

Automatic rule-based syllabication for Romanian



Toma Ștefan-Adrian*



Oancea Eugeniu*

Doru-Petru Munteanu*

* Military Technical Academy

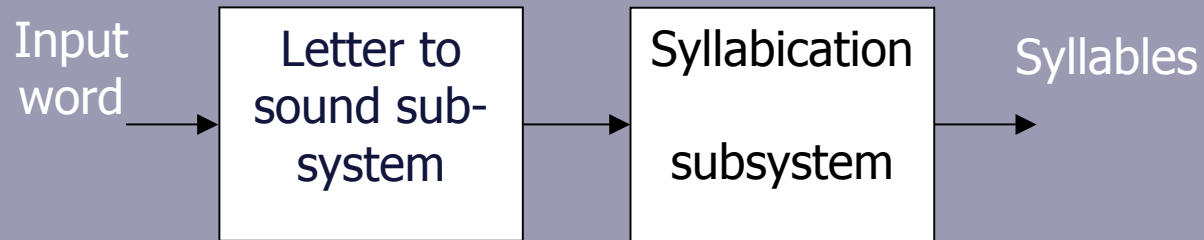
Contents

- Motivation
- Syllabication system architecture
 - The letter to sound sub-system
 - The syllabication sub-system
- Experimental results and conclusion

Motivation

- The best strategy for concatenative synthesis is to use a multi-level system: that is to have a words or sentences database and choose the appropriate word from it; if the word is not in database the system generates the word by composing it from syllables or half-syllables; if the system still can't find suitable speech units to form the required word it goes even deeper and forms the word with phones, diphones or triphones.
- Since word stress, syllabication and phonetic transcription are inter-dependent it is obvious that to determine which syllable in a word is stressed, one needs to determine the phonetic transcription and the syllables.

System description



The problem of automatic syllabication is addressed by first making phonetic transcription using a rule-based letter-to-sound system and then applying a set of 24 rules to split the word into syllables. Hence, the system is made of a letter-to-sound (LTS) subsystem and a syllabication subsystem.

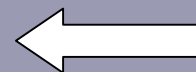
The letter to sound sub-system

Phoneme	Example	
	Graphic	Phonetic
[e]	NEANT	n e a n t
[e_X]	MEA	m e_X a
[e e_X]	ACEEA	a tS e e_X a
[j]	EA	j a
[j e]	ESTE	j e s t e
[=]	CEAS	tS a s



