Computer Speech and Language (1998) **12,** 393–410
Article No. la980104

Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French

F. Yvon[††], P. Boula de Mareüil[*], C. d'Alessandro[*], V. Aubergé[†], M. Bagein[‖], G. Bailly[†], F. Béchet[‡], S. Foukia[‡‡], J.-F. Goldman[§], E. Keller[¶], D. O'Shaughnessy[**], V. Pagel[‖], F. Sannier[†], J. Véronis[‡‡] and B. Zellner[¶]

[*]LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France, [†]ICP, CNRS ESA 5009, DU, 1180 avenue Centrale, BP25 - 38040 Grenoble Cedex 9, France, [‡]LIA, Université d'Avignon, 339 chemin des Meinajaries, BP 1228, 84911 Avignon Cedex 9, France, [§]LATL, Université de Genève, 2 rue Candolle, 1211 Genève 4, Switzerland, [¶]LAIP, Faculté des Lettres, Université de Lausanne, 1015 Lausanne, Switzerland, [‖]TCTS, Faculté Polytechnique de Mons, 31 boulevard Dolez, 7000 Mons, Belgium, [**]INRS-Télécommunications, 16 Place du Commerce, Ile-des-Soeurs, Verdun, Quebec, H3E, 1H6, Canada, [††]ENST & CNRS, URA 820, 46 rue Barrault, 75013 Paris, France, [‡‡]LPL, CNRS ESA 6057, 29 avenue Robert Schuman, 13621 Aix-en-Provence Cedex 1, France

Abstract

This paper reports on a cooperative international evaluation of grapheme-to-phoneme (GP) conversion for text-to-speech synthesis in French. Test methodology and test corpora are described. The results for eight systems are provided and analysed in some detail. The contribution of this paper is twofold: on the one hand, it gives an accurate picture of the state-of-the-art in the domain of GP conversion for French, and points out the problems still to be solved. On the other hand, much room is devoted to a discussion of methodological issues for this task. We hope this could help future evaluations of similar systems in other languages.

1. Introduction

Text-to-speech (TTS) synthesis systems usually involve three main submodules applying in sequence. The first stage of the synthesis, traditionally referred to as grapheme-to-phoneme conversion, consists of translating a written utterance into the corresponding stream of phonemes (including, for some languages, the encoding of lengthened phonemes, of lexically stressed syllables, of syllable boundaries, etc.). The second stage consists of computing a series of prosodic markers to be attached to this phonemic string. The last stage deals with the actual production of the speech waveforms. Speech synthesis can be viewed as a chain: the output quality depends on the quality of

0885–2308/98/040393 + 18 $30.00/0

individual components. It thus makes sense to conduct a specific evaluation for each part of this chain.

In language such as French (likewise, in English), grapheme-to-phoneme (GP) conversion is difficult. A first difficulty is that the French orthographical system is overly complex and contains, mainly for historical reasons, a large number of irregularities (Catach, 1984; Belrhali, 1995). As a result, any accurate rule-based description of the correspondence between graphemes and phonemes needs to incorporate a fairly large number of very specific rules and exceptions. These rules are further obscured by a number of so-called heterophonous homographs, i.e. word forms whose pronunciation varies according to the environment. A second difficulty is that the phonology of French presents some intriguing aspects, whose linguistic description is still subject to open controversies; these aspects need nonetheless be accounted for in a GP conversion system. A typical problem is the problem of the *mute-e* (or schwa), which may be either uttered or dropped (Larreur & Sorin, 1990), both word-internally and at the junction between successive words. Other problems of contextual variability occur in the case of glides, which may be uttered in a syllabic manner (diæresis) or not (synæresis), and in the case of *liaisons*. Liaison, here, refers to the phonetic realization of a word final consonant in the context of a following word initial vowel or *mute-h*, which can be compulsory, forbidden, or optional. An example of each of these possibilities is given in the sentence:

les enfants ont écouté
(the children have listened)

Liaison is compulsory between *les* and *enfants*, forbidden between *enfants* and *ont*, and optional between *ont* and *écouté*. In any case, predicting the accurate realization of these variable phonemes requires a subtle analysis of their phonological, morphological, and even syntactical environment. Finally, as is the case in most languages, *extra-lexical* items such as proper names, numbers and abbreviations, which are frequently found in real-world texts, also raise complex problems. A review of these difficulties, and of the solutions that have been put forward in the context of automated GP conversion, can be found, for instance, in Aubergé (1991), Béchet and El-Bèze (1996), Boula de Mareüil (1997), Divay (1990), Dutoit (1993), Gaudinat and Wherli (1997), Keller (1997), Lacheret-Dujour (1990), Laporte (1988), Marty (1992), O'Shaughnessy (1984) and Yvon (1996).

The extent to which these problems are solved, or still impair the quality of TTS, is however poorly appreciated. Therefore, the evaluation of GP conversion is an important problem for evaluating TTS systems. Quite an abundant literature exists about TTS evaluation methods (Silverman, Basson & Levas, 1990; Carlson, Granström & Nord, 1990; van Bezooijen & Pols, 1990; van Santen, 1993; Kraft & Portele, 1995; Sorin & Emerard, 1996; Klaus, Fellbaum & Sotscheck, 1997; Benoît, 1997; Pols & Jekosch, 1997). However, to the knowledge of the authors, no specific methods, corpora or state-of-the-art reports are currently available for GP conversion, and in particular for GP conversion in French.

The aim of this paper is to report joint experiments conducted in the French-speaking academic community on evaluation of GP conversions in French, for TTS synthesis. The systems involved are all relying on a rule-based approach; nonetheless, they differ greatly in the number of rules involved in GP conversion, which ranges between 500 and about 4000. In all the systems, rules may be superseded by look-up in exceptions lexica. These lexica may contain as few as a handful of very specific words, or as many as thousands of irregular word forms. These figures may be difficult to interpret, as there is no clear cut distinction between rules and exceptions. The systems also differ in the amount of linguistic pre-processing of the text: while some take advantage of a fairly accurate general purpose syntactic analysis module, others rely on *ad hoc* heuristics to perform morpho-syntactic disambiguation. Some systems also use specific modules for pronouncing proper names. However, this is not the general case. Precise description of the systems involved in this test is out of the scope of the present paper, and the reader is invited to refer to the paper cited above for more details.

