

Identifying dialectal features of the Udmurt language with the help of an internet corpus

Выявление диалектных особенностей удмуртского языка
при помощи интернет-корпуса

Timofey Arkhangelskiy

Universität Hamburg / Alexander von Humboldt-Stiftung

timarkh@gmail.com

Udmurt language

- Uralic family, Permic branch
- Udmurtia and neighboring regions
- 340,000 speakers
- Standard literary language; 4 main dialectal areas



Corpus

- Collection of texts
- Linguistic annotation:
 - metadata
 - lemmatization, morphological annotation
 - any other kind of annotation (e.g. borrowings)
- Search engine
 - corpus \neq library
 - corpus \neq Yandex/Google

Udmurt vk-corpus

- Posts and comments of Udmurt-language Vkontakte groups and users
- 2.5 million tokens in Udmurt (400 groups, 2000 users)
- Sentence-level language recognition (rus/udm), morphological annotation
- Author-related metadata: sex, birth year, birth place, current location

Udmurt vk-corpus

Мон бы пукысал али и кылзйськысал Лариса
Васильевнаез, сое можно кылзыны вечность.
Интерес не пропадёт. Тау та смена
понна котькудйзлы! Алиночка Владимировна, тон
прекрасной адями 😊

привет 😊 не надо грустить, Алёна. А вот лучше
малпаськы сессиед сярысь 😊

Алексей, 😊 точно

Udmurt vk-corpus

Мон **бы** пукысал али **и** кылзйськысал Лариса
Васильевнаез, сое **можно** кылзыны **вечность**.
Интерес не пропадёт. Тау та **смена**
понна котькудйзлы! Алиночка Владимировна, тон
прекрасной адями 😊

привет 😊 не надо грустить, Алёна. **А вот лучше**
малпаськы **сессиед** сярысь 😊

Алексей, 😊 точно

sentences in Russian

borrowed words / code switching within a sentence

Udmurt vk-corpora

- Web interface: search

Query

Word #1

Word: отстой*

Lemma:

Grammar:

Gloss:

Language/tier: Udmurt

Word #2

Word:

Lemma:

Grammar: N

Gloss:

Language/tier: Udmurt

Distance to word # 1

from 1

to 1

Full-text search: Precise match

Search sentences

Search words / lemmata



Select subcorpus



Udmurt vk-corporus

- Web interface: search results

Search result: 4 occurrences, 4 sentence(s) found in approximately 4 document(s).

inwis (group, 100-1000 members)

ой, Ирина, мынам но со **отстойной** сессия.



udmurt_ept (group, 100-1000 members)

Ту зэмзэ ке со **отстойной дыр** но(карликъёс ёвёл)



udmkenesh (group, 100-1000 members)

Эктон коркады но **отстойный фольклор**.



knyazpozdey (group, 1000-10000 members)

Для Удмурт кенеша, кстати, самой **отстойной уж** вал, кемалась ни малпаськом вал вераны, ухахахах)



