MORPHO-PHONETIC RELATIONSHIPS AND ELABORATION OF A LEXICON OF SPOKEN FRENCH

D. Dujardin (1), R. Belrhali (2), L.-J. Boë (2) and J. Courtin (1)
(1) Laboratoire de Génie Informatique, Grenoble, France
(2) Institut de la Communication Parlée, INPG-Université Stendhal, Grenoble, France

ABSTRACT

We present morpho-phonetic relationships derived from the lemmatization of the phonetic French spoken lexicon De À à Zut [1] processed by PILAF system [2]. Phonetic variants have been classified in relation to categories and locutions. The analysis has led to a redefinition of the classification of some connectors.

INTRODUCTION

The BDPHO database [3] was produced by the École Nationale Supérieure des Télécommunications (Département Signal) and the Institut de la Communication Parlée de Grenoble. It is based on a corpus of recorded speech (about 110 hours), transcribed by expert phoneticians. There were about 30 speakers, over 300,000 sounds and 102,000 words. BDPHO was constituted by restitution of 7,590 orthographic forms (corresponding to 7,221 phonetic forms, including 1,386 variants). BDPHO was developed on a Macintosh computer, in the HyperCard environment; it contains three parts of one corpus in their orthographic and phonetic forms and two orthographic-phonetic lexical bases.

The De À à Zut dictionary [1], was produced on the basis of this database. This dictionary of forms include the number of occurrences found in the corpus, the different pronunciations and the structure of the phonetic words presented as a cohort string (CV, CVCV...).

THE DIFFERENT ELEMENTS

The Corpus

Origin and description

Three parts constitute the BDPHO Phonetic Database and include exactly 304,752 sounds.

* Corpus n° 1 (86,360 sounds) made up of recordings from radio programmes was made in the Institut de Phonétique de Grenoble under the supervision of R. GSELL.
* Corpus n° 2, with its 201,281 sounds, is the most important one; it was put together by A. MALECOT for a project at the University of California, Santa Barbara. It contains fifty half-hour free conversations on various subjects with members of the Paris "intelligentsia" (professors, lawyers, doctors, artists...).
* Corpus n° 3 (17,111 sounds) was collected by J. VAN EIBERGEN [4] at Institut de la Communication Parlée. It includes the transcriptions of 16 short conversations by 16 speakers of various linguistic and socio-professional origins (4 teenagers, 8 adults aged twenty to sixty and 4 sixty to eighty). These are simple conversations in unformal situations. The language used is spontaneous.

Codification

Because of their different origins, these corpora have not been encoded in the same way, we have defined a representation of a subset of IPA used for the French language and a few additional characters.

Contents

The whole corpus contains 102,137 lexical occurrences, and 7,221 different phonetic shapes. The possible number of combinations of "polysounds" is far from being used completely; if "bisounds" are very numerous (87% of the possible ones are represented, "trisounds" and above all, "quadrillsounds" are few (respectively 30% et 3.8%). Some words and sequences of words are very frequently found (alors, parce que, il y a, de la, avec, et par, crois, voir...). Liaisons which bring about the d, n, p, r, t, z sounds represent about 6% of the occurrences of words in the corpus and are very unequally distributed, three of them [n, t, z] totaling over 90%. The comparison of the first 50 occurrences in the corpus with those published by G. ENGWALL [5] and those of the Listes Orthographiques de Base (LOB) by N. CATACH [6] shows both strong similarities (over 30 commun shapes or entries), and the emergence of some words which are specific to spoken language (ça, y, alors, très, oui, enfin, parce que, oui, ma, quand, puis, euh) which bring about the [d, n, p, R, l].

There are important differences in favor of spoken language (pour, est, c', pas, on, ça, ce, il, bien, alors, très, oui, enfin, parce que, fait, si, même, là, euh) and very rarely the opposite (the negation ne is often omitted). Concerning the cohorts, 25% of them cover 90% of the possibilities and 44 cohorts represent 95% of the total.

In order to manage the database [7] we have decided to adopt the HyperCard environment and the HyperTalk programming language for its ease of use and adaptation by non-specialists, but also for reasons of distribution (no licence needed, because there are also stand alone versions of HyperCard).

