HOW A FRENCH TTS SYSTEM CAN DESCRIBE LOANWORDS

Sannier Frédérique, Belrhali Rabia, Aubergé Véronique

Institut de la Communication Parlée, ESA CNRS 5009
Université Stendhal/INPG, domaine universitaire
38040 Grenoble Cedex
Tél.: 04 76 82 41 17 - Fax: 04 76 82 43 35
e-mail : (sannier,belrhali,auberge)@icp.inpg.fr

ABSTRACT

Our aim is here to give a survey of the phonographical behaviour of the loanwords introduced into the French lexicon, through the observation of the systematic functioning of the French letter-to-phone TOPH system. We thus define sub-systems, isolated into lexicons. The observation of the utterances graphemic functioning, made it possible to delimit classes. The second part of this study more specifically deals with the loanwords inflexion paradigms, for which as well different functionings are drawn.

1. INTRODUCTION

Our general framework is the description of the relations between the graphic and the phonetic codes (1), (2), (5). We adopted an inductive methodology based on the observation of an extensive attested sample: the graphic and phonetic forms of (9). The size of the corpus (59,000 lexical items in (9)) allows us to generalize to the French language in its whole the functionings that emerged from (9). The analysis reveals, for a set of graphemes, contextual constraints delimiting "lexical areas", that is to say sub-groups of lexical items for which the transcription follows a specific system of regularity. The transcription process, inside these linguistic areas, results from the projection on the French language of the source language system. In this paper we more specifically focus on the French pronunciation of loans. We present in the first part sub-systems belonging to a particular source language, represented by lexicons extracted from (9). We will then give a first description of the transcription of the loans inflected forms. The whole of these descriptions is implemented into a TTS written in Toph language (1) and has been systematicaly tested on a sample of the French language equivalent to (9).

2. LOANWORDS: SPECIFIC SYSTEMS

2.1. What is a loanword?

Borrowing a word from another language is a phenomenon experienced by every language in contact with others, with the help of a privileged media: its users. It has to be distinguished from interference or alternance between languages (8). Let us propose a first definition, inevitably reducing: borrowing consists in the translation in a given linguistic system of characteristics (in our case phonetic ones) peculiar to another linguistic system. There are many motivations: (a) the referent is borrowed at the same time as its denomination (we will develop the example of *tomate* ("tomato") further on), (b) the word does not exist in the extralinguistic reality of the borrowing culture (*djemââ*, 1870, arabic, notables assembly representing a douar, in Northen Africa), (c) at another level, a political one, if we may say so, the word is borrowed on the pression of a dominant linguistic community (this particularly concerns the computing domain).

If a word recently borrowed is easily recognizable because still stamped with exoticism in its pronunciation (and "exoticism" is taken into its broader acceptation, dealing with everything that comes from the exterior, the interior being a particular linguistic zone), others loose the (morphological, phonetical...) trace of their exogene origin. Let us go back to the example of *tomate*. This lexical item comes from Nahuatl, one of the ideographic Aztec languages. It thus underwent an alphabetization, first via Spanish (passing from *xitomatl* to *jitomatl*), then a modification in the pronunciation, slipping from the final [tl] to [t a], then [te], which became [t] in French.

The lexical item morphology can lead to aberrations: if *choucroute* (a typical dish from the East of France) definitely results from the composition of *chou* ("cabbage") and *croûte* ("crust"), it was nonetheless generated by the phonetical attraction of the german item *Sauerkraut*. The assimilation is here completely fulfilled: the written and phonetic morphologies entirely belong to the French linguistic system. However, the lexical items of the French general lexicon showing the written succession ‹tch› spelt [tS] at the initial, doing so, keeps the trace of their foreign origin, since no non-borrowed lexical item comprises it *(tchin tchin* ("cheers"): 1902, Pidgin-English of Canton; *tchervonetz* : XX°, Russian).

2.2. The Toph Processing of Loanwords

The TTS written in Toph language allows a specific processing of loanwords (provided these lasts show deviations to the norm), by two means: (a) the creation of individual rules, thus proving the grapheme-phoneme correspondance singularity (‹eo› => [i] exclusively in *boat people*, 1979, English); (b) the creation of lexicons, resulting from a systematic filtering of (9) through the general French system. The methodology is as follows: the minimal transcription units showing a deviation to the primitive grammar used as a basis are gathered in lexicons, which are the trace of a unified specific functioning. As a

result, we delimit the lexical items coming from the same source language and at the same time show the same phonetic transcription for each given unit.

