

On Phonetic Transcription for Romanian – a Comparison of Five Algorithms

Ștefan-Adrian Toma
Traian Bîrsan
Felix Totir
Eugeniu Oancea

TRANSCRIERE
REDUCERE
ARBORE
FONETICĂ
ACCIUNȚĂ
CLASIFICARE
ARBORULUI
SILABE
DE
CAUTARE
DIPLOMATIE
CAUTARE
LIMBAJ
NATURAL
PROCESARE

OVERVIEW

Romanian phonetics

Presentation of the LTS tested

- expert system
- neural networks
- decision trees
- support vector machines
- pronunciation by analogy

Test setup

Experimental results

Conclusion

TRANSCRIERE
REDUCERE
ARBORE
DE CAUTARE
SILABE
FONETICA
ACCIUNTI
SPECAULUI
SILABE
ARBORE
CLASIFICARE
LIMBAJ NATURAL
PROCESARE

ROMANIAN PHONETICS

- Romanian is a major Romance language, along with Italian, French, Spanish and Portuguese.
- Mostly phonetic
- 31 letters corresponding to 34 phonemes: 22 consonants, 7 vowels, 4 semivowels and 1 non-syllabic vowel.
- There are letters pronounced in several ways.
- Semivowels are the most difficult to identify.

TRANSCRIERE
REDUCERE
ARBORE
DE CAUTARE
SILABE
FONETICA
ACSCENT
DE CAUTARE
LIMBAJ NATURAL
ARBOR
SILABE
PROCESARE

THE LTS ALGORITHMS

Expert system (ES)

Neural networks (NN)

Decision trees (DT)

Support vector machines (SVM)

Pronunciation by analogy (PbA)

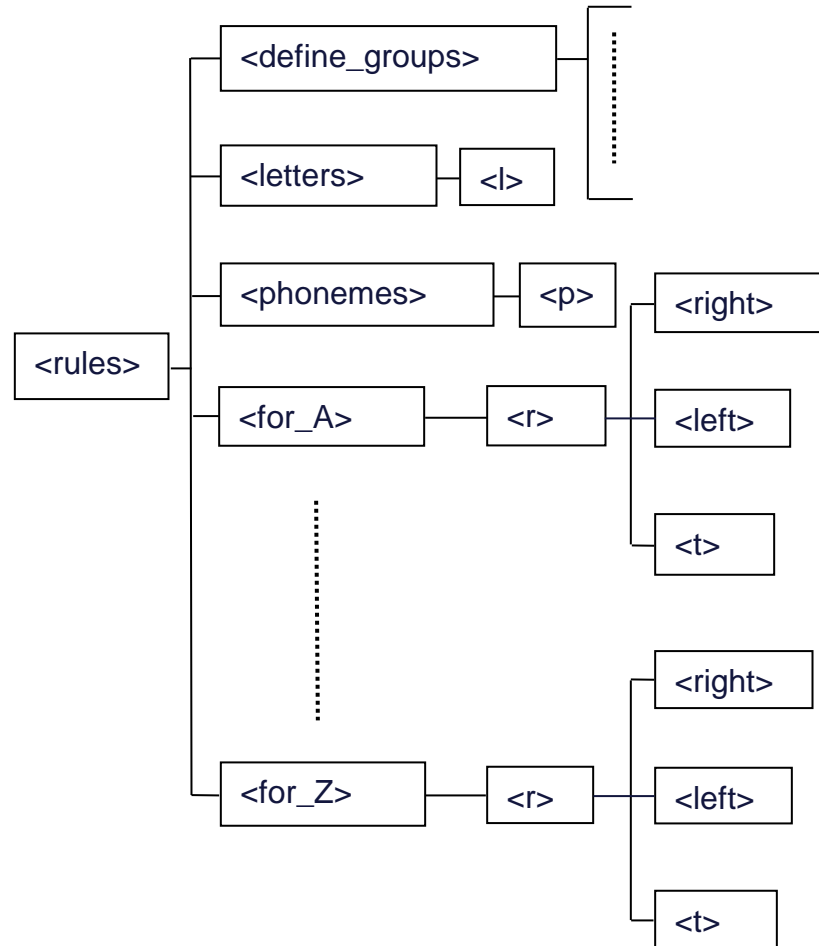
TRANSCRIERE
REDUCERE
ARBORE
FONETICA
CLASIFICARE
SILABE
ARBOR
CLASIFICARE
SILABE
DECAUTARE
LIMBAJ NATURAL
PROCESARE

