6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TOWARDS THE GENERATION OF FRENCH PHONETIC INFLECTED FORMS

Frédérique Sannier, Véronique Aubergé
Institut de la Communication Parlée
Université Stendhal/INPG, Domaine Universitaire
38040 Grenoble Cedex 9
(sannier,auberge)@icp.ing.fr

ABSTRACT

After having given the reasons why it is legitimate to dedicate a special attention to the oral code with regard to the written code, we propose an approach to the automatic generation of French phonetic inflected forms. We shall examine it first inside the general context of automatic phoneticization, with a special attention to the phoneticization of loanwords. We shall then give the methodology we adopted to build up our generation system.

Keywords: orthographic code, oral code, French oral inflection, loanwords, automatic generation.

10.21437/Eurospeech.1999-528

1. INTRODUCTION

The process of text generation trivially consists of two components, an expert reasoning system answering the question *"what to say?"* and a linguistic generation module answering the question *"How (to say)?"*. Then comes the stylistic stage, *"How(how(what to say))"*, which will always be decoded by the human speaker, wether it was explicitly controlled or neglected in the generator. If we compare speech and text generation, the first question that comes out deals with the nature of speech and orthography: are they two modalities or two mediums?

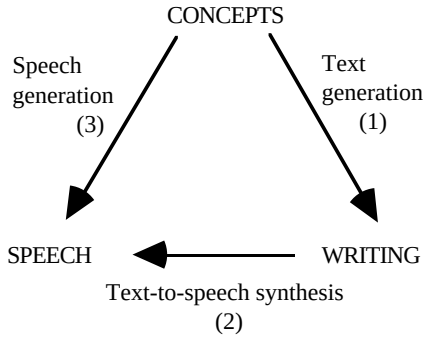

Figure 1 : speech/orthography : two modalities?

Despite the fact that French is inherently an alphabetic language (in Latin the oral code and the orthographic code had a quasi one-to-one relationship), a partial autonony between speech and orthography was demonstrated [4, 5, 6, 8, 2]. Equally, from the syntactic point of view, the structure of speech and the structure of texts clearly show specificities [3]. In other words, if speech could be obtained by means of a simple translation of orthography (code transcription and style reconstruction), speech generation ((3), figure 1) would be equivalent to the reading of generated texts, that is to say to the interlocking of stages (1) and (2) (in which the generated text together with its structures would be transferred to the synthesizer).
If the point is speech generation and its specificities, it is necessary to reduce the representation of orthography to the underlying parts common to the speech underlying representation, and certainly not to make it look like speech. Consequently, choosing to go directly from concepts to speech is not a processing economy measure but is justified by the fact that the linguistic decisions are not the same as in concept-to-text-to speech production.
This partial autonomy between speech and orthography gives all its interest to the field of phoneticization on which we worked and which constitutes the first part of our study (with a special attention to loanwords), and to the oral inflectional morphology, to which we then dedicated ourselves, by the study of the automatic generation of phonetic inflected forms.

2. GENERAL METHODOLOGY

The first part of our work consisted in observing the behaviour of the phonetic transcription of the French inflected forms. This stage took place in the succession of a long-term job, the methodology of which having been settled to give a general modelization of French phonetic transcription. This methodology is an experimental one using bootstrapping: a nucleus primitive grammar written in the phoneticization Toph language was first elaborated on a theoretical basis [1], and allowed us to filter a comprehensive (Le Robert 1) orthographical-phonetical lexicon of French canonical forms. From this filtering emerged a grammar extended to the whole of these canonical forms. It is on the basis of this canonical forms grammar that we were able to bring to the fore , by filtering an

inflected forms database, which areas are phonetically specific to inflection.

In parallel to this phoneticization study, an automatic generation system of orthographical inflected forms of an adjectival and nominal reference lexicon was proposed. This system uses the lexical DELA categories [7] and is implemented on a Hypercard environment.

Finally, we developped, in the same way, a tool generating the phonetic inflected forms from phonetic basis.

3. THE PHONETICIZATION OF LOANWORDS

Areas specific to the treatment of loanwords which cannot still be considered as being assimilated by the target language (French), emerged from the canonical forms phoneticization grammar. Equally, in the phoneticization grammar extended to the inflected forms, various behaviours specific to the inflection of those [9] were observed. These behaviours may be classified into five configurations.