The foreign languages represented in the grammar are presented below. They are given in increasing order of representation in the grammar, for the same grapheme or not. English (105), German (45), Latin (29), Italian (together with Corsican) (28), Arabic (22), Spanish (22), Greek (18), Dutch (13), Russian (10), Portugese (10), Japanese (7), Hungarian (7), Hebrew (6), Chinese (5), Turkish (4), Scandinavian (5), Egyptian (3), Hindi (3), Sanskrit (3), African languages (2), Lapp (2), Persian (2), Swedish (2), Vietnamese (1), Polish (1), Ameridian languages (1), Malay (1), Singhalese (1). A particular attention is required for the interpretation of these data: each occurence refers to a given realisation of a given grapheme, that is to say we do not give any indication about its covering. A single rule can apply to a highly varied number of lexical items. Example: the rule sounding the consonant ‹t› in the final position is activated on items belonging to many different languages (*smart [smaRt]*, English, 1898; *pschent [pskEnt]*, Egyptian, 1830). Conversely, a single lexical item can generate several different rules (*yohimbehe [jOimbe]*, bantu, 1908). Generaly speaking, the weight of loanwords on Toph varies according to the assimilation degree: the more assimilated the lexical item, the less deviant to the pronunciation rules of French. Which makes it less ″detectable″. It also means that the list of languages we give is far from exhaustive: the lexical items belonging to regular languages (regular to the French system) or the lexical items that are totally assimilated do not appear.

2.3. A Typology of Phonetic Transfers

Following Holden's work (7), in which he distinguished three functionings and insinuated that there could exist others, we can give a description, more precise since we take into account the graphic dimension.

Some lexical items prove the existence of what we will call sub-systems *(tabac* ("tobacco"), from the Arouat (an Haitian language) word, *tsibatl*, via the Spanish *tobaco*) inside a broader correspondance system ((*c* at the finale is not sounded (*clerc*, ¨clerck¨)). These transcription sub-systems, particular to borrowings follow four configurations:

First configuration. The unit also belongs to the French system. It is pronounced in the target language after an immediate transfer *(yacht [jak])*. The units, then, are the same as the French general system, which logically implies that these loanwords do not generate a specific processing in Toph TTS.

The three following cases, on the contrary, have all major repercussions on our grammar. They are concerned with cases for which the pronunciation of the unit shows deviances from the uncontextual prononciation.

Second configuration. The unit also belongs to the French system but the sound target are modified *(pace maker [pEsmEkœR])*. Let us precise that English has got a particular status (3).

Third configuration. The sound target of the unit borrowed is a contextual generalization of the general system (*bled [blEd]*, (¨the middle of nowhere¨) military slang from the Algerian Arabic, late XIX°) or were totally intergrated *(wagon [vagõ]*, 1826, *warranter [vaRA$tE]*, 1874).

Fourth configuration. The inherited unit corresponds to a sound unit resulting from a transfer without any analogy with the French system. There are three of these units in French: [x], [ø] and [h]. These consonants only show in loanwords, the first one in lexical items coming from Spanish (but Spanish inherited them from Arab) or Arab. The second one can be found in items coming from English and Spanish and the third one from Russian.

We present below exhaustively the graphemes for one particular foreign language that is processed in the grammar, Spanish.

Conversion	Exemples datés
o(y) --> [O]	arroyo 1855
ch -->[tS]	chorizo XIX°
a(y)-->[a]	ayuntamiento 1846
o(ill) --> [O]	oille 1673
(o)ill --> [j]	oille 1673
u -->[w]	guanaco 1766
j --> [x]	navaja 1840
l-->[j]	llanos 1598
e-->[e]	jerez XVIII
e(z) --> [E]	jerez XVIII
z-->[s]	zapatéado 1845
u(n) -->[u]	ayuntamiento 1846
e(n) -->[E]	ayuntamiento 1846
s("#") -->[s]	llanos 1598
z--> [z]	zapatéado 1845

Table 1: In the following recapitulatory table we present the graphemes for the lexical items coming from Spanish.

2.4. The Polyphons

In the grammar is performed a specific processing of polyphons. For each lexical item we propose the different pronunciations attested by (9). Following this reference, *design* is rewritten [dezajn] and [dizajn]. In most cases, a single lexical item *(pull-over)* (¨pullover¨) shows a pronunciation regular to the French general system [pylOvER], and another one, more faithful to the original [pulOvœR].

Here the list of the phonetic conversions of Spanish polyphons.

ien -->[jEn] / [E~]	pronunciamiento 1838
un -->[un] / [õ]	pronunciamiento 1838
ll -->[ll] / [j]	olla-podrida 1590
ch -->[S] / [tS]	chistera 1905
x -->[x] / [k] / [gz]	xérès XVIII
j--> [x] / [J]	navaja 1840
ñ --> [ñ] / [nj]	cañon 1877
on -->[On] / [õ]	cañon 1877
j--> [d3] / [3]	banjo 1859

Table 2: The polyphons of Spanish.