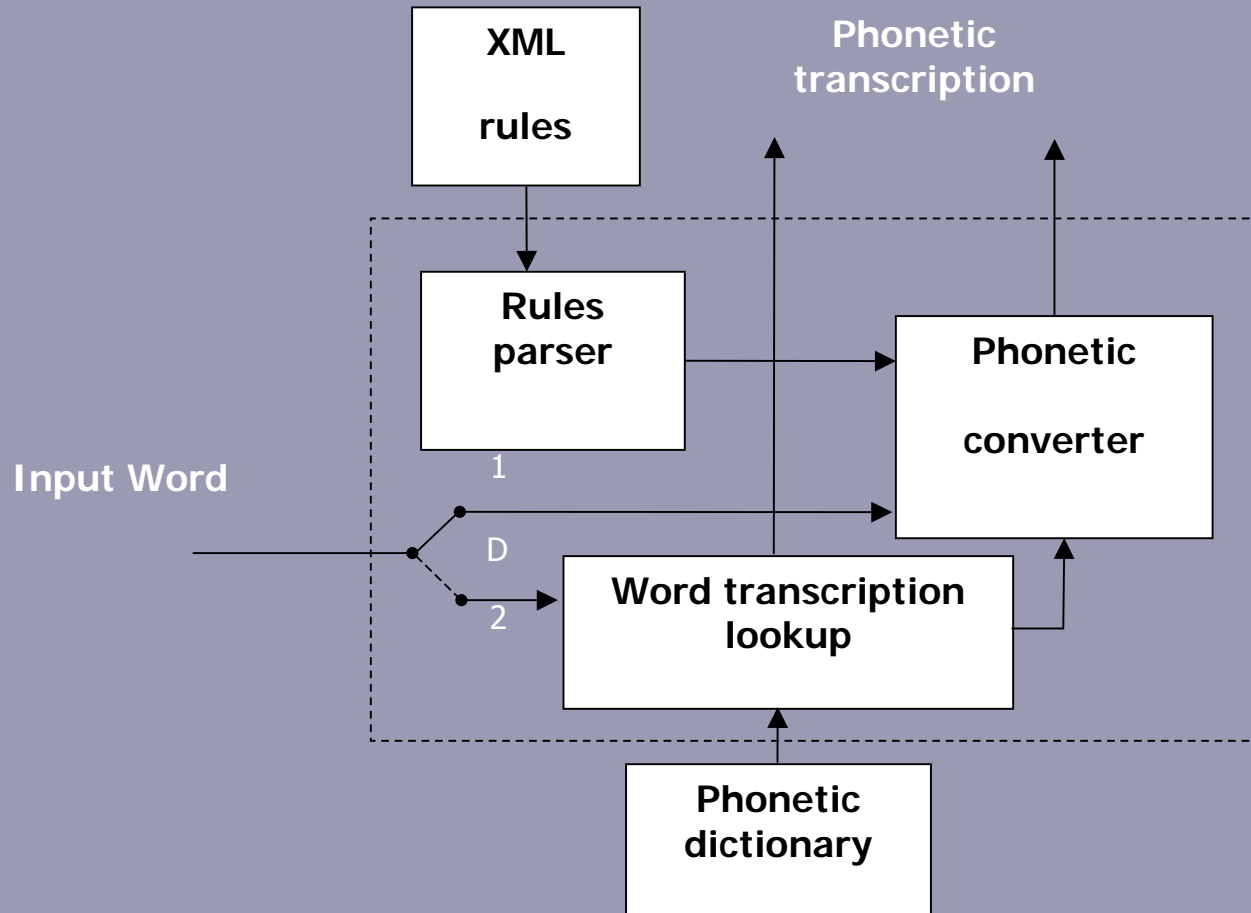
Phonetic values
of the letter E

Phoneme	Example	
	Graphic	Phonetic
[i]	MIRARE	m i r a r e
[j]	IARBĂ	j a r b @
[i_0]	ORICE	o r i_0 tS e
[=]	CIOBAN	tS = o b a n