The keypoints and main features of the present work are the following:

Objective evaluation: For a linguistic task such as GP conversion, it seems better to use automatic evaluation tools, based on enriched text corpora, rather than subjective tests, which are better suited to the assessment of complete TTS systems.

Diagnostic approach: Following the grid proposed in Gibbon, Moore and Winski (1997: chapter 12), we preferred a diagnostic approach. System developers need to systematically and objectively evaluate the behaviour of their programs at a very detailed level of precision, in order to concentrate their efforts on the most defective part of their system. Surprisingly enough, very little has been done so far to provide system developers with suitable, i.e. "objective", diagnostic evaluation methodologies.

International evalution: Eight different systems were tested in the experiments. They were provided by teams from Canada, Belgium, France and Switzerland in the following universities or research institute: ENST (Paris), LIMSI (Orsay), ICP (Grenoble), LIA (Avignon), INRS (Québec), TCTS (Mons), LAIP (Lausanne) and LATL (Geneva). LPL (Aix-en-Provence) was in charge of organization and corpora development.

Methodology design: One of the main issues, and one of the most difficult tasks, for GP converter evaluation was the design of a suitable methodology. This is discussed in detail below.

Corpus design: The corpus format and content designed for these experiments will be useful for other systems as well, since they will be at the disposal of the scientific community.

Strictly speaking, GP conversion refers to the process of converting a stream of orthographical symbols into an appropriate symbolic representation of the corresponding sequence of sounds. This output usually takes the form of a series of phonemic symbols. The usefulness of an automated GP conversion device has been demonstrated in various natural language and speech processing applications, such as the correction of spelling errors, speech synthesis and large-vocabulary speech recognition. Depending on the target application, the specifications of a GP converter are slightly different, and this should be taken into account in the design of the evaluation methodology. For instance, whereas speech synthesis systems usually output one single phonemic string for each input, GP converters used in the context of speech recognition ought to produce multiple pronunciations of their input, in order to model properly the speech variability. In this project, the evaluation methodology is tuned to the specific task of speech synthesis.

Even in this context, GP conversion modules can be given quite different tasks. This task can be as restricted as the pronunciation of *isolated items* from a potentially large list, as is typically the case for reverse directory inquiry systems, or can consist of the pronunciation of *richly annotated* linguistic representations, as in the context of dialogue systems, or more generally, concept-to-speech systems. Text-to-speech systems put other kinds of requirements on their GP conversion module, since their input may potentially be any kind of written text (news articles, e-mail, etc.). Ideally, an evaluation of GP conversion systems should include a specific experimental design for each of these tasks. Our experiments are nevertheless restricted to newspaper or book reading.

This paper is organized as follows. In the next section, we present the methodology. Section 3 describes the corpus design and content. Section 4 presents the results obtained for the eight systems. Section 5 discusses the methodology and summarizes the results obtained, which give an accurate state-of-the-art of GP conversion in French. Directions for future studies are also envisaged.

2. Method

At first glance, evaluating a GP conversion device looks quite straightforward, and is just a matter of comparing the output of a given system with one reference transcription, and of counting the discrepancies, the number of which is used as a measure of the quality of a system. This is what is traditionally done in the numerous studies dealing with the evaluation of self-learning techniques on the grapheme-to-phoneme conversion task (reviewed for example, in Damper, 1995). These studies however only consider the transcription of isolated entries taken from phonetic dictionaries, and there are good reasons to think that this kind of methodology is not directly suitable to our specific needs. The capabilities of GP conversion systems for French must go far beyond the pronunciation of isolated words. In addition to GP conversion rules and exception dictionaries for isolated words, they must include various pre-processing and phrase level phonology modules, which have to be evaluated as well.

2.1. Task and corpus target

The choice that was made was to evaluate our systems on running texts rather than on lexica. Various kinds of texts were considered (novels, newspapers, e-mails), and the decision was made to work on articles from the French newspaper *Le Monde*. The rationale was that this kind of resource was massively available (thanks to *Le Monde* who granted us the right to use its text material for research purposes) and that this kind of text contained a full assortment of the typical difficulties of GP conversion: complicated structures, citations, dates, proper names, acronyms and typing errors. Furthermore, a phonemic transcription of this kind of text corpus was unanimously felt to be of great relevance for the speech synthesis community. This corpus is described in Section 3.

2.2. Alphabet and variants

Another question we had to face was that of the definition of a common phonemic alphabet, a necessary step for making our systems comparable. A phonemic alphabet specifies simultaneously two things: the inventory of phonemes used in the transcription, as well as their representation. An examination of the alphabets used in the participating systems revealed that they were not only using different notational schemes, a harmless problem, but also that they were not encoding exactly the same kinds of phonemic distinctions. For instance, some systems did not keep the opposition between /a/ and /ɑ/ or between /ɛ̃/ and /œ̃/; some identified several types of schwas, noting schwas that are usually dropped, and schwas that are usually realized, with different symbols, etc. The definition of a common symbol list was therefore an important concern.

Consequently, the decision was made that the list of phonemic symbols to be used by all the systems would correspond to the intersection of the existing alphabets, and that these would be represented in the well-known SAMPA (Gibbon, Moore & Winski, 1997), which has already been used in the context of TTS assessment. The common alphabet that we decided to use is given in Table I. All the participating systems were consequently adapted to stay within this specific alphabet.

2.3. Evaluation protocol and scoring procedures

Stemming from the experience of the GRACE project (Paroubeck, Adda, Mariani & Rajman, 1997), devoted to the evaluation of morpho-syntactic analysers, the evaluation protocol consists of submerging, in a much larger corpus, the text on which the results are analysed (12 000 sentences in our case). This portion is of course secretly selected by the organizer. The task given to each participant consists of phonetizing the entire text within a restricted time frame (2 days in our case).

The alignment and subsequent comparisons between the different outputs and the reference corpus were performed on a per sentence basis, rather than on a per word basis. In fact, it was much easier to find agreement on a workable definition of a sentence than of a word (think about elisions, abbreviations, numbers and compound words). Irrespective of the possibility of agreeing on an alignment between orthographic and phonemic words, one major difficulty was that each system works with its own set of implicit assumptions regarding the exact definition of what a word is. These assumptions usually interact strongly with the way the transcription rules work, making a redefinition of this notion within each system quite problematic.

