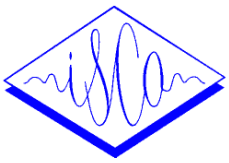

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

FROM LEXICON TO RULES: TOWARD A DESCRIPTIVE METHOD OF FRENCH TEXT-TO-PHONETICS TRANSCRIPTION

Belrhali R., Aubergé V. & Boë L.-J.

Institut de la Communication Parlée
URA CNRS n° 368 INPG/ENSERG Université Stendhal
BP 25 38 040 Grenoble France

ABSTRACT

This study presents a linguistic analysis of phenomena in French text-to-phonetics transcription and its goal is text-to-phonetics transcription dictionaries of words and compound words. The descriptive language TOPH (*Transcription Orthographique PHonétique*) [2] is the formal and computational tool organizing the description proposed here, and it permits the integration of rules and lexica into a text-to-phonetics grammar.

A minimal grammar, constituting the core of the phoneticization process, has been enlarged by systematically exploring a representative lexicon of French, the ICP's dictionary *Le 60 000*, with phonetic transcriptions being taken from the dictionary *Le Petit Robert* [13]. Several linguistically defined classes of morphemes or words emerged from this analysis, and each class is represented by a lexicon and activation rules. Through this extension, the rate of successful formalization has been improved to nearly 100% for the dictionary as a whole.

10.21437/ICSLP.1992-146

I. INTRODUCTION

Phoneticization of French has a long history behind it; one of the first works on the subject was, for example, that of Grammont [9]. Its study has been motivated by a number of reasons: linguistic description [6, 10, 16] with possible applications in language teaching, the study of phonetic variants and its use in speech and speaker recognition, machine correction of spelling errors [11] and text-to-speech synthesis [8, 14, 12].

The goal of this study is to propose a linguistic description of transformations of conventional spelling into phonetic form, the whole being organized as a computable algorithm. This description is formalized in the TOPH language [1] and is based on dual grapheme and phoneme representations that are related through a rewriting process which is 1) by rule and 2) defined by lexica.

We present here the method used to systematize in a nearly exhaustive manner grapheme-phoneme associations found in a representative French dictionary, the ICP's *Le 60 000*. We will attempt to classify these associations according to the linguistic level of the information conveyed (that is, according to the graphemic, morphophonemic, lexical, categorical, syntactic or semantic level). We will show that a solution does not require a making explicit of the attributes of each of these linguistic levels. The architecture of TOPH imposes on the grammar one representation which is sufficient under most conditions: rules activate lexica. Lastly, we will present the characteristics of the grammar elaborated in a HyperCard environment.

II. METHOD

The central core [2] is a grammar consisting of base rules in which essential types of lexica were identified. This initial description was used to transcribe phonetically the dictionary *Le 60 000* (which in the beginning contained only words in conventional spelling). Every phonetic word was then systematically verified and corrected by referring to the dictionary *Le Petit Robert 1* [13]. This new material was analyzed in order to extend the grammar with additional irregular formations and to complete the lexica [3]. It should be pointed out that classification of lexica by rules, which are in turn subject themselves to different formalizations, may be considered *a priori* from two independent points of views 1) the purely logical organization: rules (and lexica) have been grouped together whenever possible; we will describe in the next paragraph what the notion of conciseness means in the TOPH language 2) the linguistic organization: the structure of a rule corresponds to a linguistically identifiable mechanism and elements of one and the same lexicon share the same linguistic behavior. One of the more remarkable (and reassuring!) points brought out by this study is that the purely logical organization is generally congruent with the linguistic interpretation.

III. THE TOPH LANGUAGE

TOPH is a language and software environment developed at the ICP for multilingual synthesis [2]. It transcribes texts in usual typography without prior constraints on their form. One may note, for example, that in the HyperCard environment chosen in order to elaborate, on the one hand, a grammar of French and, on the other, a data base for the dictionary *Le 60 000*, the phonetic codes employed are those of the IPA.