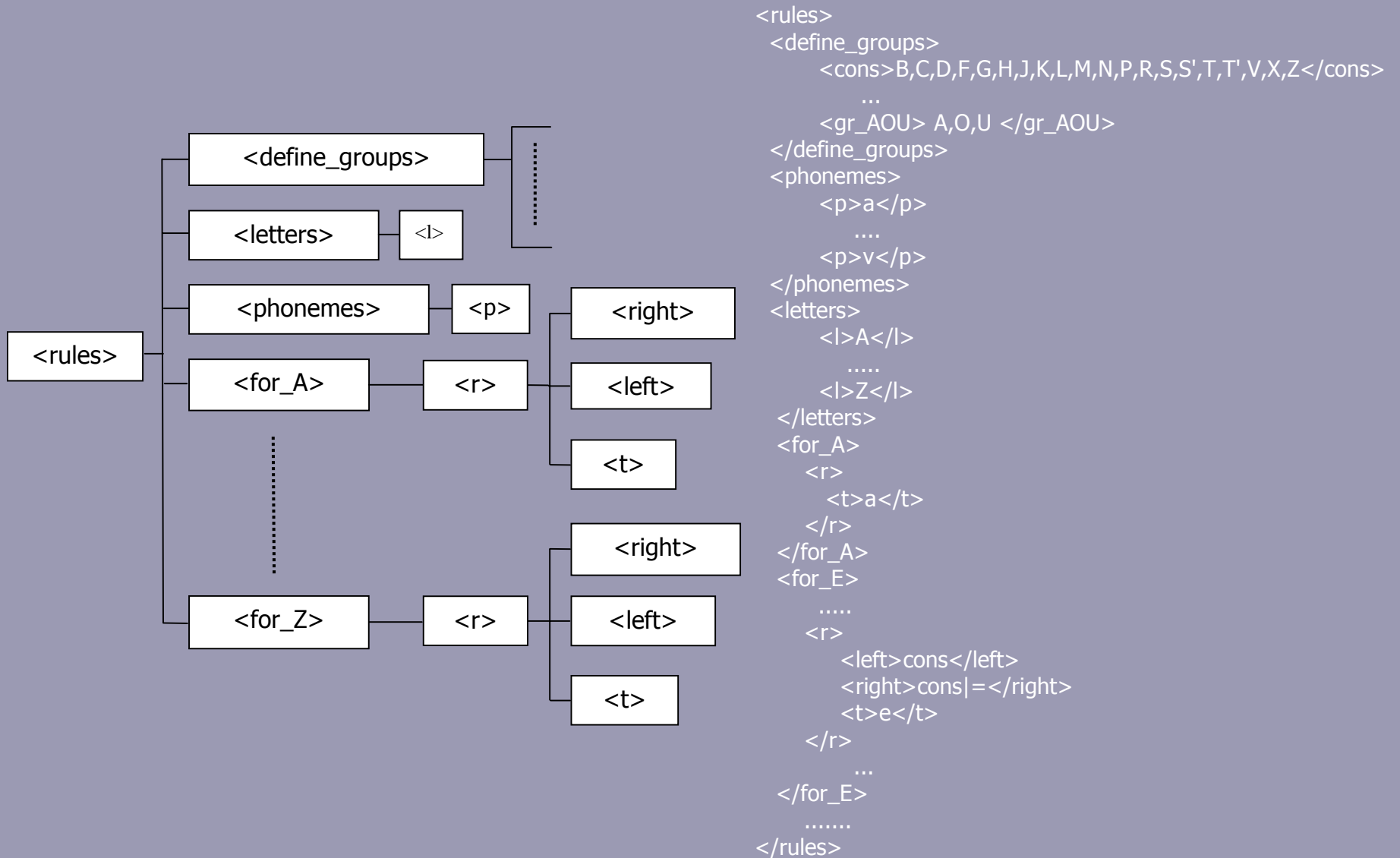


Phonetic values
of the letter I

The letter to sound sub-system



The letter to sound sub-system



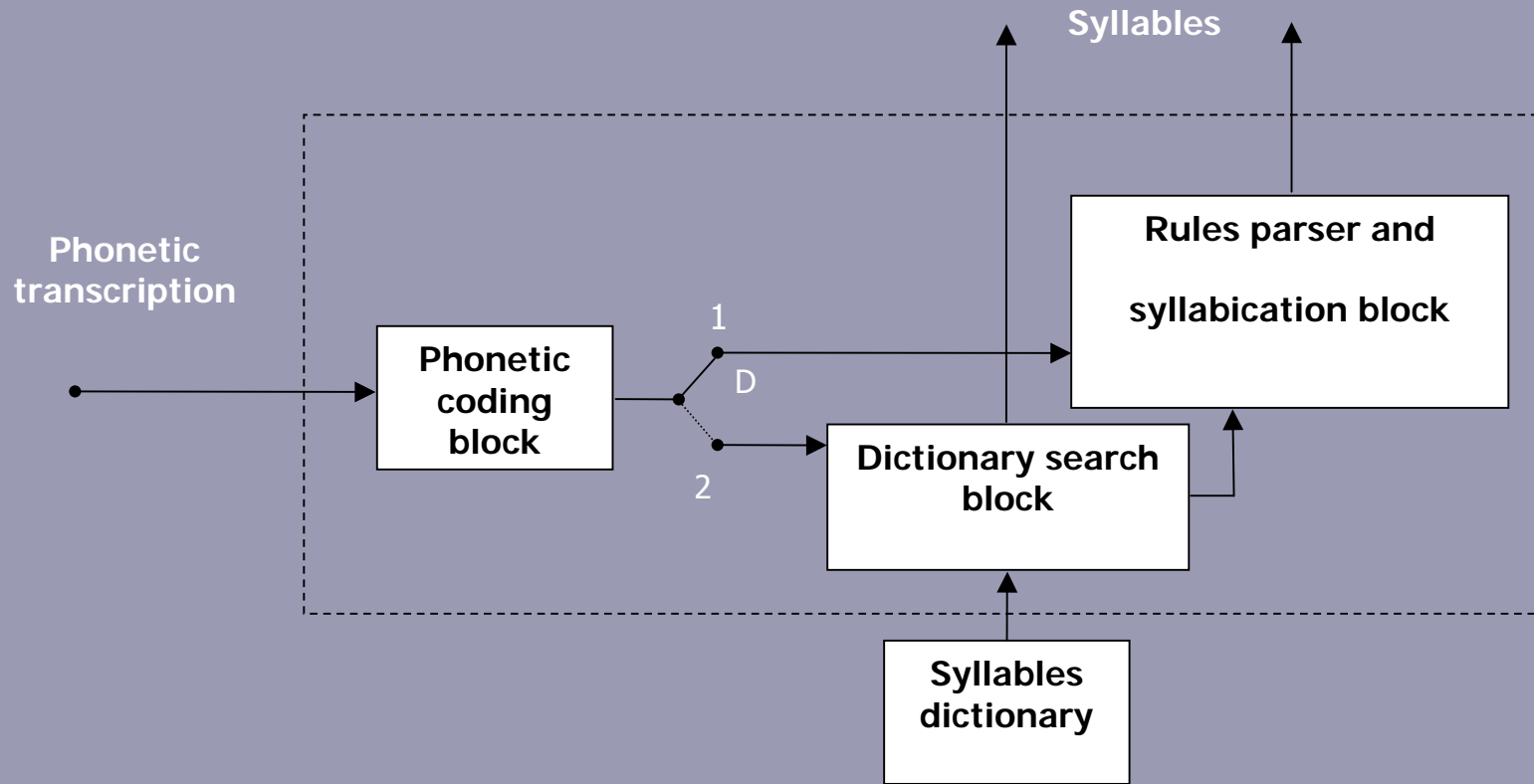
```

<rules>
  <define_groups>
    <cons>B,C,D,F,G,H,J,K,L,M,N,P,R,S,S',T,T',V,X,Z</cons>
    ...
    <gr_AOU> A,O,U </gr_AOU>
  </define_groups>
  <phonemes>
    <p>a</p>
    ...
    <p>v</p>
  </phonemes>
  <letters>
    <l>A</l>
    ....
    <l>Z</l>
  </letters>
  <for_A>
    <r>
      <t>a</t>
    </r>
  </for_A>
  <for_E>
    ....
    <r>
      <left>cons</left>
      <right>cons|=</right>
      <t>e</t>
    </r>
    ...
  </for_E>
  .....
</rules>
  
```

The letter to sound sub-system

We tested the system on a list of 4779 words consisting of the most 5000 frequent words from a text corpus made of 93 books representing genuine Romanian literature, foreign literary works translated into Romanian and scientific texts. minus foreign words. The overall letter error rate is 0.72 % while the overall word error rate is 4.79 %.

The syllabication sub-system



The syllabication sub-system

```
for current_phoneme=1 to end_of_word
{
phonetic_context(current_letter)=function_of_phonetic_code;

if (phonetic_context(current_letter)= VCSV)
{ split_phonetic_context(V-CSV);

  continue for_loop; }

elseif (phonetic_context(current_letter)= VCCSV)
{ split_phonetic_context(VC-CSV);

  continue for_loop; }

elseif (phonetic_context(current_letter)= CCVVC)
{ split_phonetic_context(CCV-CV);

  continue for_loop }

}
```

Word	C-V-S notation	Syllabication
<i>STE A</i>	C C S V	CCSV
<i>M E R E</i>	C V C V	CV/CV
<i>B U G E T</i>	C V C V C	CV/CVC
<i>O A I E</i>	S V S V	SV/SV
<i>P A I E</i>	C V S V	CV/SV

