Comparing the functional domain of linguistic elements across languages: ParaViz and the place of R

Ruprecht von Waldenfels, ruprecht.waldenfels@gmail.com

Instute of the Polish Language, Polish Academy of Sciences
ruprecht.waldenfels@gmail.com

Spotkanie miłośników R w Krakowie - May 13th 2015

Outline

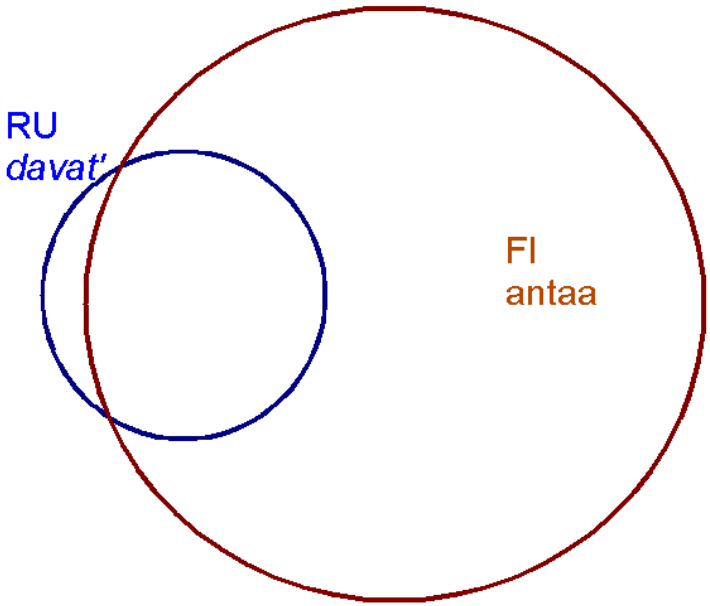
Outline

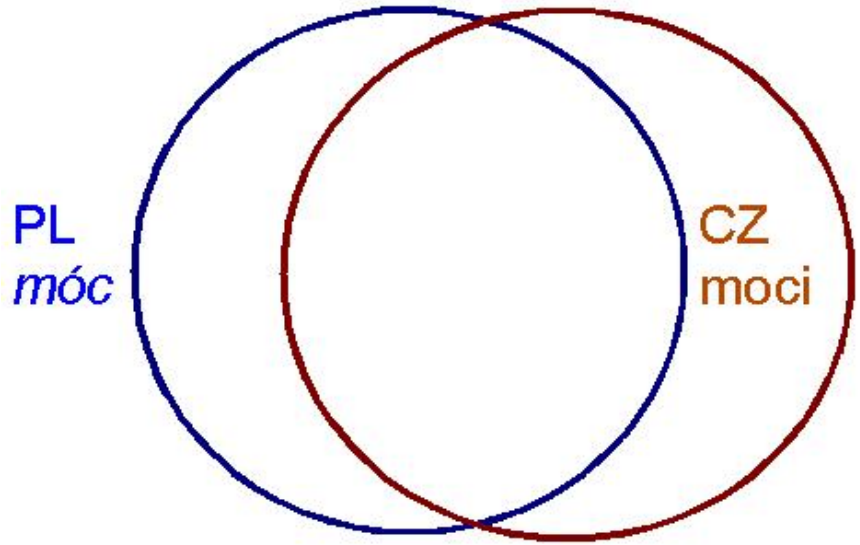
Usage-based approach

- parallel corpora as a tool to compare **functional domains** of comparable elements bottom-up, starting from actual usage

Usage-based approach

- parallel corpora as a tool to compare **functional domains** of comparable elements bottom-up, starting from actual usage





Outline

UNIVERSITÄT BERN

ParaSol: A Parallel Corpus of Slavic and other languages

Overview

The ParaSol, formerly known as the Regensburg Parallel Corpus (RPC), is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages, developed jointly by Ruprecht von Waldenfels and Roland Meyer, at the Institute of Slavic Languages and Literatures, University of Bern and the Institute of Slavistics, Regensburg University, respectively. We gratefully acknowledge support by research assistants and server hosting at Regensburg University.
The corpus is in the process of being reorganized. Many new texts have been added, and the corpus has been rechristened. Hopefully, there will be a full relaunch with many improvement towards the end of 2009.

The corpus is "work in progress" and open in terms of languages as well as texts. We ask interested linguists using this resource to contribute by adding new texts or languages.

