# Evaluating the pronunciation of proper names by four French grapheme-to-phoneme converters

*Philippe Boula de Mareüil*[1], *Christophe d'Alessandro*[1], *Gérard Bailly*[2], *Frédéric Béchet*[3], *Marie-Neige Garcia*[4], *Michel Morel*[5], *Romain Prudon*[4], *Jean Véronis*[6]

[1] LIMSI-CNRS, Orsay, France; [2] ICP, Grenoble, France; [3] LIA, Avignon, France; [4] ELDA, Paris, France; [5] CRISCO, Caen, France; [6] DELIC, Aix-en-Provence, France
mareuil@limsi.fr; evasy@elda.org

## Abstract

This article reports on the results of a cooperative evaluation of grapheme-to-phoneme (GP) conversion for proper names in French. This work was carried out within the framework of a general evaluation campaign of various speech and language processing devices, including text-to-speech synthesis. The corpus and the methodology are described. The results of 4 systems are analysed: with 12-20% word error rates on a list of 8,000 proper names, they give a fairly accurate picture of the progress achieved, the state-of-the-art and the problems still to be solved, in the domain of GP conversion in French. In addition, the resources and collected data will be made available to the scientific and industrial community, in order to be re-used in future bench-marks.

## 1. Introduction

This paper presents the first results of the EVALDA/EvaSy project dedicated to the evaluation of speech synthesis systems for the French language. Organised by the European agency ELDA, this evaluation campaign is intended to expand upon the AUPELF (now AUF) campaign of 1996-1999, the only previous evaluation campaign for text-to-speech (TTS) systems for the French language [1]. The EvaSy evaluation campaign is subdivided into three components: grapheme-to-phoneme (GP) conversion, prosody and global quality of the synthesised speech. The issue of this paper is the evaluation of the GP module of four systems: those of CRISCO, ICP, LIA and LIMSI. References to those systems can be found in [2]: they all rely on rule-based approaches, possibly completed by lexicon look-up (up to thousands of entries), except one system (referred to as the Lab4 system to keep the results anonymous), which implements the ID3 algorithm as [3]. A fifth laboratory, DELIC, was in charge of the corpus production.

The overall quality of a TTS system depends upon the voice used, on the prosody generated, but also on GP conversion. It was shown that the majority of GP errors, for the best operational systems, stem from proper names [1]. Within the framework of the joint AUPELF evaluation campaign, during which substantial resources were provided, the scores of 99.7% correct phonemes (99.1% correct words) obtained on newspaper texts let us think that a reference transcription of such corpora is very costly, to finally and laboriously bring to light few errors. In an evaluation task limited to a list of proper names, we can expect quite different rates (80-90%). That is the reason why, without excluding the importance of other aspects, this article concentrates on proper names.

Proper names, within which very different orthographic and phonetic systems coexist, raise a ticklish albeit crucial problem for TTS synthesis and automatic speech recognition (ASR), because their pronunciation strongly depends on their origin and usage. In foreign proper names especially, a conflict appears between respecting the original spelling and approximating the original pronunciation by following the French conventions. Orthographic idiosyncrasies are striking, reflecting the mono-referential character of proper names (denoting a unique entity) [4]. As a matter of fact, some people make it a point of honour to keep an original pronunciation of their name, whereas others, maybe for the sake of integration or assimilation, opt for a pronunciation of their name more in adequacy with the habits of the country. Our geographical competence or linguistic knowledge (for instance in foreign languages) can also have an influence. If a text deals with a person or a town that is assumed to be German (typically a celebrity such as Berger) will most probably be pronounced in a way that violates the basic rules of French pronunciation (namely [bERgER] instead of [bERZe]).

This linguistic mosaic, resulting from the usage diversity, is illustrated in large databases of proper names, for 11 European languages [5]. To phonetise them automatically, several solutions have been proposed since then: e.g. rule-based expert systems [6,7] and machine-learning models by analogy [8,9,3,10]. These types of techniques, which are both represented within our laboratories, deserve to be assessed and compared.