A TOPH language grammar consists of two parts: 1) one part, which may be empty, of set declarations 2) another part constituted by rules whereby the previously defined sets may be called. Although described in identical syntax, the sets have several different linguistic functions:

a) typographical identification, for example a set "#" of word separations may be declared by:

"#" = (/(, /), -, / , /, , ., :, ;, !, ?, ", ')

(the character "/" neutralizes syntactic characters of TOPH)

b) phonetic classification, for example, the set of nasal consonants may be declared by "nasal consonant" = (n, m, ɲ)

c) formation of irregular words, for example, "lexica of those "-s" pronounced at the end of the word":

(as, bus) ("ace, bus").

A rule describes the association of any input string, of which the minimal unit of meaning is the typographical

substring (graphemes), with any output string (phonemes). The rules are integrated into classes, a class being defined by the first character of the string to be transcribed. During the transcription process, the rules are examined in the order in which they are written and the first validated rule is then applied to the input text. Left and right contexts of the string to be transcribed may weight the rewriting process. A context is defined by:
1) a graphemic sub-string
2) a set defined in the declarations
3) the enumeration of 1) or 2) noted ","
4) the concatenation of 1) or 2) noted "+"

The following illustration exemplifies the notion of conciseness mentioned in the preceding paragraph. These two rules:

("#" + a) +s+ ("#") = [s]

(the grapheme "s", preceded by a word boundary + "a", and followed by a word boundary is pronounced [s])

("#" + bu) +s+ ("#") = [s]

are merged into one:

("#" +a,bu +s+ ("#") = [s]

Then the enumeration of contexts is described by a lexicon, which is finally called on in the last rule:
("#" +"lexica of those "s" pronounced at the end of the word") +s+ ("#") = [s]

It is apparent that the difference between rule and lexicon becomes fuzzy in this type of terminal rule because most of the mechanisms may be interpreted as the association (under contextual constraints) between graphic form and phonetic form, and therefore as lexica of associations, for example: the grapheme "o" is associated with [o] and the word "as" ("ace") is associated with [as].

IV. CLASSIFICATION OF PROCEDURES AND SOLUTIONS

The process of phoneticization conveys linguistic information on the graphemic, lexical, morphophonemic, categorical, syntactic and semantic levels.

4.1 The graphemic level

The graphemic level corresponds to a regular rewriting of a graphemic string into a phonetic string. Thus it is a phonotactic level which ignores both the global form of the word and higher more complex levels. For example:
• the string "eau" ("water") is transcribed as [o];
• if "s" is found between two vowels, it is rewritten [z]; otherwise it is rewritten [s].

4.2 The morphophonemic level

French lexical units are composed of at least one lexical base and a number of affixes. Thus a graphic word must be considered, as far as its phoneticization is concerned, as being made up of different types of morphemic units:

4.2.1 prefix + base
• "asocial" ("asocial") is pronounced as "a" and "social", "s" is transcribed [s] instead of [z];
• "polyacide" ("polyacide") is pronounced "poly" + "acide", "ya" is transcribed [ia] instead of [ja].

4.2.2 base + suffix

Here we meet the problem of the different pronunciations of the graphic forms "tie", "tion", where "t", when not preceded by "s", is transcribed either [s] or [t]. One may assume that "t" is pronounced [s] when "tie" and "tion" are suffixes. But when these suffixes are derivations, there subsists a problem of categorical ambiguity (see 4.4). For example, "t" is pronounced [s] in "acrobatie" ("acrobatics") and [t] in "sotie" ("farce"). When word lists taken from either Juilland's dictionary [10] or the dictionary *Le 60 000* are analyzed, it is clear that this rule is not valid in all cases. To solve this kind of problem, we propose the systematic creation of specific lexica.

4.2.3 The construction base + base

This is a marginal construction in French, but one may find traces of it in, for example: "tournesol" ("sunflower") which is pronounced as "tourne" ("turn") and "sol" ("sun", derived from "soleil").

