MORPHO-PHONETIC RELATIONSHIPS AND ELABORATION OF *De À à Zut* 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 10 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. MALÉCOT 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 been not 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, "quadrisounds" 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] totalizing 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* (LOBO) 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, moi, quand, puis, euh*). There are important differences in favor of spoken language (*pour, est, c', pas, on, ça, ce, y, bien, alors, très, oui, enfin, parce que, fait, si, même, là, euh*) and very rarely the opposite (the negation *ne* 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 <>.

[abɔʀ] abord < 42 > d'abord
(approach) (first)
[abɔʀɑ] abord < 1 > tout d'abord
(approach) (at first)
[akɔʀ] accord < 31 > d'accord
(agreement) (all right)
[tʀavɛʀ] travers < 13 > à travers
(failing) (through)
[fil] fil < 2 > coup de fil
(thread) (phone call)
[fil] fils* < 2 > fils de fer
(threads) (wires)

* the form "fils" has a heterophone homograph [fis] fils (son).

• selective removal

which seems obvious when there are liaisons

[tu-t] tout à fait (quite) < 56 >
tout à l'heure (later) <19 >
[tu] tout de même < 50 >
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 [ɛ] of the form *est* (is) is a constituent of this locution, whoever the speaker may be.

• modification of the frequency of about 200 items.

[ajœʀ] ailleurs (elsewhere) < 115 >
d'ailleurs (by the way) < 100 >
par 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 { }).

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.

Occ	Var Phon	nb Occ	nb Loc	nb OcLoc	nb Mc	nbOcMc	%
à	a	1833	62	436	9	19	24,82
c'	s	1927	4	128			6,64
œ	s	197	17	171			86,8
	se	745	21	315			42,28
d'	d	1129	38	303	12	19	28,5
de	d	124	15	53	2	4	45,96
	de	3151	87	537	19	84	19,70
des	de-z	295	7	8	2	4	2,71
	de	985	14	32			3,24
du	dy	538	26	136			26,2
en	ɑ̃	1024	48	364	3	3	35,83
	ɑ̃-n	323	3	26			8,04
est	t	3	1	2			66,6
	ɛ-t	219	2	21			9,5
	ɛ	765	15	176			23
	e-t	550	3	103			18,72
	e	1396	3	21			1,5
l'	l	1444	9	104			7,2
la	la	1666	21	97	1	1	5,8
là	la	409	11	115			28,11
le	l	113	3	6	2	4	5,3
	lœ	1453	19	68			4,67
les	le-z	345	3	33			9,5
	le	962	3	8			0,83
pas	pa-z	202	1	3			1,48
	pa	1313	4	141	1	1	10,81
qu'	kœ	19	2	9			47,36
	k	891	32	334			37,48
que	k	63	7	10			15,87
	kœ	1797	45	730			40,62

This table presents the percentages of occurrences of units included in the composition of locutions.
Occ = Occurrence, Var Phon = Phonetic variant nb Occ = number of occurrences nb OcLoc = number of occurrences of locutions
nb Mc = number of compound nouns nb OcMc = number of occurrences of compound nouns

Morpho-phonetic relationships

We have noted:

• the enhancement of a relation between the phonetic variants and the lexical category of homographs of the orthographic string. For the form *fait* (event or made) < 269 > the phonetic variant [fe] which appears 265 times in the locution *tout à fait* (quite) < 34 > et *tout compte fait* (all things considered) < 1 > and occurs 105 times as a verb *faire* (to make) in the third person singular of the indicative present and 125 times as a past participle, whereas [fɛt] which appears in *au fait* (by the way) < 1 >, *de fait* (de facto) < 1 >, *en fait* (in fact) < 33 > and occurs 15 times as a commun noun as well whoever the speaker may be.

The phonetic variant [tus] of the form *tous* (all) corresponds to the pronoun in its 25 occurrences whereas only 3 pronouns are encountered among the 122 occurrences of the [tu] variant, the others correspond to the undefined adjective. For each form the ambiguities are resolved by a syntactic parser. These results have been recalculated for each phonetic variant which is more interesting since this is a study of spoken language.

If we consider the phonetic variant [fe] which corresponds to the occurrences *fais* < 39 >, *fait* < 265 >, *faits* < 6 >, *fée* < 6 >, *fées* < 5 > there can be no ambiguity about the noun, it can only be *fée* (fairy).

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

"tout" "compte" "fait"
adji subc ppas
pnp 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 [akɔʀ] accord < 31 > which is completely absorbed in the locution *d'accord* occurs in 15 adverbial locutions and 15 speech supports.

The occurrence *ça* < 1069 > considered as characteristic of spoken language 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

[1] Boë L.-J., Tubach J.-P. (1992), *« De À à Zut » Dictionnaire phonétique du français parlé,* Ellug, Grenoble.

[2] Courtin J., Dujardin D., Kowarski I. (1992), *PILAF: Software Tools for Lexicography and Text Research.* COMPLEX'92: 2nd Conference on Computational Lexicography and Text Research, Budapest, Hungary, pp. 93-109.

[3] Boë L.-J., Tubach J.-P. (1992), *BDPHO: une base de données lexicale orthographique-phonétique du français parlé,* Séminaire Lexique du GRECO-PRC Communication Homme-Machine, Toulouse, pp. 111-119.

[4] Van Eibergen J. (1985), *Corpus d'un français vernaculaire à caractère spontané et impératif,* Bulletin de l'Institut de Phonétique de Grenoble, vol.15, pp. 35-74.

[5] Engwall G. (1984), *Vocabulaire du roman français (1962-1968). Dictionnaire des fréquences,* Almqvist & Wiksel International, Stockholm, Suède.

[6] Catach N. (1984), *Les listes orthographiques de base (LOB),* Nathan, Paris.

[7] Dujardin D., Belrhali R., Boë L.-J., Courtin J. (1994), *Lemmatisation du dictionnaire du français parlé "DE À à Zut".,* 20e JEP, GFCP-SFA, Lannion, pp. 297-302.

[8] Courtin J., Dujardin D. (1991) *Paramètres linguistiques du français dans le système PILAF.* Rapport Technique RT 67, Laboratoire de Génie Informatique, Grenoble.