

BDLEX Lexical Data and Knowledge Base of Spoken and Written French

G. Pérennou and M. de Calmès

SUMMARY

This document presents the BDLEX project (Base de Données LEXicales- Lexical Data Base) developed within the context of the GRECO-CNRS on Spoken Communication. The project is centered upon the phonological and morpho-syntactical levels of written and spoken French and is intended for use in applications involving the automatic processing of speech and texts.

INTRODUCTION

The language industries and more particularly, man-machine communication in natural language, require electronic lexicons of ever-increasing complexity, with a content and organization obligatorily adapted to the applications in view. The idea of constituting lexical data bases to provide a whole class of users with special lexicons sprung from this basic finding.

BDLEX (Base de Données LEXicales - Lexical Data Base) which is currently being developed within the context of the GRECO-CNRS of Spoken Communication, is centered on the phonological and morpho-syntactical levels of written and spoken French. Other research projects exist, aiming at the creation of formalized lexicons for French, although these have different objectives, the closest being that of M. Gross (1975), which was part of a lexicon-grammar. The first version of our base, BDLEX-0, was completed in 1986 (Pérennou, 1986) and the second BDLEX-1 is due for completion in 1987.

A number of applications are envisaged, such as : machines for recognizing dictated speech, speech synthesis, checking and automatic correction of spelling and typographical mistakes, computer assisted education (CAE).

This paper will describe the general structure of BDLEX, continuing with an examination of the phonological and morphological levels and the transfer of the various morphological relations to a data base in the form of a relational diagram.

STRUCTURE OF BDLEX

The BDLEX data base (abbrev. DB) has a relational type organization. BDLEX may be questioned through the French Transpac network or the normal telephone network, thereby facilitating team work.

The diagram in Fig. 1 gives the general structure of BDLEX. BDLEX-0 comprises :
- 7,000 lexical entries and 150,000 inflected words.

BDLEX-1 comprises 25,000 lexical entries generating 350,000 inflected forms and also a derivational morphological component.

Access may be obtained through the VORTEX system ; spelling and typographical mistakes are then tolerated.

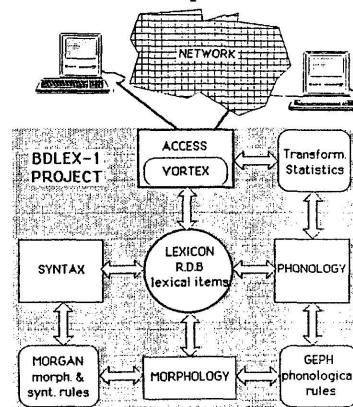


Fig.1 BDLEX components.

Laboratoire CERFIA - Université Paul Sabatier - 118, route de Narbonne - 31062 Toulouse CEDEX - FRANCE

PHONOLOGICAL COMPONENT

Each entry has an underlying associated phonological representation. The role of the GEPH expert system is to associate a phonetic representation with each entry, expressing its pronunciation in a given phrasal context, taking into account the dialect represented by a set of phonological rules.

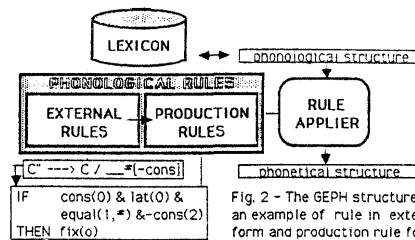


Fig. 2 - The GEPH structure and an example of rule in external form and production rule form.

Underlying phonological representation

What underlying form should be associated with the lexical entries? The answer to this question is often delicate and will, in any case, depend on the phonological processing planned. In BDLEX, the various fragments of phonological components supplied by a group of phonologists --- terminal consonants, nasality and semi-vocalization (Dell & Plénat 1985, Dell 1986), French vowels with a double timbre (Lambert & Rossi, 1986) - constitute the basis for the determination of the underlying forms.