4.3 The lexical level

A set of words may share the same rewriting specificity. Very often this fact can be explained in that the set has undergone the same process under similar circumstances (loan words of like origin or similar use). These sets of irregular words will then be placed in the appropriate lexica. We will qualify these as exceptional words.

It is then possible to construct a complete list of lexica and to determine and verify their relation to linguistically identifiable classes.

We will present in section five the name and size of the more important lexica drawn from the ICP's dictionary *Le 60 000*. The description of each is given in Belrhali and Libert [3]. Excluding some rare exceptions, the lexica contain loan or "learned" words. For example:
• "ch" is pronounced [k] in learned words (which are generally of Greek origin);
• "er" is pronounced [e] in verbal inflection: "chanter" ("sing") as well as at the end of polysyllabic nouns and adjectives: "boulanger" ("baker"), "léger" ("light");
• "er" is pronounced [ɛʀ] at the end of monosyllabic nouns and adjectives as in "cher" ("dear") as well as at the end of loan words such as "bulldozer", "bitter", "bookmaker".
Exceptions are unusual and are brought about by historical accident or frequent use:
• "ch" is pronounced [ʃ] in "trachée" ("trachea"); current use has differentiated it from cognates where "ch" is pronounced [k]: "trachéotomie";
• "pt" is sometimes pronounced [pt] in "dompter" ("tame"), but the standard pronunciation given in the dictionary *Le Petit Robert 1* is [t]. The word's pronunciation and typography has evolved diachronically: the initial form "domter" (cf. Latin "domitare") has given rise to the present form "dompter" by analogy to "compter" ("count");
• the dictionary *Le Petit Robert Oral-Ecrit* gives "dilemme" ("dilemma") as [dilɛm], but also as [dilɛmn] by analogy to "indemne" ("unscathed") [ɛ̃dɛmn];
• "carrousel" borrowed from Italian "carosello" in the 17th century has maintained the 's' and is pronounced [kaʀuzɛl].

4.4 The categorical level

From the standpoint of phoneticization, several graphic markers (derivations and inflections) are ambiguous at the strictly phonotactic level. In particular, the endings "ent", "tions" have different pronunciations following word category:
• "ent" is not pronounced when it constitutes the verbal inflection of the third person plural present: "président" ("preside");
• "ent" is pronounced [ɑ̃] at the end of nouns, adjectives adverbs and a few verbal forms: "vent" ("wind"), "lent" ("slow"), "souvent" ("often"), "sent" ("feels");
• "tions" is pronounced [tjɔ̃] when it is an inflectional marker of the third person plural of the imperfect tense:

"chantions" ("sing"), except in "balbutions" ("stammer"), "argutions" ("quibble"), "initions" ("initiate") where it is pronounced [sjɔ̃];
• "tions" is pronounced [sjɔ̃] at the end of plural nouns, "rations" ("rations").

If a linguistic description consists of reducing these ambiguities by categorizing the word under question, a practical solution would be to create lexica of "non-verbal" forms when the forms are not homographic (cf. section five). But when two linguistic units are homographic, the categorical ambiguity is multiplied by a lexical ambiguity that can not be reduced by the TOPH grammar:
• "se fier" ("trust") [fje] verb vs. "fier" ("proud") [fjɛʀ] adjective;
• "ent": "président" ("preside") [pRezid] verb vs. "président" ("president") [pʀezidɑ̃] noun;
• "tion": "portions" ("carry") [pɔʀtjɔ̃] vs. "portions" ("portions") [pɔʀsjɔ̃] noun;
• inflectional ambiguity: "un os " ("a bone") [œ̃n-ɔs] singular vs. "des os" ("bones") [dez-o] plural.

4.5 The semantic level

Finally, there exists a class of ambiguities which can not be resolved by the syntax due to categorical and lexical homomorphism. Only the semantic or *a fortiori* the pragmatic levels permit the reduction of this kind of ambiguity. Such is the case for the nouns:
• "les fils" ("the sons") [fis] vs. "les fils" ("the threads") [fil];
• "jet" ("gush") [ʒɛ] vs. "jet" ("jet") [ʒɛt].

