CENTRE FOR TRANSLATION STUDIES at the University of Leeds

Development and evaluation of phonological models for cognate identification

Bogdan Babych (b.babych@leeds.ac.uk)

- Using internal feature structures of characters
  - Current methods of cognate identification treat characters as atomic units (e.g., Levenshtein edit distance)
  - Engineering best-performing feature configuration for large-vocabulary cognate identification (rank candidates in >100k word space)

Graphemic-Phonological features
Uk: "*жовтий*" (*zhovtyj*) = 'yellow'

ж (zh) 'type:consonant', 'voice:ff-voiced', 'maner:ff-fricative', 'active:ff-fronttongue', 'passive:ff-palatal'
о (o) **'type:vowel'**, **'backness:back'**, **'height:mid'**, **'roundedness:rounded'**, 'palate:nonpalatalizing'
в (v) **'type:consonant'**, 'voice:fl-voiced', 'maner:fl-fricative', 'active:fl-labial', 'passive:fl-bilabial'
т (t) 'type:consonant', 'voice:pf-unvoiced', 'maner:pf-plosive', 'active:pf-fronttongue', 'passive:pf-alveolar'
и (y) **'type:vowel'**, 'backness:front', **'height:closemid'**,**'roundedness:unrounded'**, **'palate:nonpalatalizing'**
й (j) 'type:consonant', 'voice:xm-sonorant', 'maner:xm-approximant','active:xm-midtongue', 'passive:am-palatal'

Feature representations for corresponding characters in Ru: "*жёлтый*" (*zheltyj*) = 'yellow'.

ё (io) **'type:vowel'**, **'backness:back'**, **'height:mid'**, **'roundedness:rounded'**, 'palate:palatalizing'
л (l) **'type:consonant'**, 'voice:lf-sonorant', ' maner:lf-lateral', 'active:lf-fronttongue', 'passive:lf-alveolar'

ы (y) **'type:vowel'**, 'backness:central', **'height:closemid'**,**'roundedness:unrounded'**, **'palate:nonpalatalizing'**

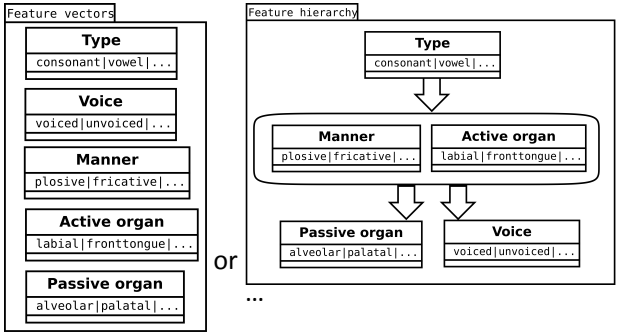
Calculation of the Phonological Levenshtein for Uk "*жовтий*" (*zhovtyj*) = 'yellow' and Ru "жёлтый" (zheltyi) = 'yellow':

0.0	1.0	2.0	3.0	4.0	5.0	6.0
1.0	**0.0**	1.0	2.0	3.0	4.0	5.0
2.0	1.0	**0.2**	1.2	2.2	3.2	4.2
3.0	2.0	1.2	**1.0**	2.0	3.0	4.0
4.0	3.0	2.2	2.0	**1.0**	2.0	3.0
5.0	4.0	3.2	3.0	2.0	**1.2**	2.2
6.0	5.0	4.2	4.0	3.0	2.2	**1.2**

cf.: Metric calculated for Uk "*жовтий*" (*zhovtyj*) = 'yellow' with Ru "*жуткий*" (*zhutkij*) = 'dismal':

0.0	1.0	2.0	3.0	4.0	5.0	6.0
1.0	**0.0**	1.0	2.0	3.0	4.0	5.0
2.0	1.0	**0.2**	1.2	2.2	3.2	4.2
3.0	2.0	1.2	**1.0**	1.2	2.2	3.2
4.0	3.0	2.2	2.0	**1.8**	2.2	3.0
5.0	4.0	3.2	3.0	2.8	**2.0**	3.0
6.0	5.0	4.2	4.0	3.8	3.0	**2.0**