3. THE INFLEXIONAL SYSTEMS OF LOANWORDS

3.1. Linguistic Concerns

We shall now draw attention to a particularly interesting point when dealing with loanwords, that is their inflexional systems. This study field deserves a particular interest for at least one reason: the inflexional paradigm of a loanword often gives clues about its degree of integration, even if this statement will need to be qualified.

The inflexional paradigm seems to be a good material to establish in synchrony a kind of graduated scale indicating the degree of integration of one loanword or another coming from a given source language. Going back to the example of *tomate*, it is a singular feminine lexical item which forms its plural following the most common procedure in French, that is adjoining a graphic *s*. The pronunciation, for its part, is not modified, which is not in accordance with the Spanish system rules to pronounce plurals in *s*, and is still farther from the Nahuatl system, which forms its plural by adjoining a vowel assimilated in French to [i], *(tomatli)*. Very few French speakers would, however, identify the item *tomate* as a loanword.

Different configurations were distinguished and will be described hereafter in extension. In the first configuration, the inflected form is borrowed at the same time as the canonical form and is phonologicaly tranferred into French *(yacht(s)man)*. Secondly, the canonical form is inflected in accordance with the French inflexional system, following the canonical form phonological transfer *(scout)*. The third operation is the one where the inflexional attributes (mainly the distinction female/male) of the form are modified after the transfer *(blinis)*. Finally, the form is inflected according to a system borrowed from the source language, which may lead to erroneous forms *(tenniswoman)*.

3.2. The Inflected Forms Automatic Generation

Some tools were developped, in a Hypercard environment, to generate automaticaly the inflected forms of two lexicons, a lexicon of nouns, which can also be used (or not) as adjectives, and a lexicon of adjectives, which can also be used (or not) as nouns (the functionings are not connected).

These lexicons are extracted from a larger one because it also includes verbs, le "60 000" of ICP. In these generators, each lexical item is submitted to operations described beforehand in (4), which classes according to the inflexional paradigm, the gender, the number and the grammatical category of the item, the difference in the pronunciation between the canonical form and the inflected form(s).

The units we handle are not linguistic units, but substitution minimal units, even if the two sometimes overlap.

Let us take the example of *pécheur* ("male sinner") and *pécheresse* ("female sinner"). The substitution minimal units are *-ur/-resse* (meaning : *ur* substitutes to *resse)*: the functioning unit and the linguistic unit do not overlap. The division into morphemes leads to *péch-eur/péch-eresse*. Contrary to *carbonaro* and *carbonari* for which the units overlap (-o/-i).

There is no morphological analysis module even if the substituted units are clues about the formation mode of the inflected forms when they appear with recurrence.

For instance, the inflexional behaviour of *jazzman (jazzmen)* is the same as for *rugbyman (rugbymen)*, ("rugby player", "rugby players"), *policeman (policemen)* and fifteen other lexical items of our classification. These items, indeed, belong to the same sub-system. This example shows the recurrence of an English plural creation mode and illustrates an interesting phenomenon. If the target language speaker is conscious about the fact that the item belongs to a given source language, here English, then this speaker may deduce the inflexional behaviour of an unmarked form, at least partly morphologicaly identical. This sometimes creates erroneous forms (erroneous compared to the source language forms), as we may show further on.

3.3. The Sub-systems

The automatic phonetic transcription of our lexicon with Toph, together with all the corresponding inflected forms, made it possible to distinguish the major cases of transfers, graphical and phonological. We will expose them following a gradation that could be the one used priorly.

First case. The target language takes in its whole the inflexional paradigm of the loanword subjected to borrowing. We mean from the graphic point of view. Then it is submitted to a phonological transfer as faithful as possible to the source language.

For example the paradigm *yacht(s)man, yacht(s)men, yacht(s)woman, yacht(s)women,* introduced into French in 1859, borrowed from English, has been integrally transferred into French. The pronunciation of the uninflected form, as well as the pronunciation of the inflexional morphemes, is close to the English one. An intermediate level of treatment might be the "rush" example: the canonical form is also directly transferred, but the inflexional morpheme has a

mixed treatment: its pronunciation is done the French way. Example: English => rush, rush**es**/[rUS], [rUS**ez**], French => rush, rush**es**/[RπS], [RπS].