V. THE FRENCH TOPH GRAMMAR

We have not included in the grammar the lexicon of words ending in "-ent" [ɑ̃] because it contains too many words (2,688) in comparaison with other lexica. It will be replaced by the lexicon of verb forms in "-ent". Having not yet been compiled from our dictionary *Le 60 000*, the precise number of words in this lexicon is not known, but it should be of much smaller size. The grammar now constitutes twenty lexica containing 1,650 words and a set of 1,270 rules.

Though it is difficult to decide whether a rule defines an exception or a regular rewriting, it is estimated that 600 of them describe about 200 exceptional cases and the other 670 characterize the basic rewriting process. Classified according to linguistic level, the following lists are examples of lexica classified according to the lexical, morphonemic, categorical and syntactic levels.

5.1 The lexical level

Lexica of words:
• beginning in "ai" pronounced [e] and not [ɛ], 25 words:
"aider", ("help"), "aigri" ("sour")...
• ending in silent "-c", 20 words:
"banc" ("bench"), "blanc" ("white")...
• ending in a pronounced "-d", (loanwords), 42 words:
"background", "barmaid"...
• containing an "-e" pronounced [e], 52 words:
"referendum", "revolver"
• ending in "-et" pronounced [ɛt], 16 words:
"cricket", "gadget"...
• ending in "-ey" pronounced [ɛ], 13 words:
"jockey", "bey" ("lord")...
• containing a "-gu-" before "i" pronounced [g], 7 words:
"anguille" ("eel"), "déguiser" ("disguise")...
• ending in "-g" which is not pronounced, 24 words:
"bast(a)ing" ("beam"), "bourg" ("town")...
• beginning with an aspirated "h-", 403 words:
"ha", "hâbleur" ("boaster")...
• ending in "-ing" pronounced [iŋ], 166 words:
"bowling"," brainstorming"...
• ending in "-d" which is not pronounced, 22 words:
"cul" ("bottom"), "outil" ("tool")...
• containing "-oo-" pronounced [u], 18 words:
"hollywood", "lambswool"...
• ending in "-p" which is not pronounced, 15 words:
"beaucoup" ("many"), "camp" ("camp")...
• beginning with "kw-" pronounced [kw], 53 words:
"quadragénaire", "quadragésimal"...
• ending in "-s" pronounced [s], 510 words:
"bus", "acinus"...
• ending in "-t" pronounced [t], 194 words:
"abject", "abrupt"...
• ending in "-un-" pronounced [ɔ̃], 22 words:
"acupuncture", "carborundum"...
• ending in "-x" which is not pronounced, 44 words:
"afflux", "eaux" ("waters")...
• beginning with "gn-" pronounced [ɲ], 8 words:
"gnon" ("punch"), "gniaf" ("cobbler")...
• ending in "-er" pronounced [ɛʀ], 86 words:
"africaner", "container"...
• ending in "-er" pronounced [œʀ], 33 words:
"bitter", "bookmaker"...

5.2 The morphophonemic level

• Lexicon of prefixes ending in a vowel and which cause the following "s" to be transcribed as [s]:
"bi-, carbo-, entre-, homo-, ultra-"...
• Lexicon of prefixal bases (Greek stems) ending in a vowel and, as above, which cause a following "s" to be transcribed as [s]:
"bio-, éco-, hypo-, -, thio-, zoo-"...
• Lexicon of prefixes ending in "y" which cause the following "a" to the pronounced [ia]:
"poly-, oxy-, tachy-",...
• Lexicon of prefixes ending in "i" which cause a following "a" to be transcribed as [ia] and not [ija]:
"di-, bi-"...