THE EXPERT SYSTEM

TRANSCRIERE
REDUCERE
ARBORE
CLASIFICARE
PROCESARE

FONETICA
ACSCENT
LIMBAJ NATURAL

ARBORE
DE CAUTARE
SILABE



THE EXPERT SYSTEM

TRANSCRIERE
REDUCERE
ARBORE
FONETICA
ACCENT
DESPICACITARE
SILABE
ARBORE
CLASIFICARE
LIMBAJ NATURAL
PROCESARE

M	= = M A M
A	= M A M A
M	M A M A =
A	A M A = =

THE EXPERT SYSTEM

TRANSCRIERE
REDUCERE
ARBORE
ARBORE
FONETICA
CLASIFICARE
CAUTARE
DE CAUTARE
LIMBAJ NATURAL
SILABE
PROCESARE

```
<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>
    ....
    <r>
      <right>A</right>
      <right>=</right>
      <t>e_X</t>
    </r>
    ....
  </for_E>
  .....
</rules>
```

102 rules

NEURAL NETWORK BASED LTS

Tool: WEKA – open source, easy to use
 Data presented as ARFF files. It is possible to use Weka classes in Java application.

M	= = M A M
A	= M A M A
M	M A M A =
A	A M A = =

M,=,=,A,M,m

A,=,M,M,A,a

M,M,A,A,=,m

A,A,M,=,=,a

Feature vectors

TRANSCRIERE
 REDUCERE
 ARBORE
 CLASIFICARE
 DEPRECAUTARE
 LIMBAJ NATURAL
 PROCESARE
 SILABE

NEURAL NETWORK BASED LTS

Several configurations were tested until a network with 23 neurons in one hidden layer and a learning rate of 0.1 produced the best results (regarding both error rate and training time).

The parameters were found using an automatic grid search on a subset of the training set. We then applied the parameters on the whole training set.

TRANSCRIERE
REDUCERE
ARBORE
DECAUTARE
SILABE
FONETICA
ACCENT
DECAUTARE
LIMBAJ
NATURAL
ARBORE
CLASIFICARE
PROCESARE

NEURAL NETWORK BASED LTS

Increasing the number of neurons in the hidden layer leads to some small improvement but the training time made it impractical for us.

In light with the results obtained by other authors we decided to keep 23 neurons, and not increase this number.

NN used by other authors for LTS conversion in Romanian.

D. Burileanu. "Basic research and implementation decisions for a text-to-speech synthesis system in Romanian". *International Journal of Speech Technology*, Vol. 5(3), pp. 211-225, 2002.

TRANSCRIERE
REDUCERE
ARBORE
FONETIC
ACSCENT
CLASIFICARE
SILABE
ARBORE
DECAUTARE
SILABE
LIMBAJ NATURAL
PROCESARE

DECISION TREES BASED LTS

Tool: WEKA, C4.5 algorithm (J48)

Data presented as ARFF files. It is possible to use Weka classes in Java application.

M	= = M A M
A	= M A M A
M	M A M A =
A	A M A = =

M,=,=,A,M,m

A,=,M,M,A,a

M,M,A,A,=,m

A,A,M,=,=,a

Feature vectors

TRANSCRIBER
REDUCERE
ARBORE
FONETICA
CLASSIFICARE
CAUTARE
DECAUTARE
SILABE
ARBORE
CAUTARE
SILABE
LIMBAJ NATURAL
PROCESARE

DECISION TREES BASED LTS

The best results for decision trees were obtained for a pruned tree with binary splits and a minimum number of objects of one in any leaf.