Dialectology

- Phonetics
- Lexicon
- Morphology
- Syntax

*traditional
dialectology*

vk-corpus: phonetics

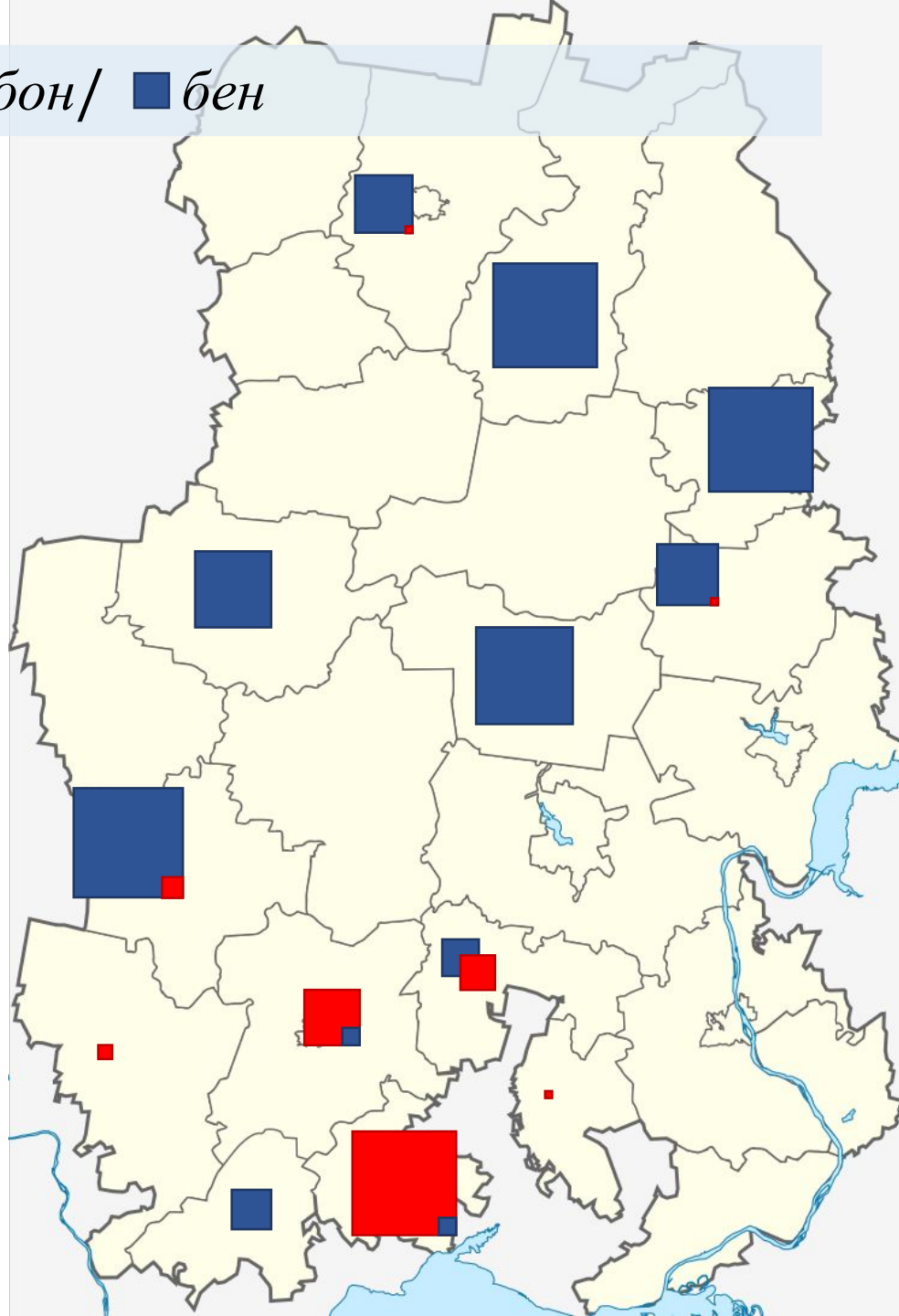
- People try not to deviate from the standard variety; orthography cannot reflect all dialectal features; the diacritics (*ü, ë, жс, ü, ö*) are often omitted



