SUBMISSION for Language Resources for European Minority Language Workshop, Granada, Spain, May 1998.

KGB PROJECT:
Tools and Resources for Breton Language Learning

Jacques Siroux, Hervé Gourmelon, Guy Mercier, Jean-Pierre Messager
CORDIAL Team, IRISA/LLI
ENSSAT
6 rue de Kerampont BP 447
F-22305 LANNION CEDEX
France
<...>@enssat.fr

ABSTRACT

This paper presents motivations and goals of the research and development project KGB in the field of the educational software tools for Breton language learning. Four tools based on speech technology are planned : a speech dictionary, a corrector of prosody, a speech synthesizer of breton utterances and an automatic speech dictation. The main steps of the realisation of the speech dictionary are described in more details. We present the methodological aspects used to build up the speech and linguistic resources (database of diphones, rules,...) needed for the software tools. The difficulties we met as well as the re-usability aspects are underlined.

1. MOTIVATIONS AND AIMS OF THE KGB PROJECT

The KGB[1] project arose from the joint initiative of TES[2] and of some members of the research team CORDIAL from IRISA Labs[3]. TES is a publisher of educational works in Breton language for all the schools in Britanny where Breton language is taught. Acting as a privileged observator, TES drew the following diagnostic about the practice and the learning of the Breton Language.

The practice of the Breton language is on the decline; in 1902, 75% of the population of Basse Bretagne practised Breton language; in 1990, the percentage was only about 17%, and now, only 2000 persons on 268 000 persons whose mother language is Breton are less 30 years old [BROU95][LAUR94].

The learning of Breton language is progressing. From 1977 to 1997, the strength of bilingual classes goes from few to 4 000 pupils. To these 4 000 pupils, one may add all those who study Breton language at school but without receiving a complete bilingual learning; today, the total amount is near 23 000 pupils. The goal of 10 000

[1] Kenaoz ar Gomz e Brezhoneg : Speech synthesis in Breton Language.
[2] TES : Ti-embann ar Skolioù brezhonek..
[3] IRISA: Institut de Recherches en Informatique et Systèmes Aléatoires, Rennes.

bilingual pupils is forseen in year 2000 and if the proportion is maintained, the total number of 100 000 pupils who will be concerned by Breton language could be reached.

But, the learned language is not always of a good quality because of the lack of linguistic environment. It is possible to live in Basse Bretagne without hear speaking Breton. So, pupils who learn Breton or in Breton, have rarely parents or professors whose Breton is the mother tongue. In such a context, it is very usual that pupils, and sometimes teachers, make serious errors both in syntax and pronunciation due the influence of the prominent French language.

So TES and CORDIAL thought that the new computers technologies could be very helpful for building up specific edicational tools. More precisely, speech technologies seem very appropriate for several reasons: the emergence of hight quality and robust technologies allows to consider their use in the very demanding educational context, the know-how and some results of the CORDIAL team in man-machine dialogue and speech-based systems can be re-used and at last, the fact that speech must be the main medium to be favoured.

The project KGB is devoted to the use of speech technologies - mainly speech synthesis- in the field of Breton language and plans to build up four softwares.

2 AIMED SOFTWARES

Four main applications have been defined taking into account on the one hand financial and human resources and on the other hand the necessary assessing steps, steps which have to be planned into projects in which human aspects are involved.

Speech Dictionary: a speech French-Breton dictionary (on CDROM) will allow to access both the written and spoken form of a word. This project is partially funded by the European Community.

Corrector of prosody: this software will provide exercises in order to improve pupils' words and sentences pronunciation (length of the vowels, tonic accent,...). It will work from recorded words and sentences and will be able to detect and correct the user's errors of pronunciation.

Speech synthesis of utterances: this software will produce speech synthesis from written utterances. It will be able to take into account some special tags introduced by teachers for pedagogical goals. Researches on this theme have been already started [MESSA98] and are partially funded by the "Conseil Régional de Bretagne".

Automatic dictation: this software will be devoted to the training of Breton understanding (speech and written). The pupil takes down a speech dictation ; the system will be able to detect the pupils's errors. This software exists already for French language [GUYO97a][GUYO97b]. It uses the speech synthesis software PROVERB which will therefore have to be replaced.

We describe now the realisation of the speech dictionary.

