

Formalizing and enhancing the feature knowledge in dialect data

Vladimir Zhobov, Petya Osenova, Kiril Simov

Sofia team

BulDialects Project

22-23 April 2007

Tuebingen, Germany

Goal

- For each variant in the data to explicate the phonetic features
- We started with the list provided by Vladimir Zhobov
- A tagset for annotation of variants in the data
- The tags to be abstract and to allow easy comparison

Steps of Processing

Step 1: Initial list of categories and values (39 categories, 307 values)

Step 2: Annotation of the basic word forms (156)

Step 3: Annotation of all variants (4504)

Step 4: Enhancing of the initial list with new categories and values

Step 5: Decomposition of the values

Step 6: Tagset defined and final annotation

Problems

- Additional features (Secondary): *As vs. Az, Aze i Azi*
- How to present the missing features – one or two
- How to compare features

Usage

- Addition to transliteration
- Alternative to transliteration
- Base for vector presentation of the site

site01: (w1:f1:v11, w1:f2:v12, w2:f7:f21, . . .)

site02: (w1:f1:v13, w1:f2:v12, w2:f7:f22, . . .)