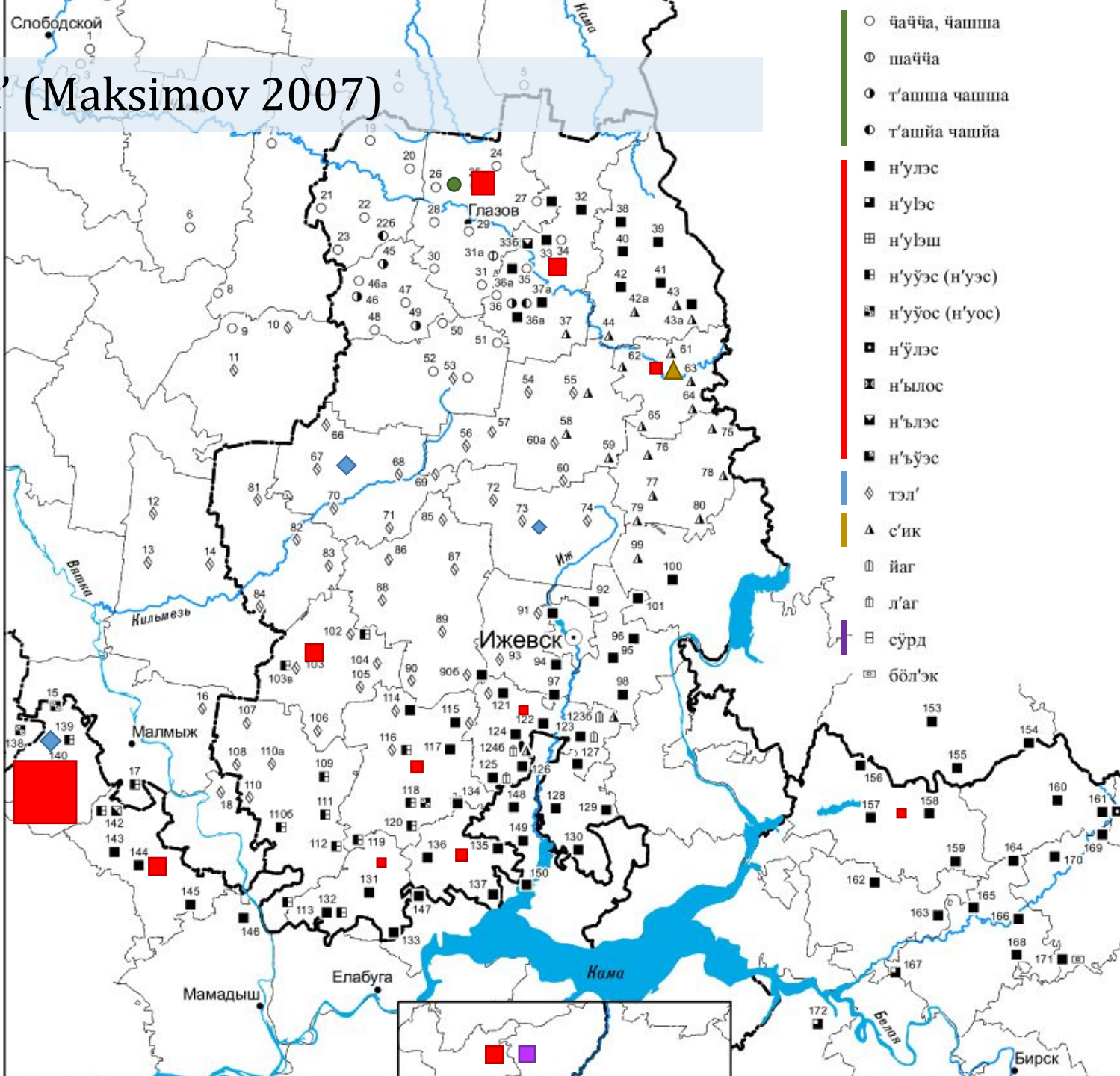
vk-corpus: lexicon

- Many people try to use the standard vocabulary
- Nevertheless, dialectal words show up quite often
- I have too few tokens for each of Udmurtia's 25 districts => only high-frequency vocabulary can be studied

Particle ■ *бон/* ■ *бен*



'Forest' (Maksimov 2007)



Подорожник (Максимов 2013)



Borrowed Russian verbs

- The standard way of borrowing a Russian verb is to use the construction $V_{inf} + [карыны]$:

Трос инты-ын снимать кар-о-м.

many place-LOC shoot.RUS do-FUT-1PL

‘We’re going to shoot [the movie] in many places.’

‘Мы будем снимать во многих местах.’

Borrowed Russian verbs

- There is a detransitivising suffix *-ськ-/-ск-* in Udmurt, which semantically is very close to the Russian suffix *-ся*:
 - passive
 - impersonal modal passive
 - generic subject/object
 - autocausative
 - reflexive
 - reciprocal

Borrowed Russian verbs

- If a reflexive Russian verb is borrowed:
 - either the light verb *карыны* has the *-сѣк-* suffix:

Кызъы дозвонитъся кар-исѣк-оно тѹ дор-ы. 😊😊😊😊

how reach.RUS do-DETR-DEB you.PL near-ILL

‘How can I reach you guys [by phone]?’

- or it does not:

со-ос ю-о, кыск-о, материтъся кар-о.

s/he-PL drink-PRS.3PL smoke-PRS.3PL swear.RUS do-PRS.3PL

‘They drink, smoke, swear.’