3.1 Minimal assimilation

The target language collects the entire inflectional paradigm of the loanword, at the graphical level, and the phonetic transfer is as faithful to the source language as possible. The series *yacht(s)man, yacht(s)men, yacht(s)woman, yacht(s)women,* the first apparition of which dates back to 1859, borrowed from English, was transferred in full to French. The pronunciation of the canonical form, as well as the pronunciation of the inflected forms tries to reproduce the English one.

3.2 Maximal assimilation

This case is orthogonal to the first. French borrows the written form of the canonical form but pronounces and inflects it the French way. Example: *boy scout [bOiskaut]* gave *scout [skut]*, and *girl scout, [g«l:skaut]* *scoute [skut]*.

3.3 Redundancy

French only takes the inflected form of the lexical item and recategorizes it into a canonical form. It thus looses the trace of the inflection and superimposes on what has become an uninflected form its inflectional morpheme. We have the example of *blinis* (here in the roman alphabet), borrowed in 1883, which comes from the Russian word *blini* ( ), plural of *blin* ( ).

3.4 The false imitation

French adopts the canonical form and fixes on it an inflectional paradigm which is thought to follow the rules of the source language. The pronunciation as well is transferred. Example: *tennisman [tenisman], tenniswoman [teniswuman], tennismen [tenismEn], tenniswomen [teniswumEn].*

3.5 The mixed functioning

The functioning is mixed: the inflected form is the one of the source language but there is also a "French" one. For example, *réal*, borrowed from Spanish gives *réaux* and *réales*, respectively French and Spanish plurals.

The five major cases of functioning we described presuppose the independance of the morphemes, namely the base and the inflection. Against all expectations, a lexical item is thus not automatically borrowed together with its inflectional paradigm.

4. THE AUTOMATIC GENERATION OF PHONETIC INFLECTED FORMS

Let us dedicate ourselves to the automatic generation of phonetic inflected forms. We will expose in a few lines the methodology we used to elaborate the generation module.

From a formal and functional point of view, each occurrence of the orthographic-phonetic lexicon of 60,000 entry words, on which are based the orthographical forms generetors previously elaborated, shows a category entry word corresponding to the DELA categories [7]. These categories, or labels, organise the entry words according to various criteria (part of speech, nature of the inflectional paradigm, gender and number). For the generation of phonetic forms from phonetic bases, it was necessary to elaborate a similar classification, and in our concern to keep an internal coherence the same parameters were kept. However, we had to define two series of labels, for the reasons that will be exposed further on.

4.1 The simple labels

The simple label associated to each lexical item may be composed of at most 4 ranks. The first rank gives the nature of the canonical form (N[oun] or A[djective]) and it is the only compulsary rank. In the case of a masculine lexical item which has a feminine formation, the second rank gives the parameters associated to the masculine plural (MP), the third one the parameters associated to the feminine singular (FS), the fourth one the parameters associated to the feminine plural (FP). In case of a feminine lexical item, the second rank will give the

characteristics of the feminine plural. Here is the example of a label:

Label	Lexical items	Definition
N.MP0.FS15.FP0	*JaklE$, O$dE$, aRl«kE$, bERnaRdE$, baladE$, bedwE$, benediktE$, blOdE, dOgE$, dofE$, gR«dE$, jukskybE$, kRapusE$, kapysE$, kulpE$, lapE$, maRgulE$, papalE$*	Nouns, the feminine of which is obtained by replacing [E$] by [in] into the canonical form. The masculine plural is identical to the canonical form, the feminine plural is identical to the feminine singular.

Table 1: example of a simple label

Let us make it clear that the first rank label may be complex, when the lexical item has got only one gender and/or only one number (in this order). The lexical item *[dekOROm]* will thus have the label NMS (Singular Masculine Noun) since it is exclusively a masculine singular.

There are 33 labels concerning the formation of the masculine plurals (21 describing the functioning of loanwords and 14 double plurals), 37 deal with the formation of the feminine singulars (1 describing the functioning of loanwords and 2 double plurals) and 6 labels for the feminine plurals (4 describing the functioning of loanwords and 2 double plurals).

