language users about productivity and prototypes in morphology. Certainly, the lexicon must also contain ~ flus structured semantic information.

References

Aronoff, M. 1976. Word Formation in Generative Grammar. Linguistic Inquiry Monograph 1. Cambridge, Mass. and London.

Blåberg, O. 1984. Svensk böjningsmorfologi - En tvånivåbeskrivning. Unpublished Master's Thesis. Unlversity of Helsinki, Dept. of General Linguistics.

Borin, L. 1985. Tvånivåmorfologi - Introduction och användar-handledning. UCDL-L-3. Uppsala University, Center for Computational Lingulstics.

Doherty, P., Rankin, I. & Wirén, M. 1936. Erfarenheter av en implementering av Hellbergs system for svensk morfologi. In this volume.

Goldsmith, J.A. 1976. Autosegmental Phonology. Indiana University Linguistics Club. Bloomington.

Hellberg, S. 1978. The Morphology of Present-Day Swedish. Stockholm.

Hockett, C.F. 1958 (1954). Two Models of Grammatical Description. In: Joos, M. (Ed.). Readings in Linguistics. New York.

Karttunen, L. 1983. KIMMO: A General Morphological Processor. Texas Linguistic Forum 22. Unlversity of Texas at Austin, Dept. of Lingulstics.

---------- 1984. Features and Values. Chapter 2 or Shieber, S., Karttunen, L. and Pereira, F.C.N. Notes from the Unification Underground. SRI International Technioal Note 327. Menlo Park, Calif.

Koskenniemi, K. 1983. Two-level Morphology - A General Computational Model for Word-Form Recognition and Production. Publications No. 11. University of Helsinki, Dept. of General Linguistics.

McCarthy, J.J. 1981. A Prosodic Theory of Nonconcatenative Morphology. Linguistic Inquiry, Vol. 12, No. 3.

---------- 1982. Prosodic Templates, Morphemic Templates and Morphemic Tiers. In: Hulst/Smith (eds.). The Structure of Phonological Representations, Part I. Dordrecht - Cinnaminson.

Moravcsik, E. 1978. Reduplicative Constructions. In: Greenberg, J. (ed.). Universals of Human Language, Vol. 3, Word Structure.

Nyman, M. 1984. Sananmuotojen tunnistuksen ja tuoton mallintamista (Review of Koskenniemi 1983). Virittäjä 4/1984. Helsinki.