For each sentence, the scoring scheme adopted for this evaluation was to count uniformly as an error every difference with the reference transcription. This kind of scoring is arguably over-simplistic in the context of text-to-speech synthesis, since all the errors do not equally impair the intelligibility of the output speech. For lack of well-defined confusion metrics between phonemes which would numerically account for phonetic similarities, this kind of measurement was nonetheless used: at least, it represented a coherent choice with respect to our goal, which was to evaluate GP conversion devices as an independent module of a TTS system.

3. Corpus design

3.1. Transcription guidelines

The transcribed reference corpus must contain a significant number of pronunciation variants. Given the very large number of possible variants for a reasonably large sentence, the solution which has been adopted amounts to limiting the reference corpus

TABLE I. Phonetic alphabet for GP conversion assessment in French

Description	Phonetic code	IPA
[mAtin, pAs]	A	a, ɑ
[VANtardise, tEMPS]	A˜	ɑ̃
[pEtit]	@	ə
[crEUser, dEUx]	2	ø
[malhEUreux, pEUr]	9	œ
[pERdu, modEle]	E	ɛ
[Emu, otE]	e	e
[pEINture, lUNdi]	E˜	ɛ̃, œ̃
[Idée, amI]	i	i
[Obstacle, cOrps]	O	ɔ
[AUditeur, bEAU]	o	o
[rONdeur, bON]	O˜	ɔ̃
[cOUpable, lOUp]	u	u
[pUnir]	y	y
[OUi, OIseau]	w	w
[hUIle]	H	ɥ
[pIétiner, paiLLe]	j	j
[Phre, caPe]	p	p
[Terre, raTe]	t	t
[Cou, saC]	k	k
[Bon, roBe]	b	b
[Dans, aiDe]	d	d
[Gare, baGUe]	g	g
[Feu, PHare]	f	f
[Sale, taSSe]	s	s
[CHanter, maCHine]	S	ʃ
[Vous, rjVe]	v	v
[Ziro, maiSon]	z	z
[Jardin, manGer]	Z	ʒ
[Lent, giLet]	l	l
[Rue, veniR]	R	ʁ
[Main, feMMe]	m	m
[Nous, toNNe]	n	n
[aGNeau, rèGNe]	J	ɲ
[campINg]	N	ŋ

to a lattice of pronunciations. Some sample sentences, coded in SAMPA (the | symbol separates possible alternatives), are:

> *Les hommes et les animaux peuvent remuer, se mouvoir, se donner du mouvement.*
> lez Om[@z|z|]e lez Animo p9v[@|]R[@|]m[y|H]e s[@|]muvwAR s[@|]d[o|O]ne dy muv[@|]mA˜
>
> *C'est ce qui ressort d'une lecture d'affilée des nouvelles, dont la plupart ont précédé, avant, pendant et juste après la guerre, la série des grands romans : le Vent noir (1947), la Plage de Scheveningen (1952), l'Invitation chez les Stirl (1955), les Hauts Quartiers (1973).*

> s [e|E] s[@|] ki R[@|]sOR d yu[@|] lEktyR[@|] d Afile d[e|E] nuvEl[@|] dO~ lA plypAR O~ pResede AvA~ pA~da~[t|] e Zyst ApR[e|E] lA gER[@|] lA seRi d[e|E] gRA~ R[O|o]mA l[@|]vA~ nwAR[mil@|mil|diz]n9fsA~kARA~t[@|]sEt lA plAZ[@|] d@S[@|e|E] v[e|E]ni[N|ng][@|E]n [mil@|mil|diz]n9fsA~sE~kA~t[@|]d2 l E~vitAsjO~ Se l[e|E] stiRl [mil@|mil|diz]n9fsA~sE~kA~t[@|]sE~k l[e|E] o kARtje [mil@|mil|diz]n9fsA~swAsA~ t[@|]tREz[@|].

Such a format can easily be handled at the computational level and does not make any assumption regarding the scoring measures that are applied during the evaluation phase. Furthermore, this representation of phonological variants can easily be extended, to accommodate additional variants in an incremental manner.

3.2. Corpus production

In a preliminary test, a 30 000 word corpus was transcribed by the eight systems in the common alphabet (d'Alessandro *et al.*, 1997). This text has been entirely manually annotated by an expert phonetician, but only contains a single pronunciation for each sentence, except for the first 100. This dry-run mostly enabled us to verify that everyone had correctly taken into account the specific requirements on the output, both in terms of the phonemic alphabet, and in terms of the file format.

In a second phase, which took place during the summer of 1997, the organizer of the test campaign hand-crafted the test corpus transcription, which consists of articles extracted from the newspaper *Le Monde* of January 1987. About 2000 sentences were selected, with the specific concern that numbers, (foreign) proper names and acronyms should be significantly represented in the corpus. Overall, the corpus was transcribed in a relatively tolerant perspective, still in as reliable and satisfactory a way as possible.

At the orthographic level, this corpus contains about 26 000 word tokens, corresponding to 6000 different word forms. These word occurrences can be further subclassified between roughly 1500 proper names (corresponding to 1000 different word forms), 600 numerals (200 word forms) and 200 acronyms and abbreviations (90 word forms), the remaining lot being composed mainly of common words.

The specificities of the corpus in terms of its vocabulary were calculated with respect to the 1950–1990 period of the "Trésor de la langue française" corpus (Imbs, 1971). This study revealed no marked deviation in terms of vocabulary.

Once manually transcribed, the reference corpus contains a grand total of 85 000 phonemic symbols. Based on indications provided by the corpus producer, one can estimate the number of possible cases of liaisons to be approximately 1500, amongst which about 600 are compulsory. Similarly, the transcribed corpus contains 8500 cases of mute-e, further subdivided between 1000 obligatory deletions, 1500 obligatory realizations and 6000 optional deletions.