5.3 The categorical level

• The lexicon of words ending in "-tions" pronounced [sjɔ̃] contains 1,780 nouns (i.e. "non-verbs"): "abdications", "abductions", "abréviations"... Given their number, we have therefore maintained the lexicon of verb endings with the same graphic form, 825 words:

verb forms ending in "-tions" (first person plural of the imperfect) pronounced [tjɔ̃], that is, all verbs in "-tions" with the exceptions of "argutions", "balbutions", "initions";

verbs in "ter", "mentions" ("lie"), "portions" ("leave")..., verbs in "-ter", "abritions" ("shelter"), "absentions" ("absent")...
• The lexicon of words ending in "-ent" pronounced [ɑ̃], that is all "non-verbs", 2680 words:
"abaissement" ("lowering"), "abâtardissement" ("bastardizing"),...

5.4 The syntactic level

The lists are given simply as a supplement since a lexicon of these forms integrated into TOPH will not permit resolution of syntactic ambiguities:
• a lexicon of homographs ending in "-tions" pronounced [tjɔ̃] in verb forms and [sjɔ̃] for noun forms, "ablations": ("ablations") noun [-sjɔ̃] vs. ("we were ablating") [-tjɔ̃] verb;

• lexicon of homographs ending in "-ent" which are not pronounced in verb forms, but are pronounced [ɑ̃] in other categories, "couvent": ("convent") [-ɑ̃] vs. ("they brood") [-ø] noun.

VI. CONCLUSION

The dictionary *Le 60 000* has been phoneticized by a new version of TOPH grammar. The rate of successful phoneticization has been increased to nearly 100%. Failure results from those particular cases described above which can not be formalized by a grammar displaying the rewriting structure of TOPH.

In this study we have insisted on the essential role of the lexicon in linguistic description. When integrated into grammar, lexica should not merely be considered as a collection of data but as natural representations in a mixed lexicon-rule model such as that proposed by TOPH.

The goal of this study was to reveal by thorough descriptive procedures the problems involved in word phoneticization. It is now being continued by the next logical step, the study of compound words [4].

The second part of this work will be to define in the dictionary *Le 60 000* a set of lexical categories specific to speech communication (for example, to reduce ambiguity in text-to-phonetics transcription and intonation [7]).

By introducing a field of category and by adapting a model of flexion [15], we will be able to generate approximately 450,000 grammatical forms. A new grammar will be established on the basis of this lexicon of forms and it will then be completed by a study of variants taken from a data base of spoken French BDPHO [5].

ANNEXE

An extract of the phoneticization grammar: declarations and examples of rules for the grapheme "a". (**TOPH** language, version 2.1)

I. DECLARATIONS

1.1 Sets

• Linguistic

"vowel" = (a, à, â, e, é, è, ê, ë, i, î, ï, o, ô, u, ù, ü, û, y)

"consonant" = (b, c, ç, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, z)

"non-nasal cons." = (bb, cc, dd, ff, gg, ll, mm, nn, pp, rr, ss, tt)

...

• Word boundaries

"#" = (/(, /), -, / , /, , ., :, ;, !, ?, ", ')

1.2 Lexicon

"words where 'ai' is pronounced [e]"

aider, aigri, aigrir, aigu, aiguillage...

II. RULES

2.1 Exception rules

• English loanwords

("#"+ch,m,p) +a+ (ce,de,ker,se) = [ɛ]
steeple-chase, pace-maker, self-made-man

("#"+b,c,footb,w) +a+ (ll,lk) = [o]
ball, call-girl, football, walk-over

• German loanwords

("#"+schn) +au+ (zer) = [aw] *schnauzer*

("#"+sprechges) +ang+ = [aŋ] *sprechgesang*

• Italian loanwords

("#"+qu) +an+ (ti) = [an] *tutti quanti*

...

• Spanish loanwords

("#"+c) +au+ (dillo) = [aw] *caudillo*

+a+ (yuntamiento) = [a] *ayuntamiento*

• Others

+am+ (h,s) = [am] *amharique*

+am+ ("non-nasal cons.","#") = [ɑ̃] *amphibie*

("#") +an+ (h) = [an] *anhydride*

+an+ ("#","non-nasal cons.",m) = [ɑ̃] *manger*

...