For future development and diagnosis purposes, a database of proper names labelled with linguistic origins is of particular interest. With TTS [11,12] or ASR [13] in view, origins such as Dutch, English, German, Italian, Spanish, Polish, Arabic, Japanese were defined to consequently determine the pronunciation of proper names. This approach, which can be widened to language groups or families, calls for further investigation, at least to go beyond mere scores obtained by systems taken as black boxes, and to detail their performance. To us, it is more interesting to know whether machine-learning or rule-based systems stumble over, say, English names but extricate themselves from, say, Italian names than to know which is the best system.

In the following, the method is described, including the corpus selection, its manual phonetic transcription and annotation with linguistic labels, as well as the task assigned to the participants. Results are then presented and discussed. Various categorisations of the outcomes enabled us to conduct a detailed quantitative analysis of the systems' mistakes.

## 2. Method

### 2.1. Corpus design and phonetic transcription

Since it is difficult to define what a proper name is (in the case of trade marks, product and company names especially), we restricted ourselves to person names. A list of 4,115 first name–surname pairs was extracted from the French newspaper *Le Monde* from 1992-2000 (over 200 million words). This sample was obtained by considering pairs of capitalised words which appear between 100 and 200 times in the corpus. This range was kept, because more frequent names would risk to have been foreseen in the different systems, and randomly selected names would have resulted in many typos and hapaxes, in accordance with Zipf's law. The retained proper names are thus of an average difficulty. First, word pairs beginning with a capital letter other than sentence-initially were automatically extracted. Then, capitalised French common words, brand or company names, abbreviations and other mismatches were filtered out, which resulted in discarding about 25% garbage such as Premier Ministre or Air France. This work was done at DELIC.

The selected material was then hand-transcribed in the phonetic alphabet SAMPA for the French language [14], with variants as illustrated in the following examples ({@/}, for instance, designates an optional schwa):

- Kissinger kisin{dZ/g}{E/9}R
- Griotteray gRi{j/}{O/o}t{@/}R{E/e}

This was done in two steps: a first transcription was produced, which was then checked by a second expert. The experts were provided with transcription guidelines bearing on the schwa, the {e/E} and {o/O} oppositions, nasal vowels, glides and optional gemination especially. It was also advised to equate the Spanish *jota* ([x]) to [R], and the English interdentals [T] and [D] to {s/t} and {d/z} respectively. Other pronunciations are subject to variation [15]. The possible overgeneration and inconsistency of variants were extensively discussed. It was felt not to be too serious a problem, because the assessed GP conversion systems are deterministic (one single pronunciation is foreseen). And undertaking to capture the context-dependency between adjacent sequences of phonemic symbols would have rendered the transcription scheme unduly complex.

Additionally, the transcribers had access to 10 excerpts in which any first name-surname pair appeared, with 100 words to the left and to the right. Also, they could launch a Google query for the names under consideration by simply clicking on a hyperlink. Their situation was therefore close to that of a radio journalist confronted with proper names he/she has to pronounce. In this way too, our database is more than a mere word list.

### 2.2. Linguistic annotation

Our list was enriched with linguistic origin indications concerning the surnames. For this purpose, a set of 20 linguistic labels was defined, exhibiting common behaviours with regards to the strategy we appeal to, so as to pronounce proper names. The geographical competence of French people, linked to their naive linguistic knowledge, was taken into account. For instance, a French speaker should know how to recognise Spanish-like or Italian-like names, which belong to his/her neighbourhood. In return, it is not always easy, if it is feasible, to distinguish between Russian, Ukrainian and Bulgarian names, or between German, Yiddish and Dutch names. Likewise, it is very unlikely that a French speaker can distinguish between the languages of the Niger-Congo family. Moreover, these languages may share a common phonetic writing; and, as far as the pronunciation of proper names by a French speaker is concerned, they may be processed in the same way, irrespective of their branch. For instance, 'e' and 'u' will respectively be uttered [e] and [u], instead of [@] and [y].

In order to provide a proper name with a linguistic label, too, advantage can be taken of the context: besides the given name, for example, the sentence in which a family name appears may give some information about the person's nationality. The latter indication can be useful in some cases, even though it does not necessarily go on a par with a linguistic origin. For example, former Latin-American heads of state such as Fujimori, Pinochet and Stroesner are notoriously and respectively of Japanese, French and German origin. Anyway, this annotation, which remains open and tolerant, is a matter of trade-off and common sense. Hence the list of linguistic labels reported in Table 1 (for the most represented labels in the corpus): they are inspired by ISO language codes. Genetically unrelated languages may be regrouped (e.g. Albanian and Turkish, Korean and Chinese), if we are anyhow unable to distinguish them. They may also adopt common conventions. Owing to the large number of English names and their phonetic specificity, we found it necessary to distinguish English from other Germanic languages.