3. REALISATION OF THE SPEECH DICTIONARY

The external functionalities of this software are the following ones:

· Search of breton words: the user enters on the keyboard a word or a part of a word (using "jokers" is allowed), then the system displays a list of words and offers the user the possibility to display more information about one of the words (spelling, list of pronunciations, short definition);

· Speech synthesis: the user can ask the system to pronounce the word;

· Translation: the user enters a French word in order to have its Breton translation.

These functionalities need various resources and programs:

· a speech synthesizer; we chose the MBROLA software designed by the University of Mons [DUTO96][DUTO97];

· a database of diphones which has to be adapted to the MROLA software;

· a French-Breton dictionary providing the various phonetic transcriptions of each Breton word and with a friendly and easy user's interface;

· a program for computing prosodic temporal curve of each words;

· a man-machine interface; its function is to manage the interaction between users and machine, to manage the searches in the dictionary and to control the speech synthesis.

We will not provide any details here on this piece of software. We have only to note that the interaction has to be as friendly as possible because aimed users are children.

3.1 Speech Resources

The MBROLA speech synthesizer needs a dictionary of diphones and a program to compute pitch and duration values of each phonetic segment composing the word to be synthesized.

The dictionary of diphones is built up after a sequence of steps: setting up a list of phonemes, setting up a list of logatomes words and sentences including every Breton sound (diphone), recording these utterances, segmenting and labeling the speech signals and extracting the diphones. Prosodic modelling needs speech data about the pronunciation of words and utterances in order to extract the pertinent information. Figure 1 gives the main steps we carried on to obtain all the various elements.

Figure 1. Setting up the Speech Resources

Setting up the list of Breton phonemes: it has been necessary to determine the set of phonemes which covers the standard pronunciation of Breton as well as the various alternative dialectal pronunciations. The Breton language is composed of about 12 vowels and 7 nasal vowels, 3 semi-vowels and 22 consonants to which it is necessary to add about 15 diphtongues, that is about 60 phonemes; if we distinguish short vowels from long vowels the number of vowels is doubled leading to a total of about 80 phonemes. This work has been done with the help of specialists from the University of Rennes 2 and of

the University of Lampeter. Today, the variant "vannetaise" has not yet been modelised.

The SIL-IPA coding is used for working in WORD; the coding for the MBROLA sofware is SAMPA.

Setting up the list of logatomes: in order to have a representation of all the possible realisations of the diphones (taking into account the context and also the sequences "silence + phoneme" and "phoneme + silence" at the begining and the end of utterances), we built up a corpus of logatomes [ARBA96]. It contains around 3200 inputs, grouped in small sets, with the following format:

instructions for pronunciation
IPA coding *

Setting up the set of utterances: we built up a corpus which contains a list of the most frequent words, a list of usual expressions and a list of sentences extracted from school books [GROS84].

Recording: recording of these two corpus has been done in a studio on a DAT material. Two breton mother tongue persons (A. Ebrel, H. Gourmelon) pronounced each item of the corpus. Two recording sessions were necessary because it appeared after the first speech synthesis test that some sounds were missing.

Extracting the prosodic rules: we extracted by hand the data needed to build up the set of prosodic rules. In the future, we would like to extract this information by means of a machine learning process.

Extracting the database of diphones: this processing requires a great attention in order to obtain a good quality diphone databasebecause. It is necessary to extract, to segment and to label the diphones in the speech signal. The SNORRI speech processing software package developped at the CRIN labs (University of Nancy 1) [LAPR] turned out to be an essential and very helpful tool for this task. We also began to use the HTK speech recognition software fot this processing in order to obtain a better phone segmentation and labelling; but a more important database is necessary to train the HMM parameters.

Conclusion
This methodology could be reused again in order to improve the database of diphones especially if we want to deal with new pronunciations. However, we think that it will be necessary both to build up a machine learning process in order to extend the prosodic rules and an automatic process to extract diphones from speech signal; this last process can be based on speech recognition tools like the HTK software..

3.2 The computerized dictionary

The printed dictionary we chose is Favereau's one [FAVE92]. This choice was motivated by several reasons: we were already collaborating with its author (F. Favereau), the dictionary was available on a digital medium (WORD 2 format) and above all each word in the dictionary is accompanied by all its pronunciations according to the different Breton

dialects. Another advantage of this dictionary is also its French part allowing us to translate a French word into Breton. Other Breton dictionaries exist (for example [ANHE]) on digital medium but there are less complete from a pronunciation point of view.