TRANSCRIERE
REDUCERE
ARBORE
FONETICA
ACSCENT
CLASIFICARE
LIMBAJ
PROCESARE
DE
CAUTARE
SILABE
CAUTARE
NATURAL

SVM BASED LTS

Tool: libSVM (& WEKA)

Data presented as ARFF files. It is possible to use Weka classes in Java application.

M	= = M A M
A	= M A M A
M	M A M A =
A	A M A = =

M,=,=,A,M,m

A,=,M,M,A,a

M,M,A,A,=,m

A,A,M,=,=,a

Feature vectors

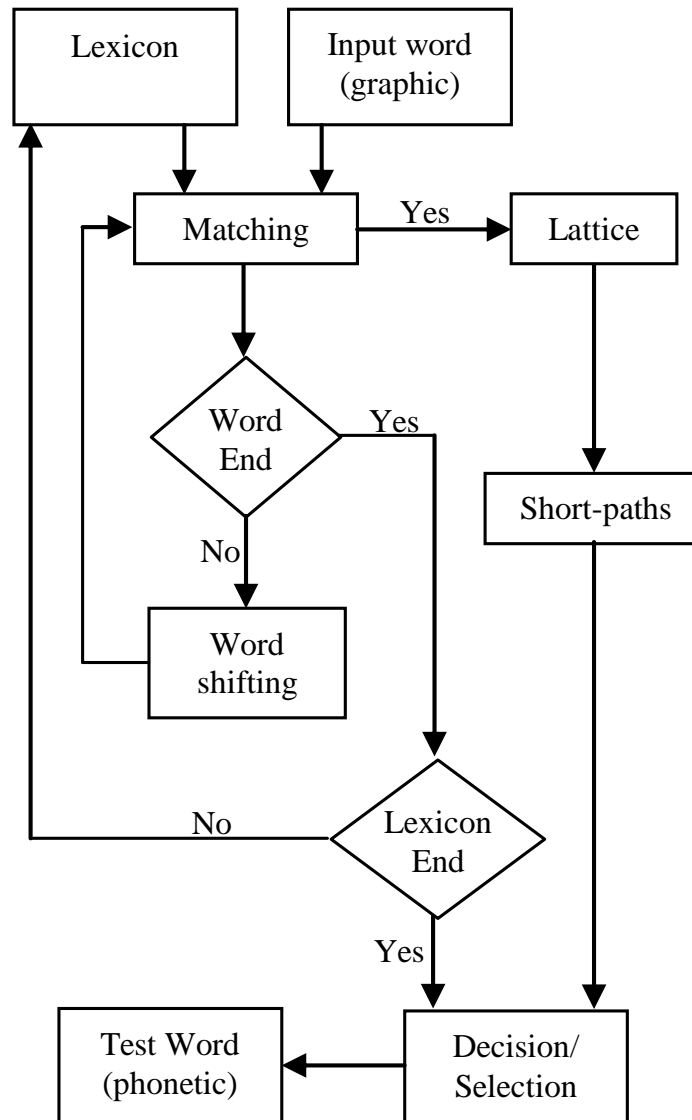
TRANSCRIERE
REDUCERE
ARBORE
DE CAUTARE
SILABE
FONETICA
ACCIUNTA
DE CAUTARE
LIMBAJ NATURAL
PROCESARE
ARBORIC
CLASIFICARE
SILABARE

SVM BASED LTS

- In our implementation of a LTS converter with SVMs, we used libSVM.
- First try – linear kernel ☹
- Second try – Gaussian kernel – ☺
- accuracy of the model is highly dependent on γ and on the tradeoff between training error and margin (denoted by C)
- grid search for finding a suitable pair (γ , C)
- Best results for $C=40$ and $\gamma=1$