*Table 1*: Linguistic labels with the proportion of the corpus they represent.

| Label | Meaning | % | Label | Meaning | % |
|---|---|---|---|---|---|
| fre | French | 51 | ind | Indian | 1 |
| eng | English | 15 | chi | Chinese | 1 |
| ger | Germanic | 10 | tur | Turkish | 1 |
| ita | Italian | 5 | heb | Hebrew | 1 |
| sla | Slavic | 4 | prt | Portuguese | 1 |
| spa | Spanish | 3 | jpn | Japanese | 1 |
| ara | Arabic | 3 | | | |
| afr | African | 2 | | Other | 1 |

The risk of being politically incorrect exists, we are aware of it, as soon as we speak about word origins. This is the very discussion about loan words, which travel and are more or less integrated (see the many etymological dictionaries and books devoted to proper names such as [16], which presents over 9,000 names of famous or contemporary personalities classified by origin, with their pronunciations). When experts transcribe proper names, they apply strategies (possibly with the help of the context, which can be exploited in delicate or ambiguous cases), which often consist of hypothesisingg origins and consequently performing phonological transfers. They were therefore asked to explicit these origins though linguistic labels. As demonstrated by a preliminary stage, this task is not necessarily harder nor more time-consuming than

phonetic transcription itself, with all its possible variants concerning the schwa or mid vowels especially. It is even more straightforward to detect the origin of names such as Chavez, Angelopoulos, Browning or Ruggero than to transcribe them.

### 2.3. Participants' task

The participants had to adapt their systems in order to output transcriptions in SAMPA. After a preliminary test, the objective of which was to discard formatting problems, the test took place during the winter 2004-2005. For each participant, the task consisted of phonetising the list of proper names within 3 hours. Once results were computed, 3 weeks of adjudication were then foreseen, to give the participating laboratories the opportunity to contest some of their errors. Errors counted for this or that system were discussed, and the reference was accordingly corrected or enriched with additional variants, to release a new version. After each phase, an alignment round was performed, between the phonemic outputs and the reference. The scoring is based on the `sclite` dynamic programming algorithm (http://ww.nist.gov/speech/tools/).

## 3. Results

### 3.1. Overall results

The adjudication phase led to correcting or adding variants to about 200 names out of 8,230, and did not change the systems' ranking. The results obtained at the term of this stage should not be considered to the nearest error. The measured figures are only indicative: they solely reflect the relative importance of certain problems and the current ability of our systems to cope with them.

Table 2 displays the raw results, after a segmentation into first names and surnames: by and large, the systems did not fare with the task as well as they did with running texts, since they achieve at least 12% word error rates. However, the 12-20% error rates on proper names are comparable to the ones obtained on the 1,500 proper names of the AUPELF text corpus [1]; they are slightly better than the ones reported by [12] for the English language. None of the rule-based systems have been superseded by the Lab4 data-driven approach. Nevertheless, it is noteworthy that the self-learning system had been trained on more native French words and that its development is not as time-consuming as is rule-management. Interestingly, the best system (the Lab1 system) is the same as in the AUPELF campaign. But contrary to the latter, we do not give percentages on phonemes here, because the higher overall error rates raise more alignment problems.

*Table 2*: Overall error rates on first names and surnames.

| %Error | Lab1 | Lab2 | Lab3 | Lab4 |
|---|---|---|---|---|
| First names | 8.4 | 10.5 | 12.7 | 13.6 |
| Surnames | 17.4 | 23.8 | 21.7 | 25.0 |
| Total | 12.9 | 17.1 | 17.2 | 19.3 |

A common trend we can observe across the different systems is that first names are generally better phonetised than are surnames. An explanation is that the participants may have watched over the pronunciation of first names which are more frequent than are surnames. Another explanation is that the reference may be more tolerant on first names than on surnames. Indeed, when an English first name also exists in French, it may be pronounced in the French way: Michael (respectively Thomas), for instance, is more inclined to be pronounced [mikaEl] (respectively [t{O/o}ma]) as a first name than it is as a surname.