Certain lexical entries are subject to morphophonological variations. For example, in French, the adjective "neuf" becomes "neuve" in the feminine form. The same may occur in English, the noun "knife" becomes "knives" in the plural. There are then two possible solutions :

- a polymorphic entry, each morpheme being subject to a condition for selection (ex. for the adjective "neuf" [new]: /**nœf**/ _ masc; /**nœv**/ _ fem).
- a phonological component with rules enabling the variations to be represented (ex. the underlying form of "neuf" is /**nœv**/ and the phonological component comprises the rule for unvoicing of final fricatives, except when there is a suffix or inflection beginning with a vowel or semi-vowel :

(DEVOI2) [-son +cont] ---> [-voix] / _ # or _ + [+cons]

In BDLEX, the second solution was adopted as this did not require the introduction of ad hoc phonological rules. Thus, the underlying form is unique when it may be used to derive the surface forms by means of the following rules :

- a) Latent consonant and mixed consonant - A latent consonant, noted as C", acts in the same way as a consonant C in a posterior non-consonantal context, it is truncated in all other cases. For example, the latent consonant /t"/ in /pətit"/ («petit») is used in the representation of the different masculine and feminine forms : /pəti/ («petit») ; /pətitə/ («petite»), derivations such as /pətitəsə/ («petitesse») and also in liaison.

A semi-fixed consonant C' acts in the same way as a fixed consonant C at the end of a group (i.e. in a _## context). In all other cases, it acts in the same way as a potential consonant C", providing just one underlying representation for words such as /sis"/ («six») or /ɔs"/ («os») (bone) etc...

The rules below govern latent and mixed consonants :

(FIX1)	C"	---	C / _____ + [- cons]	...	
(FIX2)	C'	---	C / _____ ##	(LIAI)	C" ---> C / _ # [- cons]
(DEFIX)	C'	---	C"	(TRONC)	C" ---> Λ (empty element)

These need to be completed by rules devoicing the occlusives and voicing the sibilants in liaison. Thus, «doux» and «grand» could have the respective underlying representations /**dus**"/ and /**grād**"/ required to form «douce» and «grande» whilst the liaison forms are /z/ and /t/.

- b) Regular denasalization of lexical final syllables - In French, when (TRUNC) takes effect on the latent nasal consonant behind the oral vowel, the latter is nasalized. Words such as "moyen" may therefore be represented by just one underlying form /**mwajɛ̃n**"/ in spite of the alternation [mwayɛ̃ / mwajɛ̃n] corresponding to «moyen/ moyenne» . The following rule (NAS), placed between (CL) and (TRONC) represents this nasalization.

(NAS) V ---> [+nas] / ___N", with :
 (i, ε) [+nas] ---> ě {y,œ} [+nas] ---> œ̃
 ɔ [+nas] ---> ǔ a [+nas] ---> ǎ

Not all final nasal syllables function as previously described.

The choice of the underlying forms in French raises a number of other difficulties which we are unable to examine here, such as : semi-vocalization, the alternation in timbre of medium vowels, i.e. [ɑ/a] and [ĕ/œ̃], the pronunciation of final consonants, semi-vocalization, borrowed words, etc.

The GEPH expert system

Provided with the phonological rules placed in its knowledge base, the system processes the expressions in their underlying form to determine the surface phonetic representations of the variants in pronunciation ---see Fig.2...

The rules may be entered in phonologists' formalized terms. They are then converted to an internal form -- for an example, see Fig.2 --- The contrary is also true ; a phonetic expression may be used to find the underlying form.

These need to be completed by rules enabling a more detailed description of pronunciation. At this level, the link should be made with the BDSONS data base of French sounds, also developed within the context of the GRECO-CRNS of spoken communication.

THE MORPHOSYNTACTIC COMPONENT

Inflectional morphology

The morphological alternations resulting from inflection do not justify polymorphic processing as a rule. The lexical entry «boxeur», ("boxeuse" in the feminine) is subject to a morphological alternation rule, noted here []_{øzə}, which should specify the lexicon. The single underlying representation /bɔks-/ will be valid both for the masculine form [bɔks + masc]_{øzə} ("boxeur" [bɔksœr]) and for the feminine form [bɔks + fem]_{øzə} («boxeuse» [bɔksøzə]) if we take :

[X + masc]_{øzə} ---> [X œr] [X + masc]_{øzə} ---> [X øzə]

Moreover, this entry is subject to the following rule concerning formation of the regular plural, noted []_{z"}

[X + sing]_{z"} ---> [X] [X + plur]_{z"} ---> [X z"]

In this way /bɔksœrz"/ («boxeurs») is formed from [[bɔks + masc]_{øzə} + plur]_{z"}

Generally speaking, this is the way in which nouns, adjectives and verbs are processed in BDLEX.

