Annette Östling Andersson
Dept of Romance Languages
Uppsala University

A TWO-LEVEL DESCRIPTION OF WRITTEN FRENCH

In his thesis, K. Koskenniemi presents a new model for morphological description, two-level morphology (1). The model has two components:

1) a lexicon component
2) a rule component

The lexicon component consists of lexicons for different morphological elements, such as stems and endings. Continuation classes constitute the links between the lexicons.

The rules specify equivalences between pairs of signs on two levels, a lexical level and a surface level.

lexical level n e z + s
surface level n e z 0 0

The two-level model has already been applied to a number of languages: Finnish (2), Swedish (3), English (4), and Polish (5) are some examples. My contribution is a description of the inflectional morphology of written French.

A very important matter to decide upon when setting up a two-level description is the following: which inflectional phenomena are to be treated in the lexicon system, and which ones are best described by rules?

In order to get a survey of the different types of inflectional phenomena that occur in French, the following classification is of help:

A two-level description of written French
Annette Östling Andersson
Proceedings of NODALIDA 1985, pages 195-202

a) flexion réguliere

livre + s --> livres
parl + e --> parle

i.e. regular inflection

b) variation réguliere

nez + s --> nez
mang + ons --> mangeons

i.e. deviation from regular inflection occurring for all words with a certain structure

c) variation irréguliere

complet + e --> complète
achet + e --> achète

i.e. deviation from regular inflection taking place only for some words with a certain structure

The rule component is designed for those inflectional phenomena that are productive, frequent and do not affect long sequences of signs (6). The stem lexicon together with minilexicons for affixes account for regular inflection, as well as for those phenomena that deviate from regular inflection, but do not depend on the phonological or morphological context. The stem lexicon can also be used to list irregular word forms.

Category a), regular inflection, should thus be described in the lexicon component, whereas category b), context-dependent deviation from regular inflection, is best described by rules. The lexicon component is best suited to take care of those phenomena that belong to category c), deviations from regular inflection that are not context-dependent.

I have used 15 lexicons of endings to describe the nominal morphology. Two of these take care of category a), while the other 13 cover category b).

Stem lexicon

```
grandeur    /NO   "N F";
gra  s/A          "gras A";
gratuit     /A    "A";
```

Regular inflection

```
LEXICON /NO         number: singular or plural
LEXICON /A          adjectives; points to /NO
```

Variations irrégulières

```
LEXICON /NOx        plural -x: chou - choux
LEXICON ail/NM      travail - travaux
LEXICON eau/A       beau - bel - belle
LEXICON ou/A        fou - fol - folle
LEXICON et/A        complet - complète
LEXICON et/ADJ      net - nette
LEXICON s/A         gros - grosse
LEXICON c/A         public - publique
LEXICON nc/A        blanc - blanche
LEXICON ng/A        long - longue
LEXICON gu/A        ambigu - ambiguë
LEXICON eur/A       trompeur - trompeuse
LEXICON teur/A      multiplicateur - multiplicatrice
```

If the number of words with a certain "variation irréguliere" is > 1, I have always decided to set up a lexicon instead of enumerating inflected forms in the stem lexicon.

In order to handle category b), "variation réguliere", there are six rules. I show them here in a simplified form.

1)	s,x,z + s --> s,x,z	nez + s --> nez
	e + e --> e	malade + e --> malade
2)	er + e --> ère	entier + e --> entière
3)	x + e --> se	heureux + e --> heureuse
	f + e --> ve	passif + e --> passive
4)	al + s --> aux	principal + s --> principaux
5)	el + e --> elle	naturel + e --> naturelle
	en,on +e --> enne,onne	moyen + e --> moyenne
6)	eu,au + s --> eux,aux	jeu + s --> jeux

The inflectional morphology of the verbs is more complicated than that of the nominals.