### 3.2. Analysis by linguistic label

The linguistic labels were not used by the GP converters that were evaluated here, but as argued above, they are linked with the proper names' pronunciation. They also allow us to sort out the results by origin (see Tables 3 and 4 for the most frequent linguistic labels), and to envisage future techniques which would derive benefit from this information.

*Table 3*: Error rates on surnames for the most frequent linguistic labels (%Error/Label).

| %Error | Lab1 | Lab2 | Lab3 | Lab4 |
|---|---|---|---|---|
| fre | 5 | 9 | 13 | 10 |
| eng | 32 | 43 | 36 | 46 |
| ger | 36 | 52 | 44 | 48 |
| ita | 18 | 22 | 21 | 23 |
| sla | 27 | 34 | 17 | 44 |
| spa | 30 | 41 | 22 | 36 |
| ara | 21 | 20 | 11 | 24 |
| afr | 29 | 34 | 24 | 34 |

A certain hierarchy is respected, between the lines and the columns of Table 3. On the whole, French names turn out to be the best transcribed names, English and other Germanic names the worst transcribed ones. After these extreme cases, we have Spanish names (rather poorly transcribed) and Italian names (rather accurately transcribed). Such a difference, which was unexpected between Romance languages, is important to note, and justifies the Spanish/Italian distinction a posteriori.

If Tables 2 and 3 yield error rates for a given type of proper name, it is also of interest to pinpoint the most problematic cases and their distributions, as in Tables 4 and 5.

*Table 4*: Percentage of errors on surnames broken down by linguistic label (%Label/Error).

| %Label | Lab1 | Lab2 | Lab3 | Lab4 |
|---|---|---|---|---|
| fre | 16 | 20 | 32 | 21 |
| eng | 28 | 27 | 25 | 28 |
| ger | 29 | 22 | 21 | 20 |
| ita | 5 | 4 | 5 | 4 |
| sla | 7 | 7 | 3 | 8 |
| spa | 5 | 5 | 3 | 4 |
| ara | 4 | 3 | 2 | 3 |
| afr | 4 | 3 | 2 | 3 |

For all the systems, the percentages are higher for French names in Table 4, with respect to Table 3. This is easily understandable if we look at Table 1, which shows that

French names cover the majority of the corpus. Inversely, with respect to Table 3, percentages are lower for English and other Germanic names in Table 4. But these names represent the major source of error for all the systems. In comparison, Slavic names account for few errors: too few to draw conclusions, even though detailed inspection of Tables 3 and 4 reveals very different behaviours for these names, particularly between Lab3 and Lab4.

### 3.3. Analysis by grapheme mispronunciation

A classification of errors by grapheme mispronunciation is also needed. Scripts were written to handle the systems' failures with this end in view. Three types of problem, in particular, happened to account for a large number of errors: the vowel 'e', which is dropped or pronounced as a schwa; the digrams 'an', 'en', 'in', 'on' and 'un' which are improperly nasalised; the final consonants *-d, -g, -r, -s, -t, -x, -z* which are not pronounced. The case of the *–er* termination is particular, inasmuch as this character string often results in [e] instead of [ER] or [9R] (e.g. in Schwarzenegger).

*Table 5*: Percentage of errors broken down by grapheme mispronunciation (%Grapheme/Error) — 'e' stands for an e-{@/} substitution/deletion; 'Vn' means an erroneous nasalisetion of the digrams 'an', 'en', 'in', 'on' and 'un'; C designates the deletion of the consonants *–d, -g, -r, -s, -t, -x, -z.*

| %Grapheme | Lab1 | Lab2 | Lab3 | Lab4 |
|---|---|---|---|---|
| 'e' | 16.7 | 12.5 | 5.6 | 9.8 |
| 'Vn' | 19.5 | 22.0 | 10.7 | 16.4 |
| Final C | 23.6 | 17,3 | 8.6 | 12.0 |

In the majority of cases, 'e'-related errors correspond to [e] deletions (e.g. Corea) rather than schwa realisations (e.g. Boccanegra). They are fewer for Lab3, which in return inserts many spurious schwas (18.2% of all errors). Among the 'Vn' errors reported in Table 5, the most frequent configurations are the pronunciation [a~] instead of [an] and [e~] instead of [in], partly stemming from first names like Juan or Martin (when the surname is English). The latter name (pronounced [maRte~] in French) is a good example of the context-dependency of GP conversion. Among the final consonants which are most often mute in French, the omission of an [s] is by far the most frequent. There are 951 names terminated by an *-s* or an *-x* in the corpus (e.g. Coencas [k{O/o}Enkas]). But in the majority of cases (e.g. Dumas [dyma]), a final [s] should not be pronounced.