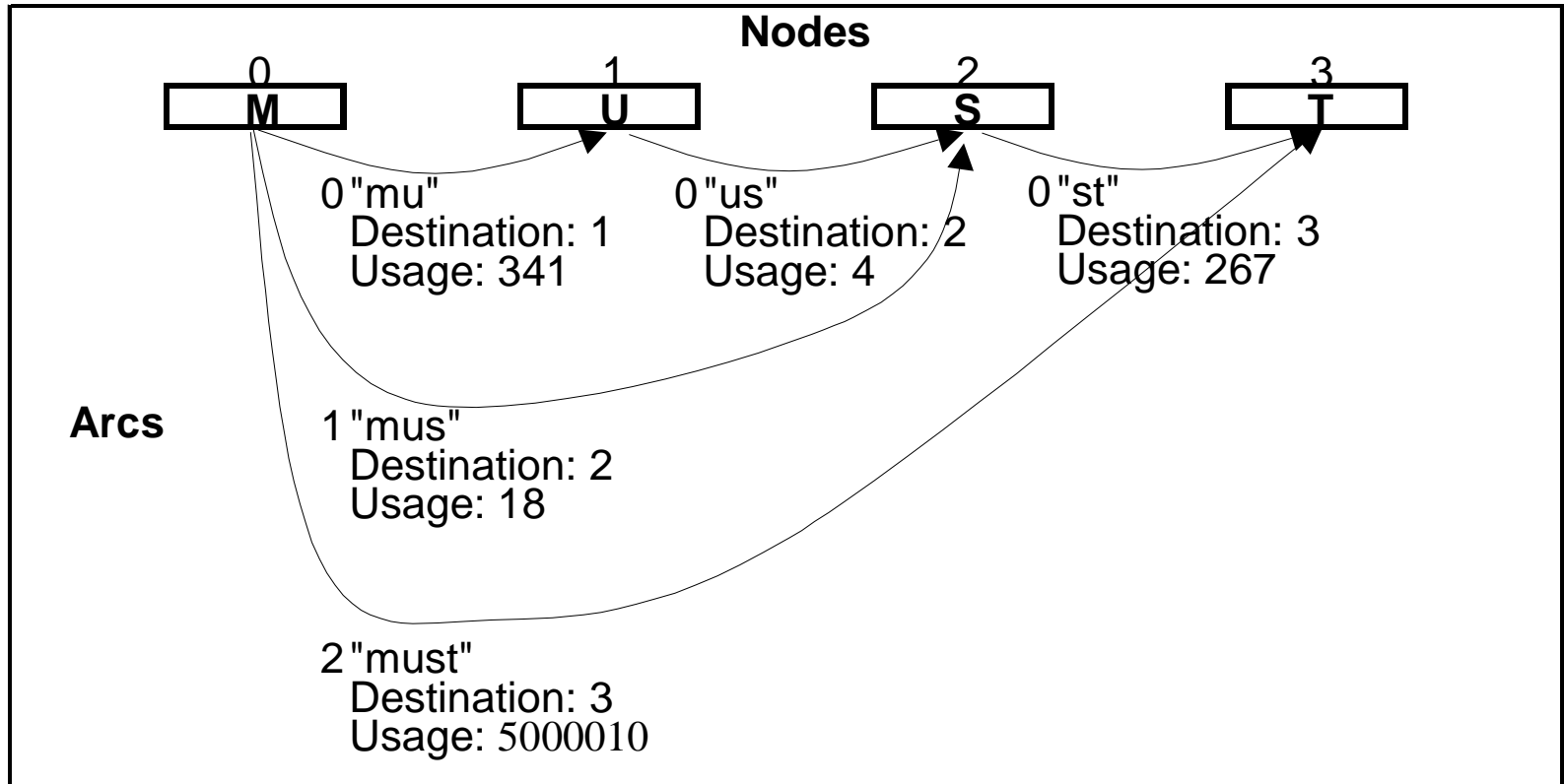
TRANSCRIERE
REDUCERE
ARBORE
FONETICA
ARBORE
CLASIFICARE
DESPICATULUI
LIMBAJ NATURAL
PROCESARE
CAUTARE
SILABE

PRONUNCIATION BY ANALOGY



TRANSCRIERE
REDUCERE
ARBORE
FONETICA
CLASIFICARE
LIMBAJ NATURAL
DE CAUTARE
SILABE
ARBORE
CLASIFICARE
LIMBAJ NATURAL
DE CAUTARE
SILABE
ARBORE
CLASIFICARE
LIMBAJ NATURAL
DE CAUTARE
SILABE

PRONUNCIATION BY ANALOGY



Decision based on most common pronunciation.

