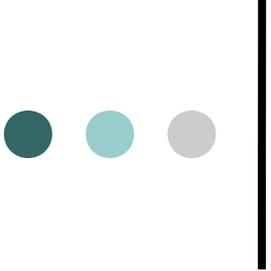




Normalization of dialects and
variants using FST technology
Phonetisaurus: an introduction

Izaskun Etxeberria
Iñaki Alegria
(University of The Basque Country)



References

Toolkit and tutorial:

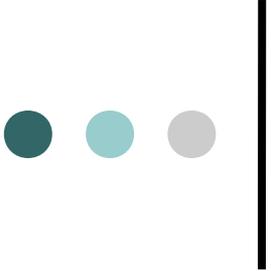
- <http://code.google.com/p/phonetisaurus/>

Additional software

- www.openfst.org/twiki/bin/view/FST/FstDownload
- www.openfst.org/twiki/bin/view/GRM/NGramDownload

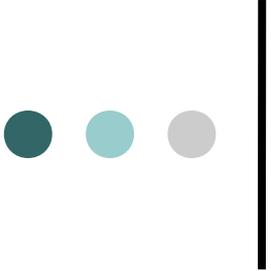
Our work

- *I. Etxeberria, I. Alegria, M. Hulden, L. Uria 2014. Learning to map variation-standard forms using a limited parallel corpus and the standard morphology. SEPLN, 52, pp. 13-20.*



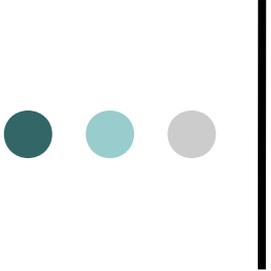
Features

- induction of weighted rules from examples
- machine-learning: noisy-channel model (usual in speech)
- we use it for grapheme-to-grapheme (g2g)
(no phonemes as usual in speech)
- (a bit) difficult to install
 - dependencies with other software
- using tuning results can be improved



Steps

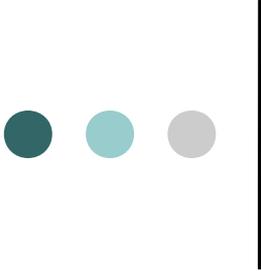
- Step 0: get the examples (training/test)
- Step 1: aligning the grapheme/phoneme sequences in the training dictionary
- Step 2: training a joint-sequence N-gram model from the alignment result
- Step 3: testing



Format of examples

- *file.train*
- Pairs of words in 2 columns (word / letter-sequence)
- Example:

```
anzatsuenak      a n t z e t s u + e n + a k  
aphetitu         a p e t i t u  
aphetituari      a p e t i t u + a r i
```

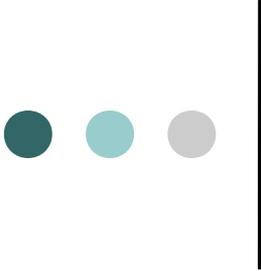


Aligning

- Command:

```
phonetisaurus-align --input=$1/$1.train  
                    --ofile=$1/$1.corpus
```

*"The result of the previous step is simply a corpus of aligned joint-sequences. This can be used directly to train a standard N-gram model using any SLM toolkit capable of outputting a standard ARPA-format model. Examples training commands are given below for the MITLM, SRILM, and Google **NGramLibrary**."*



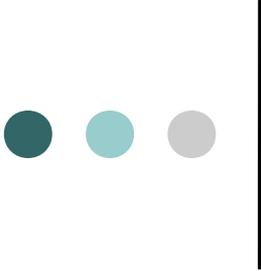
Training

- NGramLibrary:

```
ngramsymbols <$1/$1.corpus >$1/$1.syms
farcompilestrings --symbols=$1/$1.syms
                 --keep_symbols=1 $1/$1.corpus >$1/$1.far
ngramcount --order=7 $1/$1.far >$1/$1.cnts
ngrammake --method=kneser_ney $1/$1.cnts >$1/$1.mod
ngramprint --ARPA $1/$1.mod >$1/$1.arpa
```

- All together: generation of FST

```
phonetisaurus-arpa2fst --input=$1/$1.arpa
                      --prefix="$1/$1"
```



Test

- Command:

```
phonetisaurus-g2p --model=$1/$1.fst --input=$2  
--isfile --words --nbest=5 --beam=5000 >$2.out
```

- *beam*: can be changed for tuning (deep/speed)
- *nbest*: number of possibilities in the output
- format of the output (nbest=5)

```
arazitzen 10.3353 a r a t z + t e n  
arazitzen 17.3496 a r a z + t e n  
arazitzen 17.6407 a r a t z + + t z e n  
arazitzen 17.6419 a r a t z i + t z e n  
arazitzen 17.7889 a r a t z + t e + n
```