The PILAF system

The PILAF system (Procédures Interactives Appliquées au Français) is a user-friendly system for parsing the French language produced by the TRILAN team (Traitement Informatique de la Langue Naturelle) [2]. It is a part of a linguistic toolbox implemented on microcomputers. Its adaptability implies that it is not only general, parametrable and portable but also that it should be easy to integrate to different systems. For the lexical level, the PILAF system proposes modules for morphological parsing, generation of flexional forms from a root and a lemmatiser. It is based on a database composed of two dictionaries and linguistic data including a validation-saturation grammar defined by a set of rules, as well as lists of models, of lexical categories and of variables [8]. All of these data are manipulated by means of specialized editors.

RESULTS

VARIATIONS OF FREQUENCIES

In order to recognize grammatically each phonetic variant and to associate a lemma it has been necessary to take into account compound words and fixed locutions. This work has entailed modifications of lexical entries. Thus, if almost 500 compounds are created, over 1,000 are modified. Often the majority of locutions include tool-words.

Here are some of the findings:

* complete removal of about a hundred phonetic variants. In the following examples the number of occurrences is indicated by numbers between <>

< [aba] abord < 42 > d'abord
[aba] abord < 1 > d'abord
< [ako] accord < 31 > d'accord
< [traver] travers < 1 > à travers
< [fil] fil < 1 > coup de fil
< [fil] fill* < 2 > fils de fer
* the form "fils" has a heterophone homograph [fis] fils (son).
* selective removal which seems obvious when there are liaisons
< [tu-f] tout à fait (quite) < 56 >
< [tu-f] tout à l'heure (later) < 19 >
< [tu-f] tout de même < 50 >
< [tu-f] tout de suite (immediately) < 20 >
* but is less obvious in the case of n'est-ce pas (isn't it) where, among the 67 occurrences only the phonetic variant of est is a constituent of this locution, whoever the speaker may be.

* modification of the frequency of about 200 items:
< [ajrre] ailleurs (elsewhere) < 115 >
< [ajrre] ailleurs (by the way) < 100 >
< [ajrre] ailleurs (otherwise) < 5 >

The study of the quantitative variations of components of locutions pinpointed the emergence of "kernel" occurrences which appear in a number of different locutions (number between 1). The study of the quantitative variations of components of locutions pinpointed the emergence of "kernel" occurrences which appear in a number of different locutions (number between 1).

même (same) < 17 >, peu (few) < 12 >, moment (moment) < 8 >, temps (time) < 8 >, fois (times) < 5 >, mesure (measure) < 5 >. It is also possible to study the incidence on the tool-words which often have high frequencies.
The occurrence of the [u] variant as a verb is ambiguous in the noun, it can only be 

"fait" (fairy).

- Lemmatization also leads to deletion of a large number of morphological homographs. The majority of compound words and locutions is not ambiguous, some homographs of components are removed by their absorption into a locution.

"tout" "compte" "fait" adj subc ppas

ppn verb adjq

"tout compte fait" loca

Creation of new lexical categories

In the first version the emergence of words characterizing spoken language appeared (alors, ça, y, il y a, très, oui, enfin, parce que, moi, quand, puis, euh). It has led to the adaptation of morphological classes to needs of spoken language.

* creation of the class of:

- presentatives c'est, c'était, il y a..., pauses, euh, hein ...
- speech support alors, ben, quoi ...

- pragmatic connectors ça va, ça va pas, ça y est, c'est ça, c'est bon, c'est fini...

- Nevertheless, it would seem that in these informal conversations, the class of pragmatic connectors is not adequate.

On the other hand, the phonetic variant [aks] accord < 31 > which is completely absorbed in the location d'accord occurs in 15 adverbial locations and 15 speech supports.

The occurrence ça < 1069 > is considered as characteristic of spoken language and has only one phonetic variant [sa]. But it is a component which belongs to most of the newly created classes such as ça serait, presentative, ça y est, pragmatic connector, tout ça pause, comme ça support of speech.

Among the 90 occurrences comme ça also 3 cannot be considered as entities comme ça vient.

CONCLUSION

The flexibility of the system has led to a redefinition of the classification of some connectors, to a better knowledge of qualitative and quantitative phonetic variants. The presence of contexts has led us to the reconsideration of sequences of connectors, and to further study of speech markers and markers of conversation structure, and to refine our analysis of speech behaviour.

ACKNOWLEDGEMENT

The lemmatization of this dictionary is a part of the GDR: Programme de recherche concertée Communication Homme-Machine, Pôles Parole et Langue Naturelle and was a component of the ORALL project (Organisation et Accès à de Larges Lexiques en vue du traitement de la Parole), now the Groupe de Travail (GT5) Lexique du PRC Informatique.

REFERENCES