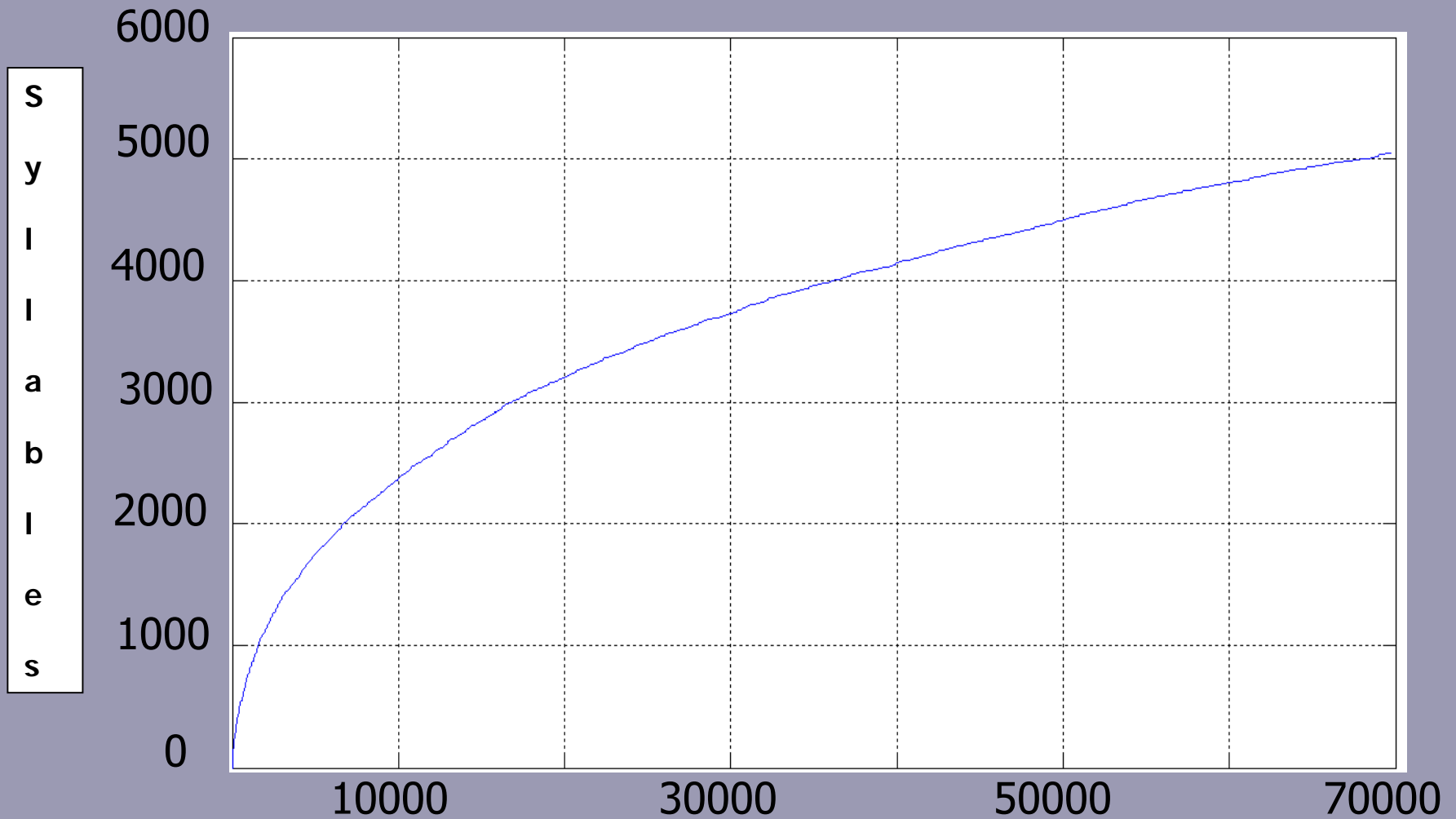
Experimental results and conclusions

The system was tested on a 72921 words database collected from the 75339 words list of the accepted words by the Scrabble Romanian Federation. Foreign words were discarded. Of the 72921 words 7696 words were wrongly syllabicated. **The obtained error rate was 10.55%.**