## 4. Conclusion

We presented a corpus and an objective evaluation methodology tuned to GP conversion for proper names in French. This practical and theoretical problem proved important, in particular for English and other Germanic names. Another major contribution of this work is that it was suited to examine error types automatically (e.g. 'e'-related). The resources which enabled us to establish the grid of analysis will be put at the disposal of the scientific community, to serve as a bench-mark for other domains and other languages. The construction of pronunciation dictionaries for ASR and reverse dictionary inquiry would be concerned in the first place. A list of proper names with their pronunciations might also be helpful to learners of French as a foreign language. It would arguably improve by being completed with actual recordings. Finally, the applied nature of this work will not exempt us from carrying out research on the phonology of loan words and proper names.

## 5. Acknowledgements

This work was financed by the French Ministry of Research in the context of the Technolangue programme.

## 6. References

[1] Yvon, F. Boula de Mareüil, P., d'Alessandro, C. *et al.* "Objective evaluation of grapheme-to-phoneme conversion for text-to-speech synthesis in French", *Computer Speech and Language,* 12(4): 393-410, 1998.

[2] d'Alessandro, C. & Tzoukermann, E. (Eds), *Synthèse de la parole à partir du texte, Traitement Automatique des Langues,* 42(1), Hermès, Paris, 2001.

[3] Black, A., Lenzo, K. & Pagel, V., "Issues in building general letter-to-sound rules", *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, 1998 (pp. 77-80).

[4] Gary-Prieur, M.-N., *Grammaire du nom propre*, Presses Universitaires de France, Paris, 1994.

[5] The ONOMASTICA Consortium, "The ONOMASTICA interlanguage pronunciation lexicon", *Eurospeech*, Madrid, 1995 (pp. 829-832).

[6] Divay, M. & Vitale, A.J., "Algorithms for Grapheme-Phoneme Translation for English and French: Applications", *Computational linguistics*, 23(4): 495-524, 1997.

[7] Boula de Mareüil, P., *Étude linguistique appliquée à la synthèse de la parole à partir du texte*, PhD thesis, University of Paris XI, Orsay. 1997.

[8] Yvon, F., *Prononcer par analogie : motivations, formalisation et évaluation*, PhD thesis, ENST, Paris, 1996.

[9] Bagshaw, P., "Phonetic transcription by analogy in text-to-speech synthesis: Novel word pronunciation and lexicon compression", *Computer Speech and Language,* 12(2):119-142, 1998.

[10] Damper, R.I., Stanbridge, C.Z. & Marchand, Y., "A Pronunaciation-by-Analogy Module for the Festival Text-to-Speech Synthesiser", *4th ISCA Workshop on Speech Synthesis*, Pitlochry, 2001 (pp. 97-102).

[11] Béchet, F. & El-Bèze, M., "Automatic assignment of part-of-speech to out-of-vocabulary words for text-to-speech processing", *Eurospeech*, Rhodes, 1997 (pp. 983-986).

[12] Llitjos, A.F. & Black, A.W., "Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names", *Eurospeech*, Aalborg, 2001 (pp. 1919-1923).

[13] Bartkova, K. & Jouvet, D., "Language based phone model combination for ASR adaptation to foreign accent", *ICPhS*, San Francisco, 1999 (pp. 1725-1728).

[14] Gibbon, D., Moore, R. &Winski, R. (Eds), *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, 1997.

[15] Boula de Mareüil, P., Yvon, F. *et al.*, "A French phonetic lexicon with variants for speech and language processing", *LREC*, Athens, 2000 (pp. 273-276).

[16] Maes, P., *La prononciation des langues européennes*, Éditions du centre de formation et de perfectionnement des journalistes, Paris, 1993.