4. Evaluation results

4.1. Global results

The computation of results has been subject to a two-stage procedure. Scores were produced according to a first release of the reference corpus. An adjudication phase then took place, giving the participating teams the opportunity to contest some of their

TABLE II. Global performance of the eight systems

Labs	Number of phonemes	Correctness	Accuracy	Sentence correctness
A	83841	97·1	93·0	21·2
B	84250	94·9*	94·4*	53·2
C	85850	97·7	97·3	57·7
D	85554	98·4	97·8	59·6
E	86338	99·2*	98·8*	69·0
F	86205	99·2	98·7	72·1
G	86938	99·3	99·0	76·0
H	86047	99·6	99·5	89·1

For each system, this table gives successively the total number of phonemes predicted, the system's correctness and accuracy and the percentage of entirely correct sentences.

* These figures largely underestimate this system's performance, which have been severely degraded in terms of phonemic correctness and accuracy, by a non-negligible number of entirely incorrect sentences. A further examination of these sentences revealed that they were either incorrectly formatted, or had been wrongly aligned with the reference corpus.

errors. The main causes of disagreements were related to the insufficient description of variants in the reference corpus.

A new version of the corpus was then produced, in order to take into account these additional variants, and new scores were accordingly computed. Even if these figures cannot be taken at face value, for lack of a complete and general agreement on all the errors, they still reflect accurately the level of performance of the eight participating systems.

The raw results, obtained at the term of this adjudication phase, are displayed in Table II. This table distinguishes between correctness and accuracy: the former gives the percentage of phonemes correctly predicted, whereas the latter also takes into account the percentage of spurious insertions. It is also important to note that systems significantly differ in their treatment of optional phonemes: this fact is reflected in the important variability (nearly 5%) in the total number of phonemes produced.

Overall, the eight systems fared relatively well with the task at hand, since they all achieve at least 97% phonemes correct. There still exists significant differences, which are better reflected when one considers the percentage of entirely correct sentences: for this score, results vary between 20% and 90%.

These results, however revealing, are nonetheless of little interest for system developers. What is really needed is a detailed classification of errors enabling us to pinpoint the most problematic cases for each system. This kind of analysis is developed in the next section.

4.2. *Error analysis*

There exist various dimensions along which a system developer may wish to classify its transcription errors. A first important piece of information is to get a list of the word tokens which incurred a transcription error. A second type of information is related to word grammatical classes: in fact, many systems use different transcription rule sets for different types of words, and one may need to identify which rule set is the less reliable one. Getting a classification of errors according to the grammatical tag of the erroneous word is therefore a desirable result. GP conversion systems often comprise several submodules, which perform specific linguistic analysis, such as segmentation, pre-processing, morphological analysis, etc. Obviously, identifying which submodule is responsible for which error is also a key concern for the system developer. As far as pre-processing is specifically concerned, it is a fact that the typographical characteristics of a word token (is it capitalized, entirely uppercase? does it comprise arabic or roman numerals? . . .) often determines its processing in the GP conversion system. Furthermore, classifying errors along that specific dimension also provides useful indications regarding the robustness of the system when confronted with unexpected typographical inputs, and is consequently very relevant in the context of a detailed analysis. Finally, errors may also be classified in terms of their "seriousness": the confusion of two vowels that differ only in timbre will probably not impair the output speech as much as the mispronunciation of a well-known proper name will. In fact, for lack of a good definition of the seriousness of an error, which ultimately depends on the speech synthesizer used, only the first four dimensions were eventually explored.

To the best of our knowledge, very few attempts have been made at defining a sensible grid for classifying and analysing transcription errors (see, however, Laver, McAllister & McAllister, 1989; Cotto, 1992). In order to make such an analysis possible, a common grid was defined for classifying errors. Errors have consequently been manually annotated according to the following dimensions and categories:

- Related orthographical form. Identifying the word in error is relatively easy, except for a (small) list of problematic segmentations (numbers, compounds, etc.).
- Typographical characteristics of the word in error. We considered the following categories: lower case word, capitalized word, upper case word, arabic number, roman number and digit-bearing strings.
- Major grammatical categories. In fact, for the purpose of this evaluation, we simply distinguished between: proper names, acronyms and abbreviations, numbers and symbols, and lexical items. The choice of this restricted grid was mainly motivated by practical reasons: since there was no agreement on a grammatical tag-set, it was felt that this set of rather uncontroversial categories was sufficient to permit useful comparisons between systems.
- Error type. We specifically distinguished the following types: error on a phomeme liaison, error on a schwa, error on a loan word, error caused by the incorrect identification of a morpheme boundary, error caused by a typing error and error induced by an insufficient pre-processing of the text.

This annotation scheme is unsatisfactory in many respects, since it confuses various levels of errors (on a phoneme, on a word). A consequence is that in some cases, the appropriate class of an error might be subject to variation between annotators. For instance, an incorrect pre-processing may result in a liaison error; a morphological boundary may be missed in a borrowed word, etc. Furthermore, this grid fails to distinguish between types that are relevant for a specific system: for instance, a liaison might be omitted either because the grammatical tag of the word is incorrectly identified, or because the liaison rules fail to take that specific configuration into account, or because the grapheme-to-phoneme rules fail to predict a final latent consonant.

Nonetheless, this categorization still reflects the main difficulties of GP conversion

Table III. Fields of the four-axis grid used for the error analysis

Word axis	Tag axis	Error type axis	Typography axis
Orthographical string	Proper name	Borrowed word	Lower case
	Acronym	Liaison	Upper case
	Number/Symbol	Schwa problem	First letter capitalized
	Other	Heterophonous homograph	Arabic numeral
		Spelling mistake	Roman numeral
		Pre-processing	Mixed letters and figures
		Morphological ambiguity	
		Other	

Table IV. Distribution of errors by word category for six systems

	B	C	D	F	G	H
Proper name	180 (9·6)	296 (24·0)	204 (22·7)	113 (15·2)	168 (25·6)	131 (49·4)
Acronym	105 (5·6)	30 (2·4)	73 (8·1)	15 (2·0)	11 (1·7)	7 (2·6)
Number/symbol	113 (6·0)	157 (12·7)	55 (6·1)	90 (12·1)	50 (7·6)	30 (11·3)
Other	1469 (78·7)	752 (60·9)	565 (63·0)	526 (70·7)	428 (65·1)	97 (36·6)
Total	1867	1235	852	744	657	265

For a given system and word category, each cell of this table contains the absolute number of errors for this category, and the corresponding percentage of the errors for the system.

for French, and most annotators found it convenient and useful for diagnosing their system. Moreover, it appeared that controversial classification cases were sufficiently rare to make this grid an appropriate tool for comparing the strengths and weaknesses of the various systems on objective bases. A summary of the annotation scheme is given in Table III.