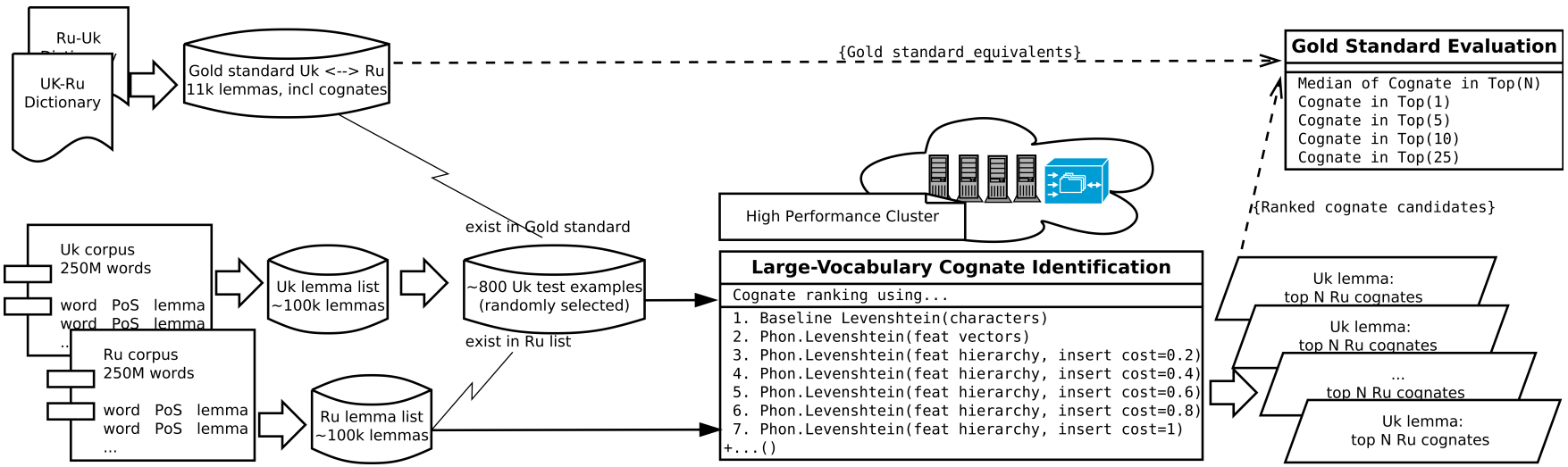
Optimal feature topology… ?



… optimal cost for insertion, deletion … ?

CENTRE FOR TRANSLATION STUDIES at the University of Leeds

Automated evaluation framework for phonological feature engineering



Evaluation results

Best performance for the settings:

- Feature hierarchy +
- Insertion/deletion = 0.8
- Improvement over baseline 'atomic' character Levenshtein distance
  - +20% median cognate rank
  - +16.5% cognate in top 1
  - + 3% cognate in top 5

Experiment	Median top N	Top 1	Top 5	Top 10	Top 25
Baseline Char.Lev	50	206	328	360	382
Phon.Lev FeatVectors	87.5	**215**	289	319	349
Diff with Base L	-75%	+4.4%	-10%	-11%	-9%
Phon.Lev Hierarchy:					
Phon.LevH i=0.2	125.5	216	291	315	342
Phon.LevH i=0.4	54.5	230	307	334	367
Phon.LevH i=0.6	*48*	235	328	354	385
Phon.LevH i=0.8	**40**	**240**	**337**	**359**	**391**
Phon.LevH i=1.0	47.5	**240**	334	359	385
Best Improv. over BaseL	**+20%**	**+16.5%**	**+3%**	**0%**	**+2%**