TRANSCRIERE
REDUCERE
ARBORE
CLASIFICARE
DESPICARE
CAUTARE
LIMBARE
NATURALE
SILABARE
PROCESARE

TEST SETUP

- 2 phonetic dictionaries
- “training” dictionary 15,517 words and 114,044 letters
- test dictionary - most frequent 4779 words (31,483 letters) in 93 Romanian books (literary works and science literature)
- SAMPA notation
- although they seem small, they retain the important characteristics of Romanian pronunciation

TRANSCRIERE
REDUCERE
ARBORE
DE CAUTARE
SILABE
FONETIC
ACCENT
LIMBAJ
NATURAL
ARBORE
CLASIFICARE
PROCESARE

EXPERIMENTAL RESULTS

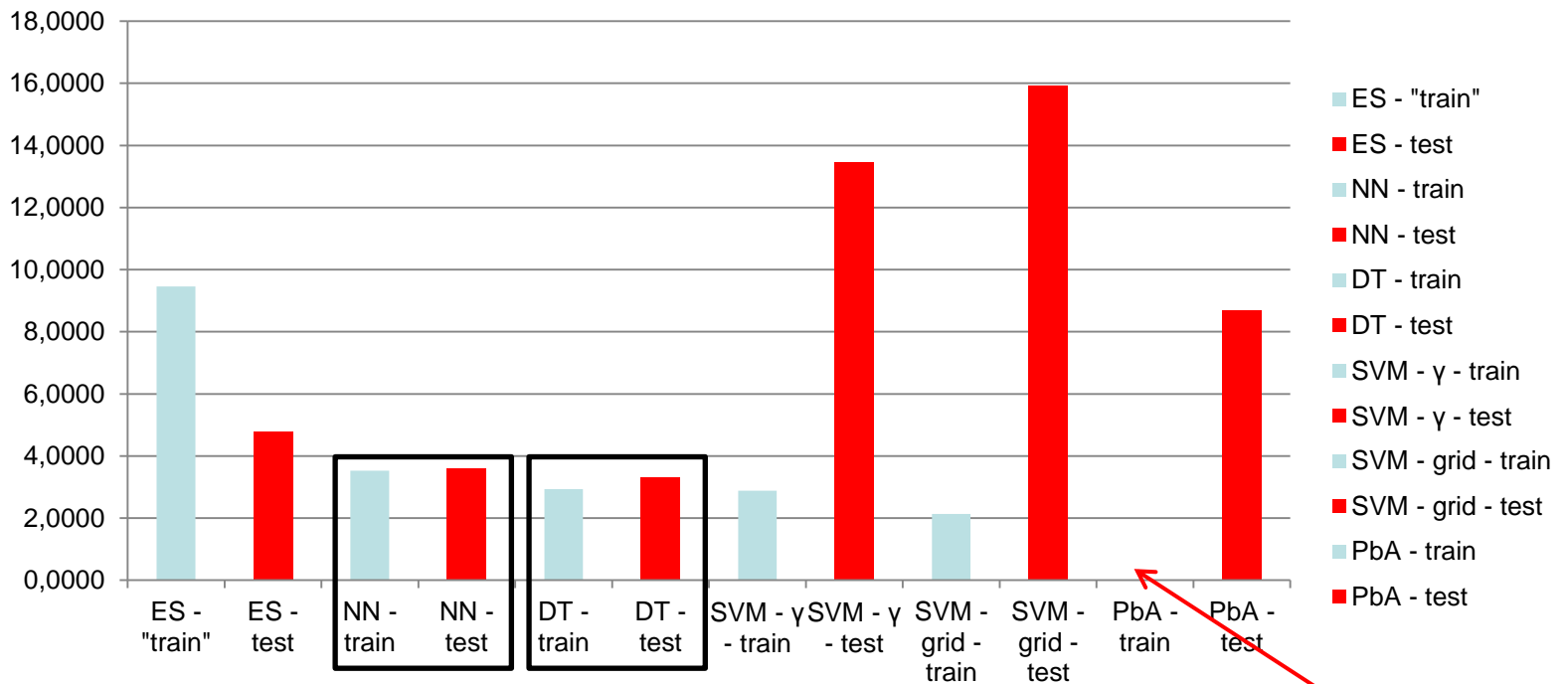
LTS conversion method		Train	Test
		15.517	4779
Rule-based	LER	1.3	0.72
	WER	9.46	4.79
MLP 1 layer	LER	0.5393	0.6096
	WER	3.5294	3.6185
Decision tree	LER	0.4419	0.5177
	WER	2.9313	3.3273
SVM γ search	LER	0.4226	3.100
	WER	2.8787	13.455
SVM grid search	LER	0.3183	4.2892
	WER	2.1295	15.938
PbA	LER	0.00087	1.3213
	WER	0.0064	8.7047