The annotated mistakes provided by six of the eight participating teams enabled us to conduct a more detailed quantitative analysis of the errors. Again, the results presented hereafter cannot be taken at face value, but merely reflect the relative importance of various GP conversion problems in French, and the current ability of our systems to cope with those. It should finally be noted that, in the following tables, the (indicative) figures refer to occurrences of erroneous words. In fact, for lack of an appropriate distance metric between phonemic symbols, the automatic alignment between the reference text and the systems outputs is only approximate. This makes the errors difficult to analyse at the phoneme level.

Table IV presents the distribution of errors by word category for six systems. This table mainly illustrates the difficulties of correctly pronouncing proper names and acronyms. While these tokens only represent, respectively, 5·8% and 0·7% of the words in the test corpus, they are significantly more represented in the erroneous words. Errors on proper names represent between a half and a sixth of the total number of errors.

Numbers and numerals are another significant cause of errors, which can be further subclassified into three main categories:

TABLE V. Distribution of errors w.r.t. type for six systems

	B	C	D	F	G	H
Loan word	77 (4·1)	52 (4·2)	129 (15·1)	97 (13·0)	89 (13·5)	103 (38·9)
Liaison	111 (5·9)	123 (10·0)	76 (8·9)	49 (6·6)	38 (5·8)	15 (5·7)
Schwa	493 (26·4)	374 (30·3)	134 (15·7)	46 (6·2)	91 (13·9)	3 (1·1)
Heterophonous homograph	98 (5·2)	47 (3·8)	34 (4·0)	90 (12·1)	7 (1·1)	19 (7·2)
Typing errors	116 (6·2)	78 (6·3)	14 (1·6)	15 (2·01)	38 (5·7)	15 (5·7)
Pre-processing	537 (28·8)	187 (15·1)	79 (9·3)	19 (2·6	26 (4·3)	53 (20·0)
Morphological ambiguity	197 (10·6)	1 (0·1)	15 (1·8)	8 (1·1)	2 (0·3)	6 (2·3)
Other	238 (12·7)	373 (30·2)	371 (43·5)	420 (56·4)	366 (55·7)	51 (19·2)

For a given system and error type, each cell of this table contains the absolute number of errors for this type, and the corresponding percentage of the errors for the system.

- deletion of the final consonant of *cinq (5)*, *six (6)*, *huit (8)*, *dix (10)* before a consonant, within dates, addresses or phone numbers;
- insertion of a segment corresponding to forbidden liaisons;
- substitution, especially due to the pronunciation *un* (instead of *une*) for *1* before a feminine noun.

4.2.1. Breakdown by error type

Table V displays the distribution of errors by error type. This table may be analysed by line or by column. First, the results are quite different from one system to another: the best system is not necessarily the best for coping with the various kinds of difficulties. On the whole, it appears that the major sources of errors are foreign words (common and proper names), liaisons, and mute-e. The latter point needs to be moderated though: *e* is the most common grapheme in French, and the transcribed corpus contains 8500 schwas, a small proportion of which (approximately 2% on average) is in fact erroneously predicted.

5. Discussion and conclusion

The methodology developed for GP conversion evaluation proved to be effective. However, it is certainly not perfect. In this concluding section, we critically review the methodology we used, and point out possible improvements. This part is not particularly language dependent: the same types of methodological problems would certainly be encountered in other languages as well. The results obtained are then summarized: we think it gives an accurate picture of the state-of-the-art in the domain of GP conversion for French. The main types of problems that are still remaining are also listed. Finally, a conclusion is given.

5.1. Evaluation of the methodology

Several arguments and questions can be put forward to justify the need for more sophisticated evaluation methodologies. Some of them are:

(1) What is a reference transcription? It is difficult to define something like a reference transcription, for the ideal situation where every one would do exactly the same phonetic distinction is far from reality (Martinet, 1960). At the word level, the definition of a normative pronunciation is nonetheless possible, and most dictionaries consider that there exists one and only one acceptable pronunciation for each word of the language. However, defining a unique normative pronunciation at the sentence level seems much more controversial, which raises a reference problem.

In fact, a given input may well have much more than one single acceptable pronunciation, reflecting the fact that the production of speech is subject to variations along multiple dimensions: phonological, but also rhythmic, stylistic, sociolectal, idiolectal, etc. Independent of these factors, the problem of the definition of a reference transcription is, in our case, aggravated by the fact that the competing systems have been designed in different countries, and are thus very likely to disagree on many aspects of the phonology of French, reflecting the fact that there exists several dialects of French. The design of a reference corpus and the proper definition of transcription errors should ideally address the issue of variability with greater care.

(2) GP conversion and speech synthesis: Most GP conversion devices have been designed with one ultimate goal, that is, the production of a speech signal. Consequently, their phonological output has generally been tailored for the explicit needs of one specific synthesis device. This makes the direct comparison between different transcription devices difficult, since TTS system A may well encode a given phonological distinction directly at the acoustic level (e.g. in the diphone database), whereas it is distinguished in the phonological representations produced by system B. In other words, transcription systems may interact with other modules, which is clearly one of the limits of our modular approach of assessment. Additionally, the evaluation ought to take into account the fact that the possible disagreements with the reference transcription do not impair the intelligibility of the output speech in equal measure. This obviously raises the issue of the objective measurement of the severeness of an error.

(3) Evaluation of modular systems: For French, as for many other languages, the transcription of running texts requires that a series of linguistic analyses be performed. If an evaluation is carried out for diagnostic purposes, it has to be designed so as to provide system developers with results that enable them to identify the most deficient part of the system precisely. As a consequence, the evaluation procedure has to include means of classifying the discrepancies into categories which are meaningful for the system developer. However, this linguistic pre-processing is, depending on the system, either performed by specific modules, or directly integrated into the transcription device. For instance, the pre-processing of abbreviations can be implemented either as a two-stage procedure, where the abbreviation is first expanded, then pronounced, or alternatively, as a direct look-up in an abbreviation dictionary during the transcription phase. This makes the definition of a universal grid for classifying errors a non-trivial problem.