Second case. The written code only takes the canonic form and the French system apposes its own inflexional morphemes : *boy scout [bOi skaut]* became *scout [skut]*, and *girl scout [g´:l skaut], scoute [skut]*. A mixed functioning can also be found *(maximum [maksimOm], maximums [maksimOm], maxima [maksima])* where the inflected forms may come from either the Latin or the French systems. It may be appropriate at this stage to say a few words about the common reflexion which is done when dealing with the borrowing datation: it would justify the integration degree. The first apparition of *maximum* in French dates back to 1718, according to (9). If one only examines the examples of *maximum* and *yacht*, the theory might be acceptable. But it is given a rough handling by other examples such as *goy* (from Hebrew "christian"), which has the same characteristics as the first case (pluriel *goym [gOim]* ou *goyim [gOjim])* but whose borrowing is antecedent (XVI°). Other factors then have to be determining, one of them beeing possibly the use frequency (8) (7) (*maximum* was introduced more recently into the French lexicon but is more used than *goy*).

Third case. The French lexicon only takes the lexical item inflected form. The French system then recategorizes it into an unmarked form. It then looses its inflexion markers and superimposes on this new unmarked form its inflexional morphemes. Example: *blinis [blinis]*, borrowed in 1883 from the Russian lexical item *blini [blini])*, the plural form of *blin [blin])*. The French form then shows a redundant number. Haugen (6) evoked this phenomenon, calling it "hybrid inflection" (p.218). As for the final ‹s› sounding, it is certainly analogical.

Fourth case. The French lexicon only takes the stem and superimposes an inflexional paradigm, the closest to what is supposed to be the source language inflexion formation mode. The pronunciation is also supposed to be the source language one. Example: *tennisman [tenisman], tenniswoman [teniswuman], tennismen [tenismEn], tenniswomen [teniswumEn]*. These forms are pseudo-anglicisms, created around 1930, following the model of the lexical items having the same inflexional behaviour as *sportsman*.The English lexicon does not distinguish between genders, only between numbers: *tennis player, tennis players*.

The four main cases of functioning stress on the fact that the adaptation of foreign items of all linguistic levels to the target language patterns, such as the incorporation of verbal and nominal suffixes, gender attribution, the creation of the inflexional paradigm, etc... are hints about the degree of integration of the lexical items in this language.

4. CONCLUSION

The linguistic origin of a lexical item is decisive for the way it will be pronounced. A lexical item that has been borrowed from another language has to be pronounced following the rules that are supposed to be compatible to those of the source language. But for one given source language, we have to take into account the different stages of introduction, corresponding to different uses. The contemporary stages are gathered in our grammar into lexicons of lexical items found into (9) (description in extension). Following this idea, we could contemplate, as what is done in the general system, extending this sub-system to all the entries possibly answering the same characteristics (datation, usage...). Indeed, there are two sorts of loans lexicons, those that have been introduced into French for a long time (and go on evoluating), and those which entered French contemporarily. Contrary to the first ones, the second cannot be defined extensively in the grammar. As a consequence, such a list of words cannot be exhaustive. Each word recently borrowed follows this law. The best solution, towards which we are heading is the automatic detection of the source language.

5. REFERENCES

1. Aubergé, V., and Belrhali, R., "La phonétisation automatique du français : émergence de règles ou de lexiques ?" Revue LIDIL n°13, 1996.

2. Belrhali, R., *Phonétisation automatique d'un lexique général du français : systémique et émergence linguistique*. Thèse de l'Université Stendhal, Grenoble□3, 1995.

3. Catach, N., *Orthographe et lexicographie*, tome 1, éditions Didier, Paris, 1971.

4. Chatti, S., *Catégories lexicales du français*. T.E.R. de maîtrise, Université Stendhal, Grenoble 3, 1991.

5. Ghneim, N., *Relations entre le code de l'oral et de l'écrit : contraintes et ambiguïtés*. Thèse de l'Université Stendhal, Grenoble□3, 1997.

6. Haugen, E., "The analysis of linguistic borrowing", "Language" n°26, pp.210-231, 1950.

7. Holden, K., "Assimilation rates of borrowings and phonological productivity", "Language" n°52, pp.131-147, 1976.

8. Poplack, S. and Sankoff, D., "Le trajet linguistique et social des emprunts". "Revue québécoise de linguistique" Vol 14, pp.141-186, 1984.

9. *Petit Robert 1*, Edition Le Robert, 1991.

10. Sannier, F., and Aubergé, V. and Belrhali, R., "La phonétisation des morphèmes flexionnels du français dans le système TOPH". Actes des 1ères JST, AUPELF_UREF Francil, Avignon, pp.463-468, 1997.