SEGRE

1. BASIC INFORMATION

1.1. Tool Name

SEGRE: An automatic tool for grapheme-to-allophone transcription in Catalan

1.2. Overview and purpose of the tool

Segre is a rule-based automatic phonetic transcription system for Catalan, jointly developed by the Universitat Politècnica de Catalunya, the Universitat Autònoma de Barcelona and the Universitat de Barcelona in the framework of the Catalan Reference Centre for Language Engineering (CREL, *Centre de Referència en Enginyeria Lingüística*).

The syntax of the rules has been designed to obtain phonetic transcriptions for four major dialects of Catalan: the Central or Eastern dialect, spoken in the East of Catalonia, the North-Western or Western dialect, spoken in the West of Catalonia (including the South), the Balearic, spoken in the Balearic Islands, and finally the Valencian, spoken in the Valencian Community.

The accuracy of transcriptions of new texts, when compared with human expert generated transcriptions, is of 99.1% for isolated words and 99,39% for running text.

Segre can be considered a useful tool to model how coarticulation modifies the isolated transcription of words in real sentences. So, it is helpful not only to build speech syntesis systems but also to train subword-based speech recognition systems.

1.3. A short description of the algorithm

The transcriptor has been designed in a very flexible way, since the rules are not fed to the program, which has very little hardwired knowledge. They are defined externally and specified in a number of ASCII text files following a simple syntax for grapheme-to-allophone and allophone-to-allophone conversion rules, the latter necessary to obtain those modifications due to coarticulation phenomena across word boundaries. So, the tool provides a phonetic transcription, as broad or narrow as desired, in isolated mode (without coarticulation across word boundaries) or in text mode (with coarticulation between words).

Furthermore, and due to the fact that the rules may be tweaked in any desired way, it is also possible, for instance, to transcribe particular subdialects or to obtain more or less narrow transcriptions. This follows from the fact that there is not a closed list of allophones in terms of which words are transcribed. Instead, the allophones are given by the various rule files.

2. TECHNICAL INFORMATION

2.1. Software dependencies and system requirements

The program is written in C++ and can be compiled to run in either Linux/Unix or Windows operative systems.

2.2. Installation

```
$ tar -xzvf Segre.tgz
$ cd Segre/fonts
$ make
```

2.3. Execution instructions

```
$./segre test.cfg
```

2.4. Input/Output data formats

Text files

2.5. Input data formats

Text file with ortographic transcription.

2.6. Output data formats

Text file with phonetic SAMPA transcription

3. CONTENT INFORMATION

3.1 A test input file

Accepten a ulls clucs els ídols esportius, artístics o
aristocràtics que els venen els agents publicitaris.

3.2. the output file

```
@ k | s ` e p | t @ | n @ | ` u L s | k l ` u g | z @ l | z ` i | D u
l | z @ s | p u rr | t ` i uw s
@ rr | t ` i s | t i g | z ` O | r i s | t u | k r ` a | t i k s | k @ l z |
B ` E | n @ | n @ l | z @ | Z ` e n s | p u | B l i | s i | t ` a | r i s
```

3.3 approximation of the time necessary to process the test input file

1 second

4. ADMINISTRATIVE INFORMATION

4.1. Contact person

Name: Climent Nadeu
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 6438
Fax: +34 93 401 6447
e-mail: climent.nadeu@upc.edu

4.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

4.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya, Universitat Autònoma de Barcelona and Universitat de Barcelona. The resource is free, license-based, for research purposes of academic or research institutions; and fee license-based for companies and commercial purposes.

5. RELEVANT REFERENCES AND OTHER INFORMATION

P. Pachès, C. de la Mota, M. Riera, M. P. Perea, A. Febrer, M. Estruch, J. M. Garrido, M. J. Machuca, A. Río, J. Llisterri, I. Esquerra, J. Hernando, J. Padrell, C. Nadeu, "Segre: An automatic tool for grapheme-to-allophone transcription in Catalan",

Workshop on developing language resources for minority languages: reusability and strategic priorities, Athens, May 30th, 2000.