TRANSCRIBER
 REDUCER
 FONETIC
 ARBOR
 CLASSIFIER
 DECAUTAR
 LIMBAJ
 PROCESAR
 SILABE
 NATURAL

EXPERIMENTAL RESULTS

TRANSCRIBER
REDUCERE
ARBORE
FONETICA
CLASIFICARE
DESPICARE
CAUTARE
LIMBAJ
NATURAL
SILABE
PROCESARE

WER [%]

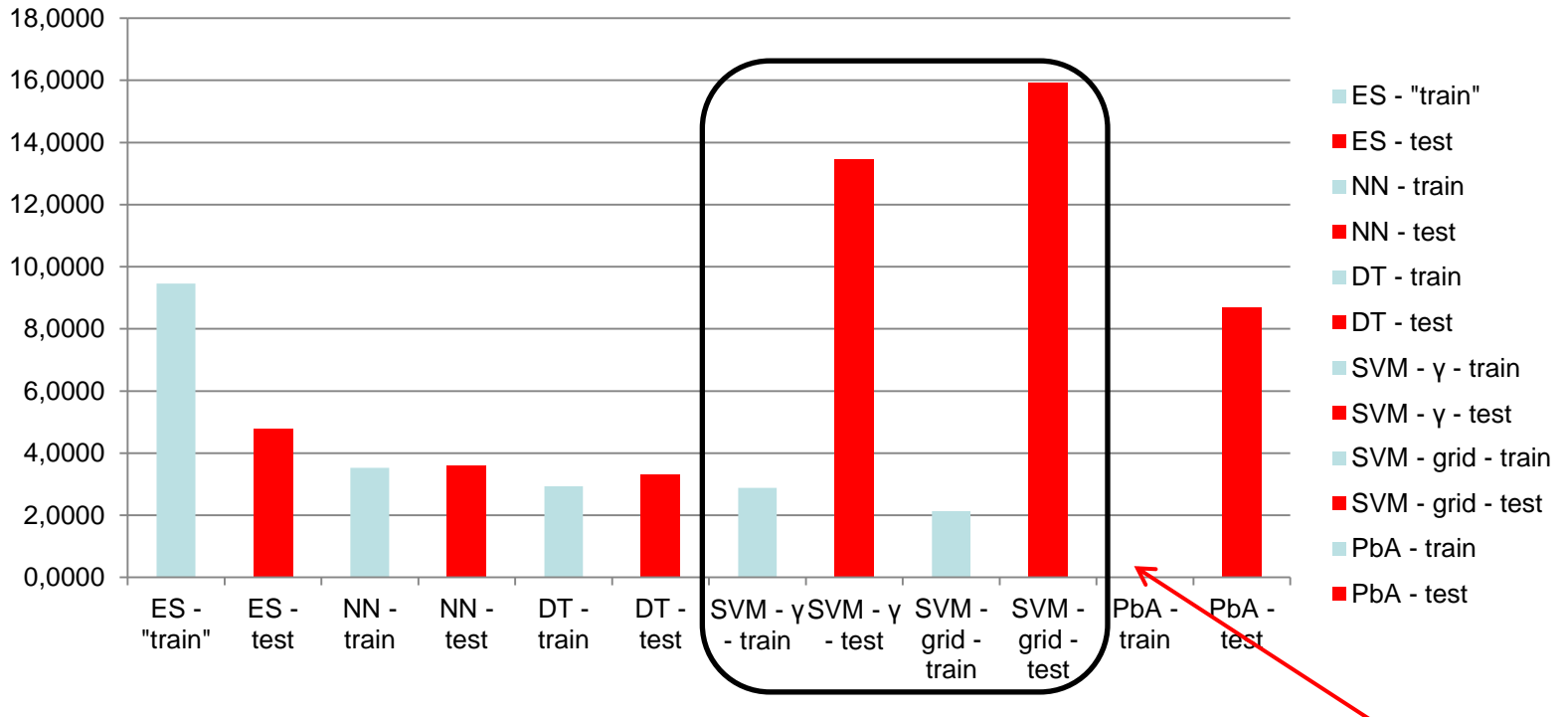


very small value

EXPERIMENTAL RESULTS

TRANSCRIBER
REDUCERE
ARBORE
FONETICA
CLASIFICARE
DESPICARE
CAUTARE
LIMBAJ
NATURAL
ARBORE
DECAUTARE
SILABE
PROCESARE

WER [%]



very small value

EXPERIMENTAL RESULTS

Burileanu, 1999	dictionary	4.000	1.000
	WER [%]	1,6	2,9
Jitca, 2003	dictionar	1.000	4.00
	WER [%]	-	5,00
Ordean, 2009	dictionar	200	1000
	WER [%]	0% la 25%	min. 5,00

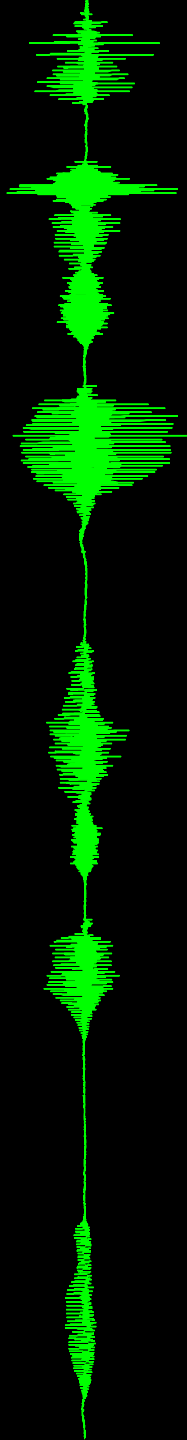
- statistical machine translation based LTS conversion which was applied to Romanian by H. Cucu, L. Besacier, C. Burileanu, A. Buzo in Investigating the role of machine translated text in ASR domain adaptation: Unsupervised and semi-supervised methods.(2011)

TRANSCRIERE
REDUCERE
ARBORE
CLASIFICARE
DESPRE
CAUTARE
LIMBAJ
NATURAL
PROCESARE
SILABE
CAUTARE

CONCLUSION

- the decision tree and neural network approaches to phonetic transcription provide the best results, for Romanian
- PbA is promising; this is still work in progress
- NN results consistent with other work

TRANSCRIERE
REDUCERE
ARBORE
FONETIC
CLASIFICARE
SILABE
ARBORE
DE CAUTARE
SILABE
LIMBAJ NATURAL
PROCESARE



Thank you!