("#"+f,p,t) +aon+ = [ɑ̃] *paon*

+aim+ ("#") = [ɛ̃] *daim*

+ain+ ("#","non-nasal cons.") = [ɛ̃] *pain*

+a+ (il+l,"#") = [a] *travail*

("#") +aî+ (né) = [e] *aîné*

...

("#"+abb) +aye+ = [ei] *abbaye*

("#"+f) +a+ (ya) = [a] *fayard*

("#"+b,c,cob,cong,rim) + a+ (yé) = [a] *bayer*

("#"+f,m) +a+ (yo) = [a] *fayot*

("#"+f) +ai+ (san) = [ə] *faisan*

("#"+qu,g) +ai+ ("#") = [ɛ] *quai*

("#"+g,s,p) +aie+ (tt,rie,"#") = [e] *paierie*

("#"+f) +ai+ (san) = [ə] *faisan*

("#") +ai+

("words where 'ai' is pronounced [e]") = [e]

+a+ (y+é,er,iste) = [e] *payer*

+a+ (yure) = [e] *balayures*

+a+ (y+"vowel") = [ɛ] *maye*

2.2 General rules

+ai+ = [ɛ]

+âi+ = [a]

+aî+ = [ɛ]

+a+(ï) = [a]

+æ+ = [e]

+à+ = [a]

+â+ = [a]

+au+ = [o]

+ay+("#") = [ɛ]

+aa+ = [a]

("#")+a+("#") = [a]

+a+ = [a]

REFERENCES

[1] Aubergé V., *La synthèse de la parole: des règles aux lexiques*, Doctoral Thesis, University P. Mendès France, Grenoble, 1991.

[2] Aubergé V., Contini M., Maret D. & Schnabel B., *TOPH: un outil de phonétisation multilangue*, Bull. de l'Inst. de Phon. de Grenoble, 16, 155-176, 1987.

[3] Belrhali R. & Libert L., *Phonétisation automatique: évaluation et enrichissement d'une grammaire de phonétisation du français*, DEA Thesis, University Stendhal, Grenoble, 1991.

[4] Berlrhali R., Libert L., Aubergé V. & Boë L.J., *Organisation de lexiques pour une grammaire de phonétisation du français: "Le 60 000" de l'ICP.*, 19e Journées d'Étude sur la Parole, Bruxelles, 1992.

[5] Boë L.J. & Tubach J.P., *BDPHO: une base de données lexicales orthographique-phonétique du français parlé*, Séminaire Lexique, GDR-PRC Com. HM, Toulouse, 1992.

[6] Catach N., *La phonétisation automatique du français*, Ed. CNRS, Paris, 1984.

[7] Chatti C., *Catégories lexicales du français*, Maîtrise, University Stendhal, Grenoble, 1991.

[8] Divay M. & Guyomard M., *Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français*, Doctoral Thesis, University of Rennes, 1977.

[9] Grammont M., *Traité pratique de prononciation française*, Ed. Delagrave, Paris, 1926.

[10] Juilland A., *Dictionnaire phonétique inverse du français*., Mouton, The Hague, Paris, 1965.

[11] Lahens F., *Un modèle stochastique pour la vérification et la correction automatique de textes: le système VORTEX*, Doctoral Thesis in Computer Science, CERFIA, Toulouse, 1987.

[12] Laporte E., *Méthodes algorithmiques et lexicales de phonétisation de textes, applications au français*, Doctoral Thesis, University of Paris VII, 1988.

[13] Robert P, *Le Petit Robert 1*, Société du Nouveau Littré, Le Robert, 1990.

[14] Prouts B., *Contribution à la synthèse de la parole à partir du texte, transcription graphème-phonème en temps réel sur microprocesseur*, Engineering Thesis, University of Paris Sud, Orsay, 1980.

[15] Rouault J., *Linguistique automatique, Applications documentaires*, Sciences pour la Communication, Peter Lang, Berne, 309 p., 1987.

[16] Véronis J., *Correction of Phonographic Errors in Natural Language Processing*, 8th Int. Conf. on Computers in the Humanities, Colombia, 1987.