(4) Phonemic alphabet: The definition of the phonemic alphabet for evaluation proved to be difficult and controversial. Two work-arounds for this problem can be envisaged. The first solution is the solution we took in the experiments: a phonemic alphabet with a minimum number of symbols that reflect the intersection between the existing alphabets.

Another, maybe better solution, would be to take as the common symbol set the union of all existing alphabets. This would allow each participant to transcribe the reference text with their own inventory, and make the comparison between different transcriptions only a matter of transducing each symbol set into a common notational scheme. In fact, an additional potential requirement put on that common alphabet could be its ability to encode phonological variants. As a test corpus has to contain multiple phonological variants, one could try to encode as much as possible of this variation directly into the alphabet.

One possible solution to the problem of describing the variability is to consider phonetic transcriptions not as a succession of atomic symbols, but as a succession of more complex units. Let us see how this would work in the case of liaisons. A liaison can be compulsory, forbidden, or optional. Using abstract symbols, it is possible to represent optional liaisons in /z/ with the capital letter /Z/, which covers /z/ and zero in its realization (a solution reminiscent of the concept of archiphonemes). As a consequence, the two variants need not be explicity listed in the test corpus. Likewise, using such equivalence classes, we could represent the variation between equally acceptable realizations of:

- consonants which are optionally deleted (e.g. the final *s* of *jadis*);
- variations between diæresis and synæresis (e.g. the *u* in *tuer*);
- variations of the intermediate timbre vowels, in an unstressed position or when there is hesitation or neutralization (e.g. the first *o* of *microcosme*);
- schwas which are optionally deleted or realized, as in *petit*.

The definition and use of this kind of alphabet have two main advantages. The first one is a very practical one: encoding (part of) the variability directly in the alphabet facilitates the transcription of the reference corpus, and makes the resulting corpus smaller. Additionally, the use of this kind of alphabet allows us to evaluate not only the accuracy of a given GP conversion device, but also its *precision*. Let us assume, for instance, that the reference pronunciation of the first *o* in *microcosme* can be either /o/ or /ɔ/. In this situation, a system capable of predicting both alternatives should be more rewarded than a system which only predicts one of the two possible outcomes, since arguably, the former is more precise than the latter.

However, this kind of encoding could only capture some cases of phonological variability, namely, the cases where the pronunciation of one single symbol is subject to variation, irrespective of its phonological context. More complex cases, where the alternative exists between several sequences of phonemes, or where there is a contextual dependency between adjacent symbols, would still need to be explicitly enumerated. Such cases are not uncommon: for instance, the final part of the proper name *Bernstein* can be pronounced [stɛn], [ʃtajn] or [stɛ̃], and this cannot be properly described using complex atomic units. Likewise, [mɛtsɛ̃] is an acceptable transcription of *médecin*, but the opening of the initial /e/ and the devoicing of /d/ are only possible if the schwa is properly deleted. In addition, it is possible that the evaluation of the precision of a GP conversion device is not central in the specific context of speech synthesis systems, and that its undertaking would render the scoring measure unduly complex.

(5) Variants and coherence: Another point is worth mentioning, which concerns a dimension along which systems were not specifically evaluated. The reference corpus contains a lot of independently specified variants. A consequence is that a system which would, whenever variation is allowed, randomly select one of the possible phonemes, would be considered as accurate as a system which sticks to a coherent elocution strategy (such as always realizing optional liaisons, always deleting optional schwas . . .). In other words, this scoring strategy fails to assess the stylistic coherence of a given transcription. This has been found to be quite a minor problem: all the tested systems use deterministic transcription rules, which makes them unable to produce this kind of "incoherent" output.

(6) Enriched corpus: Annotating errors according to several evaluation dimensions could, to a large extent, be performed automatically, thus greatly reducing the burden of the system developer. This would however necessitate enriching the test corpus annotation scheme, including for instance part-of-speech tags and an alignment between orthographic and phonemic strings at the world level, and a finer phonemic alphabet, where for instance liaison phonemes would be identified (and even subclassified between optional and compulsory liaisons), and where specific symbols would mark the places where liaison is not possible.

Several test corpora could also be generated with the same text material. In this way, systems' behaviour could also be tested by using different variants, each representing a different level of difficulty for the GP conversion task. Going from the most complex to the easiest, the first variant would be pre-segmented, the next would be free of typing errors, the next would further contain expanded abbreviations, an even easier variant would include morphosyntactic tags (or even brackets), etc. Given the availability of accurate natural language processing tools for performing these tasks, the production of these variants could be done nearly automatically.

(7) Corpus transcription: An alternative exists regarding the constitution of a reference transcription for a text corpus: ear-transcribed or hand-transcribed, which means observed or not. A comparison of reference dictionaries (Juillard, 1965; Warnant, 1987; Larousse, 1989; Robert, 1990; Boë & Tubach, 1992) reveals the obvious difference of phonemic prediction by the authors. More generally, it seems that the problem of finding an agreement between different listeners makes the constitution of ear-transcribed corpora difficult. It thus appeared more reasonable, given the resources that were allocated to the corpus building group, to go for a hand-transcribed corpus in these experiments. This choice was also plainly justified with respect to the participating GP systems themselves, whose pronunciation rules reflected more the content of pronunciation dictionaries than the actual pronunciations.

Lexica vs. running texts: Common word lexica are very valuable tools for measuring the capabilities of a transcription system to handle phenomena internal to lexemes or derivational morphemes. Moreover, their coverage enables us to test the transcription capabilities on a large scale, i.e. to evaluate overall consistency of the transcription rules. Finally, specialized lexica, such as proper names lexica, constitute suitable testing conditions for evaluating GP conversion devices on a task like reverse directory inquiry. However, the only publicly available electronic dictionary (BDLEX, (Pérennou *et al*., 1991)) was not directly suited to our needs. The only way to evaluate our systems on a lexicon would have been to develop a large coverage and/or specialized test lexicon, which was not possible, given the limited amount of resources available within the frame of this project. This kind of evaluation still remains to be performed.