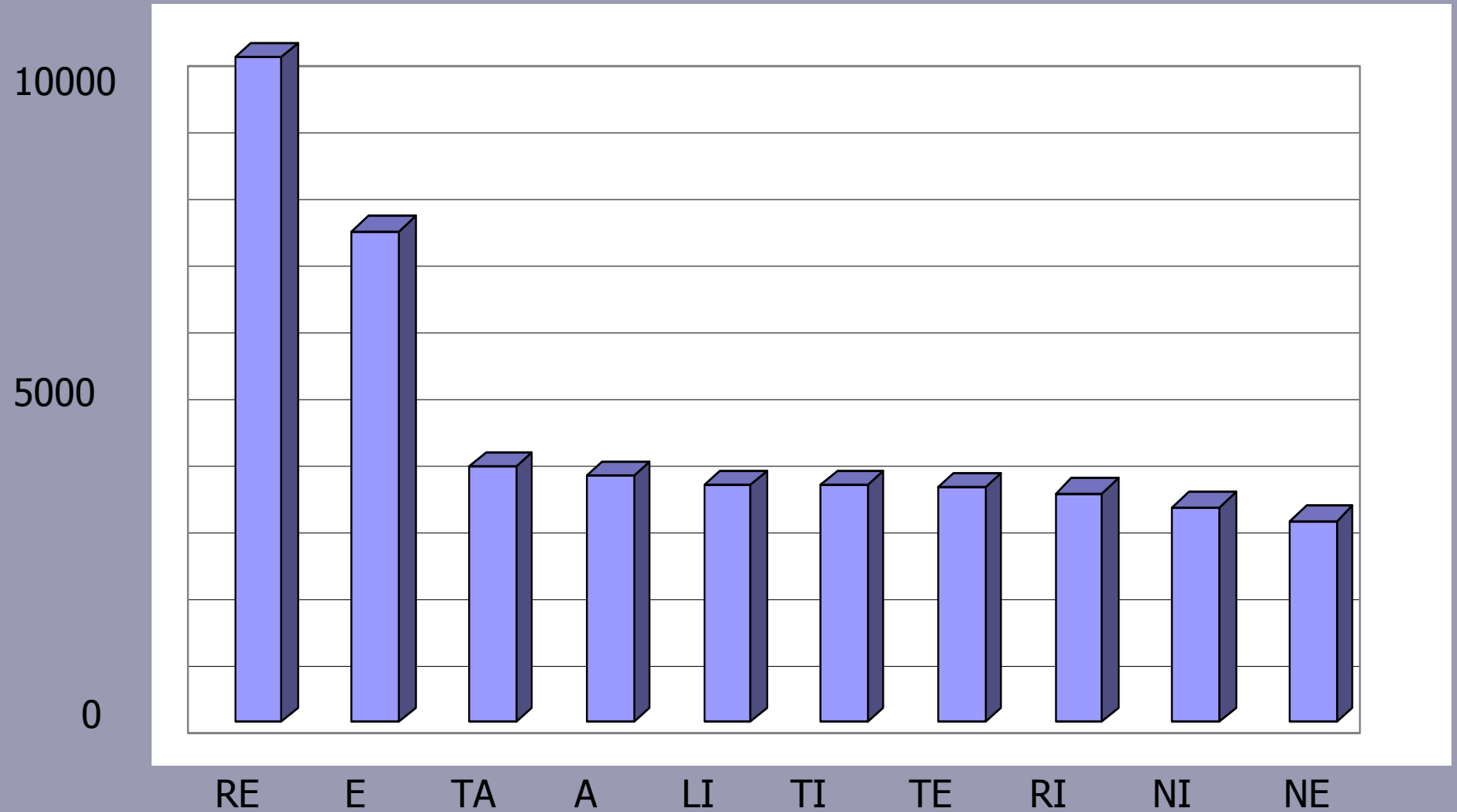
There were three types of errors:

- errors generated by erroneous phonetic transcriptions,
- errors generated by the syllabication process
- errors generated by both erroneous phonetic transcriptions and the syllabication process.

Experimental results and conclusions



Experimental results and conclusions



Thank you!