Morphosyntax

The morphosyntactic component is used for (non-deterministic) analysis and generation of inflected forms. Should these be defective in tense and/or person, this is taken into account. Processing is carried out graphically and/or phonetically. Examples :

. For the entry «boxeur», the system generates in phonetic mode :

masc. sing: bɔksœr / masc. plur : bɔksœrz"/
 fem. sing: bɔksøzə / fem. plur : bɔksøzəz"/

. The verb «faire» in the perfect tense, will be conjugated in graphic mode :

je suis fait or j'ai fait tu es fait or tu as fait

. An analysis of the word «cours» in graphic mode, gives three groups of solutions:

- 1) «cour + s» fem. noun in plur.
- 2) «cours» masc. noun in sing. or plur.
- 3) «cour + s» verb in pres. ind., 1st, 2nd. pers. or imper. 2nd pers. or pres. subj. 2nd pers.

. If [moz] is submitted to phonetic analysis, the system will put forward two solutions which are : «mot+s» (masc. noun, plur) and «m+aux» (masc. noun, plur.).

When the data base is requested to conjugate an item or to generate inflected forms of nouns and adjectives, the morphophonological rules are required. Alternations which depend neither on the context in the sentence, nor the speaker, are processed at this level.

Derivational morphology

Derivations may be interpreted as relations. Thus, "égal", "inégal" and "inégalement" (and similarly in English, "equal", "unequal" and "unequally") are described by tuples in PREFIX relations (pref, entry-1, entry-2) and SUFFIX relations (entry-1, entry-2, suf), i.e.

(IN₁₁,égal,inégal) (inégal,MENT₁₁,inégalement)

In these tuples, the fact that the affixes may be homophones, giving rise to morphophonological variations, should be taken into account. Thus, «in-» covers two prefixes in French. IN₁₁ is the negative adjectival prefix in the variant «in» (it may also be «il-» in front of an «l», «ir-» etc.) MENT₁₁ is the suffix forming adverbs from adjectives in the variant -ement. It may be felt that "entry-2" is redundant in this type of formulation and this is backed up by the example.

Other more complex cases show that the type of simplification envisaged is generally impossible unless a processing device for irregularities or allomorphy is provided, for example ____«éli» is the root of the verb «élire» separated from the «-re» infinitive termination. In the same way, «croi» is the root of the verb «croire» (éli,IBLE,éligible) (croi,IBLE,crédible).

In BDLEX, the initial stage of morphological representation selected has a descriptive aim. Each lexical entry has boundary markings specifying the nature of the term immediately following. When part of the word is an entry, it is not broken down. Any letters which need to be added or removed to obtain the identifier of the entry concerned, are given in brackets. This is illustrated in the examples below where

:X refers back to the autonomous entry X

.s designates s as a suffix

+p designates p as a prefix

;d designates d as a desinence

paie :pay(>i)e;r :pay(+er).able +im:payable +bi=game

CONCLUSION

Future versions of BDLEX should take morphology-semantics relations into account. Numerous linguistic projects are currently trying to solve the delicate problems which cause the distortions observed. We ourselves are examining object modelling of lexical entries with this in mind.

REFERENCES

Dell, F., Les règles et les sons, Paris:Hermann (1986).

Dell, F.1 Plénat, M., Semi-voyelles et consonnes finales en français, rapport interne du GRECO communication parlée du CNRS (1985).

Goss, M., Méthodes en syntaxe, Paris:Hermann (1975).

Rossi, M. 1 Lambert, M., Représentation et traitement des voyelles à timbre multiple, Actes du séminaires GRECO-GALF (1986) pp. 141-62.

G. Pérennou, "BDLEX: A Data and Cognition Base of Spoken French". Proceedings of ICASSP, Tokyo (1986) pp. 325-8.