Testing GP systems on running texts has the disadvantage that a text only covers a small number of possible phonetic forms and can therefore only test the validity of a small subpart of the transcription rules. Moreover, as running texts integrate the frequency usage of words, their use in an evaluation penalizes systems which pronounce incorrectly very frequent words. However, in comparison to a lexicon, they allow us to evaluate the ability of a GP conversion device to realize properly contextual variants (final schwas, liaisons, etc). In addition, texts are publicly available, and reflect quite well the task to be performed by GP conversion systems for TTS. As stated earlier, another advantage is that this kind of corpus contains a full assortment of the typical difficulties of GP conversion: complicated structures, citations, dates, proper names, acronyms and typing errors.

5.2. State of the art

The evaluation results presented in this paper give a fairly accurate picture of the state-of-the-art in GP conversion for French. The systems made between 0·5% and 7% of errors at the phonemic level. On average, between one and eight out of every 10 sentences contain a possible synthesis error due to GP conversion. This indicates that the problem of GP conversion is still important for practical applications in French, since even the best systems are likely to make an error every 10 sentences.

Most of the GP conversion errors can be attributed to few sources. Difficulties are due to proper names, heterophonous homographs, pre-processing, schwa and liaison. It is interesting to review all these sources of errors, because all systems did not perform equally well on all problems.

Loan words: This problem causes many errors for all the systems. It is a difficult problem which may involve specific rules, together with some knowledge of the linguistic origin of the word.

Proper names: Pronunciation of proper names is difficult to deal with. For some applications, the correct pronunciation of proper names is of the utmost interest (e.g. reverse directory applications). Even so, the best system for proper names still mispronounces one out of 10 names, leaving some room for improvement.

Acronyms, numbers and symbols: This is an important source of errors for all the systems. The best system regarding this matter is still making many errors. Moreover, correct pronunciation of this type of item may be crucial for understanding.

Heterophonous homographs: This is a classical problem in GP conversion. The best system for this task made only very few errors (7 words in error in this corpus). Particular attention is paid to the part of speech assignment module of this system. In this case, it seems that almost all the ambiguities can be solved. Thus, this problem seems to be more related to system design rather than being a fundamental problem.

Pre-processing: Depending on the task, pre-processing can be more or less difficult. In the newspaper reading experiments reported here, few pre-processing errors remain for the best system (19 words in error). However pre-processing may well be a very difficult problem that needs to be tuned to the specific speech synthesis application under study.

Schwa: For many systems, the schwa is an important source of errors. Surprisingly, the best system for this particular problem made only very few errors (3 words in error). This may indicate that a good set of rules may be able to handle this problem rather well. More probably, the evaluation protocol was too tolerant regarding the transcriptions of variants for this phenomenon, and tends to under estimate the real difficulty of the problem.

Liaison: This is also a classical problem for GP conversion in French. For the best system regarding this matter, only a few liaison errors remain. Again, this may indicate that a good set of rules may be able to handle this problem rather well when the reference encodes a standard realization of liaisons; or more probably, that the test design partly masked the complexity of the problem. In fact, realization of schwa and liaison are linked since they are subject to variation according to the same set of stylistic parameters (speaking rate, language register, etc). For lack of a complete understanding and precise description of the influence of these parameters, a more appropriate evaluation of the accuracy of GP conversion systems may be, for these specific problems, to use ear-transcribed corpora.

Other: Other sources of error vary among systems. It is interesting to point out that for the best system, the number of "other types" of errors account only for about half the number of proper name errors for the best systems on proper names.

Finally, imagine that a system could merge all the best features of the different systems (e.g. the best models for proper names and loan words, the best linguistic analyser, the best set of rules, etc.). One can guess that this system would have achieved 20% better results compared to the best system. On the one hand, this means that none of the systems tested in this project are optimal according to the state-of-the-art of all systems. On the other hand, this indicates that some problems are still difficult for all the systems.

5.3. Conclusion

One of the aims of this project is to make available corpora and evaluation paradigms that may be re-used in future work. This will enable a quantitative analysis of the results obtained, and a measurement of the progress achieved for each specific system. For this goal, a corpus was defined, based on newspaper texts. A working methodology was designed, and tests were performed, using many systems.

The discussion on methodology raised serious theoretical objections against the concept of an objective evaluation of GP conversion devices. However, in the specific context of this test campaign, we could come up with practical solutions to most of these problems. In fact, the eight systems taking part in this evaluation were found to be surprisingly similar in their overall architecture, which certainly made the agreement on an acceptable evaluation methodology much easier to reach.

This work is, to the authors' knowledge, the first large-scale joint evaluation of GP conversion for speech synthesis in French. It demonstrates that it is possible to do the task, and to establish the state-of-the-art in this domain. However, it must be emphasized that many methodological problems are still difficult to handle. We hope that this aspect of the paper, namely the discussion on pros and cons of the methodology, will help future work aiming at similar goals.

The experiments reported on in this paper were conducted in the framework of the ILOR ARC-B3 project supported by the Francil network. This project was funded by the AUPELF-UREF (Association of French-Speaking Universities). We are grateful to the French newspaper *Le Monde* who kindly provided us with electronic text material. Special thanks to Pascale Bouchet and Roseline Hamamdjan for their work on the corpus transcription.

References

d'Alessandro, C., Aubergé, V., Béchet, F., Boula de Mareüil, P., Foukia, S., Goldman J.-P., Keller, E., Marchal, A., Mertens, P., Pagel, V., O'Shaugnessy, D., Richard, G., Talon, M.-H., Wehrli, E. & Yvon, F. (1997). Vers l'évaluation de systèmes de synthèse de parole à partir du texte en français. In *Actes des Premières Journées Scientifiques et Techniques du réseau Francil de l'AUPELF-UREF* Avignon, pp. 393–397.

Aubergé, V. (1991). La synthèse de la parole: des règles aux lexiques. Doctoral Dissertation. Université P. Mendès-France, Grenoble.

Béchet, F. & El-Bèze, M. (1996). Intégration de différents niveaux linguistiques pour le traitement des mots hors-dictionnaire dans la conversion graphème-phonème automatique. In *Actes des XXIes Journées d'Études sur la Parole* Avignon, pp. 421–442.

Belrhali, R. (1995). Phonétisation automatique d'un lexique général du français: systématique et émergence linguistique. Doctoral Dissertation. Université Stendahl, Grenoble.