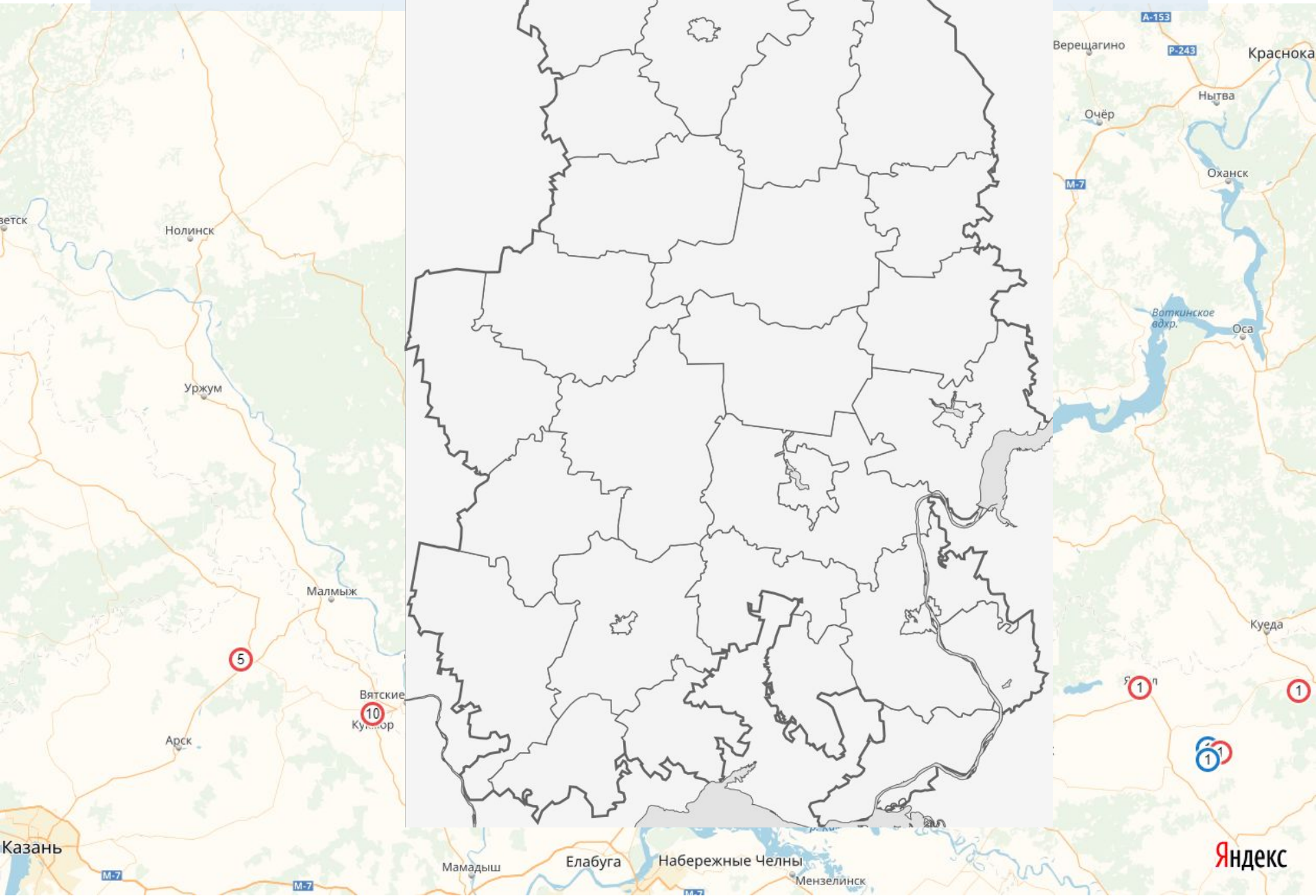
Borrowed Russian verbs

- Possible hypotheses regarding the distribution of the two variants:
 - lexical (depends on the verb)
 - depends on the meaning of the *-ся* suffix
 - depends on the aspect of the Russian verb
 - depends on the form of *карины*
 - random

Borrowed Russian verbs

- Possible hypotheses regarding the distribution of the two variants:
 - ~~lexical~~: same verbs often occur in both constructions
 - ~~depends on the meaning of -ся~~: no correlation
 - ~~depends on the aspect~~: no correlation; btw, the aspect is not always chosen according to Russian rules
 - ~~depends on the form of карыны~~: no correlation
 - ~~random~~: no, because people tend to consistently use only one of the strategies

Russian verbs: к



Borrowed Russian verbs

- The choice is clearly geographically conditioned
- The detransitive-less strategy prevails on the territory of the neighboring Tatarstan and Bashkortostan regions
- The light verb construction for verbal borrowings is exactly the same in Tatar and Bashkir (therefore, contact influence may be the driving force behind this distribution)

Conclusion

- An internet corpus can provide the data for identifying dialectal features
- The phonetic differences are almost impossible to extract from such a corpus
- Lexical features can be identified, provided the frequency is high enough
- Besides, interesting syntactic features can be identified (which is valuable, since the science does not know much about them)

Thank you for your attention!