The French verbs are usually divided into three regular conjugations. There are also about 150 irregular verbs. The endings of the different tenses and moods differ in most cases between the conjugations, and no element always signalling for example person or number exists. The ending -rai, for instance, has as its signifié futur, 1st person, singular, and -irent passé simple, 3rd person, plural. Thus it is obvious that the number of lexicons for endings becomes large. By keeping the endings as constant as possible and, when necessary, using more than one stem, a good descriptional economy is achieved (7). A regular verb, such as parler, gets one entry in the stem lexicon, while an irregular verb, vivre for instance, gets more than one. Exceptions exist, though. So far I have encountered one irregular verb that needs only one entry, courir:

```
cour      V19  "courir";
parl      V1   "parler";
viv       V27  "vivre";
vi        V6 ;
véc       V11 ;
```

To each stem is attached a reference to a continuation class (V1, etc., above). The continuation class then points to the lexicons containing the endings possible to add to the stem. V1, for example, points to 15 lexicons. There are 62 continuation classes, 57 of which are needed for the description of the irregular verbs. The lexicons I use are:

Finite forms

8	lexicons for	indicatif présent
1		imparfait
3		futur
3		conditionnel
4		passé simple
2		subjonctif présent
4		subjonctif imparfait
4		impératif
1		the infix -iss-

All of these are not complete lexicons for all six persons. One of the lexicons for subjonctif présent, for instance, consists of the endings for the singular and the 3rd person in the plural, whereas the other one contains the endings for the 1st and 2nd person in the plural.

Infinite forms

4	lexicons for	infinitif
5		participe passé
1		participe présent
2		the inflection of the participles

Four rules take care of the "variation réguliere" among the verbs:

1)	c + a,o --> ca,co	plac + ons --> plaçons
2)	g + a,o --> gea,geo	mang + ons --> mangeons
3)	oy,uy + 9e --> oie,uie (8)	employ + 9e --> emploie
4)	= + 9e --> =e (9)	parl + 9e --> parle

The contexts specified within the two-level framework are word-internal. The two-level model analyses the words as isolated units, and gives all the possible analyses. In this respect it works very well for French, whose inflectional phenomena are all accounted for in a satisfactory way, with one small exception. A handful of nouns, such as arc-en-ciel get their plural -s after the first noun: arcs-en-ciel. These plurals have to be listed in the stem lexicon.

Homographies are very frequent in French, a language that can be characterized neither as analytic nor as synthetic. The homographs can be divided into two groups:

1) homographies between inflectional forms of a word

nez	nom masculin singulier
	nom masculin pluriel

2) homographies between parts of speech

maintenant	adverbe
	verbe participe présent

The homographies of both types are disambiguated by the syntagmatic relations of the sentence. A good deal of the informational contents is thus to be found in the relations between the words. Since the two-level model only is concerned with the isolated sequences of signs constituting words, the number of words with more than one analysis gets very high for French. This is not the case for languages such as Finnish, whose morphological elements carry more information.

Often the analysis resulting from the two-level model is quite satisfactory, but for the purpose of vocabulary studies the definition of the word as a sequence of signs separated by blanks is not adequate for a language like French. It becomes necessary to push the analysis further and disambiguate the homographies. For the same purpose, discontinuous signs have to be accounted for. In French they occur for example among the verbs (ai/ /parlé) and the negations (ne/ /pas). Since this process necessarily involves syntax, it is no longer within the scope of the two-level model.

Which homographs can be separated automatically, which segments have to be identified to achieve an analysis, and how are these segments identified? These are interesting questions to which I will try to find some answers in my future research. I will also investigate how the two-level model is to be combined with a subsequent syntactic analysis.

In order to test the two-level description, I have a corpus consisting of 102 signed editorials from the daily French newspaper Le Figaro. Together they constitute 50012 running words.

At present there are in the stem lexicon about 2830 nouns, proper names, adjectives and abbreviations, and about 1400 verbs,

i.e. all the words belonging to these categories up to 30000 running words.

Notes

(1) Koskenniemi, K., Two-Level Morphology: a General Computational Model for Word-Form Recognition and Production, Helsinki 1983.

(2) Koskenniemi, op.cit., pp. 42-88.

(3) Blåberg, O., Svensk böjningsmorfologi. En tvånivåbeskrivning, Helsinki 1984.

(4) Karttunen, L. and K. Wittenburg, A Two-Level Morphological Analysis of English, Texas Linguistic Forum 22 (1983), pp. 217-228.

(5) Borin, L., A Two-Level Description of Polish Inflectional Morphology, Center for Computational Linguistics, Uppsala University (forthcoming).

(6) Koskenniemi. op. cit., pp. 20-21, 42.

(7) Cf. Maegaard, B. and E. Spang-Hanssen, La segmentation automatique du francais écrit, Paris 1978, pp. 23-27.

(8) "9" stands for "unstressed verb ending -e".

(9) "=" stands for "any sign that is not mentioned in this rule, but is mentioned in some other rule".