New Experiments

Kiril Simov and Petya Osenova

Linguistic Modelling Laboratory

Bulgarian Academy of Sciences

(http://www.BulTreeBank.org)

BulDialects Project

3-4 November 2006

Sofia, Bulgaria

Plan of the Talk

- New Experiments
- Proposal for Language Contact Characterization
- Software Infrastructure

New Experiments - Task

- To check how the compression method works with the new data
- To compare some dialects to the Bulgarian and Serbian Standard languages
- Different data generation methods

Experiment

- Selection of two sets of three sites each, such that the first set is of dialects closer to Bulgarian standard language and the second to Serbian standard language (expert judgment – Vladimir):
  - Kramolin, Sevlievo; Kravenik, Sevlievo; Zdravkovec, Gabrovo (closer to Bulgarian)
  - Aldomirovci, Slivnica; Golemo Malovo, Slivnica; Razboishte, Godech (closer to Serbian)

Permutation Method

- First, each word is segmented in phonemes and in bigrams, trigrams, etc (as discussed in Tuebingen)
- Then each permutation is generated and stored
- Restrictions – long words – too many permutations, too big impact

Permutation Method - Results

Cluster 1:

Kramolin; Kravenik; Zdravkovec – 0.49

distance to others > 0.90

Cluster 2:

Aldomirovci; Golemo Malovo; Razboishte – 0.36

distance to others > 0.90

Segmentation Method

- Similar to permutation method, but only segmentation in n-grams
- We have enough data that the n-grams to be enough

Results of Segmentation Method

Cluster 1:

Kramolin; Kravenik; Zdravkovec – 0.36

distance to others > 0.80

bigger to Serbian ~ 0.88

Cluster 2:

Aldomirovci; Golemo Malovo; Razboishte – 0.26

distance to others > 0.80

Explanation of the results

- The clustering of the dialects is reflecting the expert intuition
- Standard languages not comparable description
- Small differences in the features descriptions have big differences
- Two questions:
  - Feature encoding – granularity, one symbol – one feature
  - Feature selection – which feature are important in comparison

Proposal for Language Contact Characterization (1)

- Selection of distance metric
- Selection of set of features
- Maximization of the set of features for a given distance
- The best set(s) of features determines the features shared by the languages with respect to the given metric

Proposal for Language Contact Characterization (2)

- *DL1* – description of language L1 with set of features *F*, similar *DL2* for L2 the same set F
- *d(x,y)* - distance metric, ε a given distance
- The best set of features *Fb* for comparing the language L1 and L2 wrt *F*, *d(x,y)* and ε is such that

$Fb \in \{Y \subseteq F \mid d(DL1/Y, DL2/Y) \leq \varepsilon\}$ and

$|Fb| = \max |X| : X \in \{\mathrm{Y} \subseteq F \mid d(DL1/Y, DL2/Y) \leq \varepsilon\}$

Software Infrastructure

- CLaRK as a server supporting web services
- The Groningen and Tuebingen tool sets available by web services
- The data is recorded locally or on a server
- Processing is done in the following steps:
  - Transfer of the data to the Groningen or Tuebingen server
  - Processing with the required tools
  - Transfer of the result back

Conclusions

Here we discussed:

- The new experiments shown that small differences in the encoding play roles
- Feature characterization of language contacts
- Software infrastructure for dialect data processing