Benoit, C. (1997). Evaluation inside or assessment outside?. In *Progress in Speech Synthesis* (J. P. H. van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg, eds) Springer-Verlag, New York, pp. 513–517.

Boë, L.-J. & Tubach, J.-P. (1992). De A à Zut. *Dictionnaire phonétique du français parlé*. ELLUG, Grenoble.

Boula de Mareüil, P. (1997). Conversion graphème-phonème: de la formalisation à l'évaluation. In *Actes des Premières Journées Scientifiques et Techniques du réseau Francil de l'AUPELF-UREF* Avignon, pp. 399–406.

Carlson, R., Granström, B. & Nord, I. (1990). Segmental evaluation using the Esprit/SAM test procedures and mono-syllablic words. In *Talking Machines* (G. Bailly & C. Benont eds) Elsevier, North Holland, pp. 443–453.

Catach, N. (1984). *La phonétisation automatique du Français*. Editions du CNRS, Paris.

Cotto, D. (1992). Traitement automatique des textes en vue de la synthèse vocale. Doctoral Dissertation. Université Paul Sabatier, Toulouse.

Damper, R. I. (1995). Self-learning and connectionist approaches to text-phoneme conversion. In *Connectionist Models of Memory and Language* (J. Levy, D. Bairaktaris, J. Bullinaria & J. Cairns eds) UCL Press, London, pp. 117–144.

Divay, M. (1990). A written text processing expert for text to phoneme transcription. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP), Volume 2*, Kobe, pp. 853–856.

Dutoit, T. (1993). High-quality text-to-speech synthesis of the French language. Doctoral Dissertation. Faculté Polytechnique de Mons.

Gibbon, D., Moore, R. & Winski R. (eds) (1997). *Handbook of Standards and Resources for Spoken Language Systeme*. Mouton de Gruyter, Berlin.

Gaudinat, A. & Wherli E. (1997). Analyse syntaxique et synthèse de la parole: le project FIPSVox. *Traitement Automatique des Langues* **38**(1), 121–134.

Imbs, P. (1971). Trésor de la Langue Française. *Dictionnaire de la langue du XIXe et du XXe siècles (1889–1900)*. Editions du CNRS, Paris.

Juilland, A. (1965). *Dictionnaire Inverse de la Langue Française*. Mouton & Co., London–The Hague–Paris.

Keller, E. (1997). Simplification of TTS architecture vs. operational quality. In *Proceedings of Eurospeech* ESCA, Rhodes, pp. 585–588.

Klaus, H., Fellbaum, K. & Sotscheck, J. (1997). Auditive Bestimmung und Vergleich der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache. *Acta Acustica* **83**, 124–136.

Kraft, W. & Portele, T. (1995). Quality evaluation of five German speech synthesissystems. *Acta Acustica* **3**(4), 351–365.

Lacheret-Dujour, A. (1990). Contribution à une analyse de la variabilité phonologique pour le traitement automatique de la parole continue multi-locuteur. Doctoral Dissertation, Université D. Diderot, Paris.

Laporte, E. (1988). Methodes algorithmiques et lexicales de phonétisation de textes. Doctoral Dissertation, Université D. Diderot, Paris.

Larousse, P. (1989). *Dictionnaire de la Langue Française Lexis*. Larousse, Paris.

Larreur, D. & Sorin, C. (1990). Quality evaluation of French text-to-speech synthesis within a task, the importance of the mute 'e'. In *Proceedings of the ESCA Workshop on Speech Synthesis* ESCA, Avignon, pp. 91–96.

Laver, J., McAllister, M. & McAllister, J. (1989). Pre-processing of anomalous text-strings in an automatic text-to-speech system. In *Studies in the Pronunciation of English. A Commemorative Volume in the Memory of A. C. Gimson* (S. Ramsaran ed.). Croom Helm, London.

Martinet, A. (1960). *Éléments de Lingistique Générale*. Armand Colin, Paris.

Marty, F. (1992). Trois systèmes informatiques de transcription phonétique et graphémique. *Le français moderne LX.2*, 179–197.

O'Shaughnessy, D. (1984). Design of a real-time French text-to-speech system. *Speech Communication* **3**(4), 317–324.

Paroubek, P, Adda, G., Mariani, J. & Rajman, M. (1997). Les procédures de mesure automatique de l'action GRACE des assignateurs de parties du discours pour le français. In *Actes des Premières Journées Scientifiques et Techniques du réseau de l'AUPELF-UREF* Avignon, pp. 245–252.

Pérennou, G., Cotte, D., de Calmès, M., Ferrani, I., Pécatte, J-M. & Tihoni, J. (1991). Composantes phonologique et orthographique de BDLEX. In *Actes des Journées du GRECO-PRC CHM*. EC2 Editeur, Toulouse.

Pols, L. C. W. & Jekosch, U. (1997). A structural way of looking at the performance of text-to-speech systems. In *Progress in Speech Synthesis* (J. P. H. van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg eds) Springer Verlag, New York, pp. 519–527.

Robert, P. (1990). *Dictionnaire Alphabétique & Analogique de la Langue Française*. Société du Noveau Litré, Paris.

Silverman, K., Basson, S. & Levas, S. (1990). Evaluating synthesiser performance: is segmental intelligibility enough? In *Proceedings of the International Conference on Speech and Language Processing (ICSLP)* **90**, Volume 2, pp. 981–984. Kobe.

Sorin, C. & Emerard, F. (1996). Domaines d'application et évaluation de la synthèse de parole à partir du teste. In *Fondements et Perspectives on Traitement Automatique de la Parole* (H. Méloni ed.) AUPELF-UREF pp. 123–131.

van Bezooijen, R. & Pols, L. C. W. (1990). Evaluation of text-to-speech systems: some methodological aspects. *Speech Communication* **9**, 263–279.

van Santen, J. F. H. (1993). Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language* **7**, 49–100.

Warnant, L. (1987). *Dictionnaire de la Pronunciation Française dans sa Norme Actuelle*. Éditions Duculot, Paris-Gembloux.

Yvon, F. (1996). Prononcer par analogie: motivations, formalisation et évaluation. Doctoral Dissertation, ENST, Paris.