Using the dictionary in the software needed several previous steps (See Fig. 2).

Figure 2. Integrating the Dictionary in the System.

Transcribing: the first version of the dictionary was a simple WORD 2 file. Then the aims of the first step have been to extract the entries and their content from this file and to create data structures adapted for searching. We obtain a list of words (orthographic format) and a list of numbered definitions. A pointer on this second list is associated to each word.

A definition is made up of three parts :

· a label (the word and its variants; the variants may be shortened);

· the different phonetical transcriptions and the grammatical categories;

· a free-structure text which contains the text of the definition itself, remarks,... The different character styles convey information (for example: italic style indicates a translation).

This transcription has been realised using UNIX tools and standard languages. In the final version, typography is encoded using a sub-set of HTML. So, we will be able to develop an Internet version of the software.

Expanding: the phonetic transcriptions of the printed version of the dictionary contain abbreviations which present difficulties to be automatically dealed with. So, about twelve volunteers expanded by hand each abbreviation before recording them in the computerized dictionary. The French part was also transformed in order to be structured as the Breton part but without phonetic transcription.

Conclusion

We think that programs, tools and data structures we defined could be re-used in case of changing dictionary or version of dictionary. Anyway this was a necessary step in order to be able to access each field of the structure: orthographic, phonetic, grammatical, stress, information that is necessary in order to synthesize words and sentences.

4. CONCLUSION

We presented motivations of this project as well as the applications we plan to build up in the field of Breton language learning: speech dictionary, corrector of prosody, speech synthesis of utterances, automatic dictation. We presented the main steps of the realization of the first application (speech dictionary). We described the activities carried on and needed either by the lack of initial data (diphones) or by the format of data unusable in a computing environment. The speech

dictionary will be available (on CDROM) during the second term of 1998. Of course, this is only a first step and in the next future, much more resources will be necessary in order to develop task-oriented speech synthesis, to develop Breton speech recognition systems; grammatical and prosodical annoted databases will also be necessary. So it is important to develop tools making this tagging task more easy.

References

[ANHE95] Geriadur brezhoneg; An Here 1995.

[ARBA96] H. Ar Barzh. *Corpus de parole pour la synthèse de la langue bretonne*. TES/IRISA, 1996.

[BROU95] Fañch Broudic. *La pratique du breton, de l'ancien régime à nos jours*, Rennes, PUR, 1995, p. 351.

[DUTO96] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, O. van der Vrecken, The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, in *Proceedings of ICSLP'96*, Philadelphia, vol. 3, P. 1393-1396.

[DUTO97] T. Dutoit, *An Introduction to Text-To-Speech Synthesis*, Kluwer Academic Publishers, 1997.

[FAVE92] F. Favereau. *Geriadur Ar Brezhoneg* (Dictionnaire du breton contemporain).
Editions Skol Vreizh, 1992.

[GROS84] J. Gros, *Le trésor du breton parlé, le style populaire*, Emgleo Breiz, Brud Nevez, 1984.

[GUYO97a] M.Guyomard, J. Siroux, D. Pernici, C. Royer. A general Public Application of Pedagogic and Linguistic Vocations of Speech Synthesis: Ordictée, in *Proceedings of ROCLING X*, p. 316-324, Taipei, Taiwan, 1997.

[GUYO97b] M.Guyomard, J. Siroux, D. Pernici, C. Royer. A Speech Synthesis Application: To learn French Spelling, in *ELM BANK Publications Exeter Call'97, Theory and Pratice of Multimedia C.A.L.L.*, 1997, to appear.

[LAPR] Y. Laprie, logiciel SNORRI, CRIN, INRIA Lorraine, http://www.loria.fr.

[LAUR94] Loeiz Laurent. *La connaissance du breton*, in *Octant*, n°56-57, Rennes, INSEE, mars 1994.

[MESSA98] Jean-Pierre Messager, Hervé Gourmelon, Guy Mercier, Jacques Siroux. Research in Speech Processing for Breton Language Training, in *Proceedings of Speech Technology in Language Learning*, ESCA Workshop, Stockholm, May 1998.