We have explained the functioning of the simple labels. We shall now expound the functioning of the complex labels, the elaboration of which follows another methodology.

4.2 The complex labels

The necessity to make the distinction between simple and complex labels is the consequence of the methodology we used. This methodology having consisted in starting from a lexicon of orthographical units submitted to a phoneticization process, quite a lot of homophone units were generated. We were then faced to an alternative. The first solution consisted in generating all the occurrences. This solution had the advantage to allow an unquestionable transparency. But we chose another solution consisting in gathering all the homophones in only one occurrence, for two main reasons. The first of these reasons was motivated by our concern to stay coherent with the DELA database: DELA only shows one occurrence for each homograph, wether it refers to one or several entries in the common French dictionary Le Petit Robert, our phonetic reference. The same semantic entry which is at the same time heterograph and homophonous *(cawcher, cascher, kascher,* all pronounced *[kaSER]*) must be represented following its phonetic polymorphism (reducible to 1) and not following its spelling polymorphism, since orthography must intervene as little as possible in our representation. The second reason why we chose this option is also a matter of staying coherent with the general processes of our work which postulate, as already said in the introduction, the partial autonomy of the oral code and the written code. Giving only one occurrence consolidates this autonomy by making the reference to orthography superfluous. The complex labels are of the form AMBX.(Y), where AMB is a given string, fixed, X is the ambiguity category and Y the, facultative, substring of the ambiguity. There are 18 X categories, which take into account only the nature of the entry lexical category. The maximum number of entries for one given homophon is 6 (*[Sa]*, corresponding to *chat* (Adjective), *chah* (noun), *chas* (noun), *chat* (noun), *schah* (noun) and *shah* (noun)).

There are 237 labels for the thus called "ambiguous" lexical items, on a total number of 320. It represents 74% of all the labels. The labels of the ambiguous items are consequently the majority, but this figure does not reflect the reality expressed in absolute figures since the total number of the ambiguous lexical items rises to 5956 occurrences only for a total of 52750 lexical items. This is not surprising since each ambiguous lexical item "hides" from 2 to 8 occurrences, and the conditions for each label to share the same criteria are all the more difficult to fulfill since they are numerous.

5. CONCLUSION

The aim of our work is to propose a module generating automatically inflected forms from the reference phonetic forms. Our concern was to restrict to the minimum the morphological hypothesis based firstly on structures, in order to organize a generation based on the minimum of orthographical a priori. The system we propose (was implemented in Hypercard, on MacIntosh), then, gathers 320 labels, describing 52750 canonical lexical items. The number of labels of the system specific to the phonetical generation is thus diminished of approximately 25% compared to the system developped for the orthographical inflection system [10].

6. REFERENCES

[1] Aubergé, V., (1991), La Synthèse de la Parole : Des Règles aux Lexiques. PhD thesis, Université Stendhal, Grenoble, France.

[2] Aubergé, V., Belrhali, R., (1996), La Phonétisation Automatique du Français : émergence de règles ou de lexiques ? Revue LIDIL, n°13, Orthographe et Prononciation.

[3] Blanche-Benvéniste, C., (1990), Le français parlé. Études grammaticales, Paris, CNRS Editions.

[4] Catach, N., (1984), La phonétisation automatique du français, les ambiguïtés de la langue écrite. CNRS Editions, Centre Régional de Publication de Paris.

[5] Catach, N., (1989), Les délires de l'orthographe. Plon Éditions, Paris.

[6] Gnheim, N., (1997), Relations entre les codes de l'oral et de l'écrit. PhD thesis, Université Stendhal, Grenoble III.

[7] Laporte E., (1988), Méthodes algorithmiques et lexicales de phonétisation de textes, Application au français. PhD Thesis, Université ParisVII.

[8] Lucci, V., Millet, A., (1994), L'orthographe de tous les jours, enquêtes sur les pratiques orthographiques des français. Honoré Champoin Editions, Paris.

[9] Sannier, F., (1998a), How a French Text-to-speech System can Describe Loanwords. *Proceedings of The International Conference on Speech and Language Processing*, Sydney, Australia, volume 5, p. 2023.

[10] Viala, F., (1995), Première étape vers la generation phonétique des formes fléchies du français. DEA des Sciences du langage, univeristé Stendhal, Grenoble III.