Public access to the corpus is provided via a public web interface (beta stage).
Here's a short introduction in German and in English on how to use it, as well as the list of texts.
In order to register for use, please fill in this form.

A short outline of the corpus

- ParaSol focuses on
  - post-war belletristic texts (but also contains some legal and journalistic texts)
  - Slavic languages, but not exclusively: English and German are also included, and we hope to add more languages as the corpus grows.
  - texts that are translated into many Slavic languages, so that subsequent addition of further translations of can build on already included translations
  - for more information, see the list of texts and languages currently included in the corpus.

UNIVERSITÄT BERN

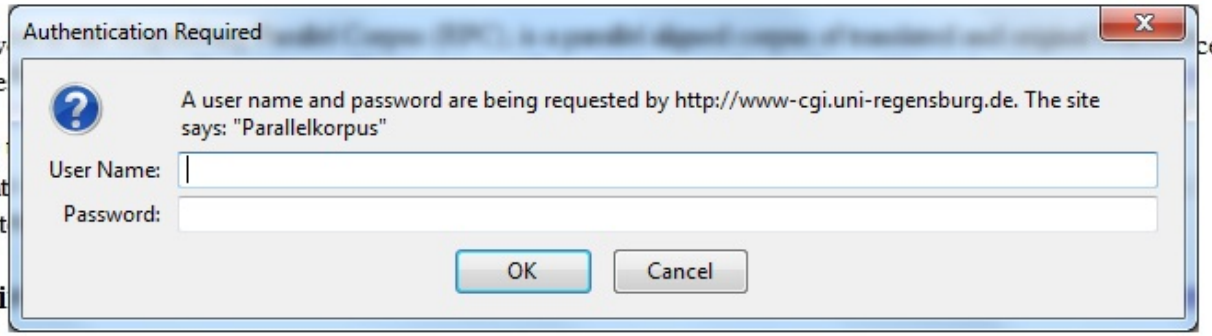
ParaSol: A Parallel Corpus of Slavic and other languages

Overview

The ParaSol, formerly known as the Regensburg Parallel Corpus (RPC), is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages, developed jointly by Ruprecht von Waldenfels and Roland Meyer, at the Institute of Slavic Languages and Literatures, University of Bern and the Institute of Slavistics, Regensburg University, respectively. We gratefully acknowledge support by research assistants and server hosting at Regensburg University.
The corpus is in the process of being reorganized. Many new texts have been added, and the corpus has been rechristened. Hopefully, there will be a full relaunch with many improvement towards the end of 2009.

The corpus is "w ce to contribute by adding new texts or language

Public access to
Here's a short int
In order to regist



A short outli

- ParaSol focuses on
  - post-war belletristic texts (but also contains some legal and journalistic texts)
  - Slavic languages, but not exclusively: English and German are also included, and we hope to add more languages as the corpus grows.
  - texts that are translated into many Slavic languages, so that subsequent addition of further translations of can build on already included translations
  - for more information, see the list of texts and languages currently included in the corpus.

Advanced querying

On this page, first specify one primary and several other languages and then choose subcorpora (texts) and queries.
You *need* to specify a query on the primary language.
You *may* also define queries on the other languages, which apply *in addition* to the query on the primary language.

For example, choosing pl (Polish) as primary and de (German), ru (Russian) and bx (BCS) as additional languages, a query on pl *[word="nigdy"]* will "nigdy" as well as all their translations in German, Russian and Bosnian/Croatian/Serbian.
An *additional query* over BCS *[word="nikad"]* will constrain the search results to (a) include only those segments that contain *"nigdy"* in the Polish ve
Additional queries may be negated: Using *![word="nikad"]* as an additional query will result in all segments that (a) contain *"nigdy"* in the Polish and

Primary language:

Slavonic				Germanic	Romance		Baltic	Others
○ BG	○ SRA	○ PLA	◉ RU	○ NL	○ FR	○ ES	○ LV	○ EO
○ HR	○ SL	○ SK	○ RUA	○ EN	○ IT		○ LT	○ EL
○ MK	○ CZ	○ US	○ UK	○ DE	○ PT			○ HU
○ SR	○ PL	○ BY		○ DEA	○ RO			

Further languages:

Slavonic				Germanic	Romance		Baltic	Others
☑ BG	☐ SRA	☑ PLA	☑ RU	☐ NL	☐ FR	☐ ES	☐ LV	☐ EO
☑ HR	☑ SL	☑ SK	☑ RUA	☐ EN	☐ IT		☐ LT	☐ EL
☑ MK	☑ CZ	☑ US	☐ UK	☐ DE	☐ PT			☐ HU
☑ SR	☑ PL	☐ BY		☐ DEA	☐ RO			

◉ **All textes** ○ **Only textes available in all languages**

	bg	hr	mk	sr	sl	cz	pl	pla	sk	us	ru	rua
☑ **lempowgwiazd**	☑						☑				☑	
☑ **sueskindparfuem**	☑	☑		☑	☑	☑	☑		☑		☑	
☑ **lemwizjalokalna**							☑				☑	
☑ **lemglospana**				☑		☑	☑				☑	
☑ **endemomo**							☑				☑	
☑ **potter4**							☑				☑	
☑ **lemastronauci**						☑	☑				☑	
☑ **bulgakovmaster**	☑	☑	☑	☑	☑	☑	☑	☑	☑		☑	
☑ **ostrovskijstal**	☑	☑	☑	☑	☑	☑	☑		☑	☑	☑	
☑ **pelevincapaev**	☑						☑				☑	
☑ **pavichazar**	☑		☑	☑			☑		☑		☑	
☑ **lempamwannie**						☑	☑				☑	
☑ **strugpiknik**	☑					☑	☑		☑		☑	
☑ **lemfiasko**				☑		☑	☑				☑	
☑ **gombrowiczferdyduke**							☑				☑	
☑ **lemsolaris**	☑	☑		☑		☑	☑				☑	☑

Russian	[tag="Vmm.*" & lemma =".*писать"]
Slovene	
Bulgarian	
Croatian	
Macedonian	
Serbian	
Slovak	
Russian a	
Polish a	
Upper Sorbian	
Czech	

34794 Ему приходится читать мне речь три - четыре раза своим благозвучным баритоном , пока она не приобретет подлинную убедительность .	Musí mi to nahlas prečítať tri , štyri razy , aby jeho plnozvučný baryt
42334 Здесь ежедневно предлагают на выбор три меню , внизу в баре сидят жиголо , которые готовы потанцевать со мной , но я больше не танцую .	Tri razy denne sa ponúka jedlo , dolu v bare sedia gigolovia , a keby netancujem - a večer prechádzajú z lesa cez lúku srnčiatka .
61523 точно зная , что , потребовав сто процентов , получит только сорок два , он , когда ему дают сорок три с половиной , трубит победу , ибо не подозревает , что ему с легкостью дали бы и сорок восемь и что основания для торжества имеет не он , а другая сторона .	A on by mohol vytvoriť kompromis , veď presne viem , keď požaduj dajú 43 1 // 2 a on triumfuje , lebo nevie , že by mu boli dali aj 48 , triumfovať , oni .

38 hits in corpus
bulgakovmaster.

	sl	sr	bg	mk	hr
2550 - Взять бы этого Канта , да за такие доказательства года на три в Соловки !	- » Tega Kanta bi bilo treba za take dokaze poslati za tri leta na Solovke !	- Tog Kanta bi za te i takve dokaze trebalo najuriti na jedno tri godine u Solovke !	— Тоя Кант за подобни доказателства може ли да го приберат за две - три годинки в Соловки !	- Да го земеме тој Кант и за таквите докази да го испратиме на едно три години во Соловки !	– Trebalo bi tog za takve dokaze godine u Solovk
13674 За ним по три в ряд полетели всадники в туче пыли , запрыгали кончики легких бамбуковых пик , мимо прокуратора понеслись казавшиеся особо смуглыми под белыми тюрбанами лица с весело оскаленными , сверкающими зубами .	Po trije v vrsti so zdirjali za njim njegovi konjeniki , zaviti v oblak prahu , konice njihovih lahkih bambusovih sulic so poskakovale , in mimo prokuratorja so se podili obrazi , ki so se zdeli še posebno temni pod belimi turbani , z veselo kažočimi se lesketajočimi zobmi .	- - Za njim - po tri u redu - projuriše konjanici u oblaku prašine , promakoše vrhovi bambusovih kopalja , kraj prokuratora proleteše izrazito tamna lica pod belim turbanima , veselo iskeženih belih zuba .	- Зад него по трима в редица полетяха сред облак прах конниците , заподскачаха върхчетата на леките бамбукови пики , покрай прокуратора се понесоха лица с весело оголени блестящи зъби , под белите чалми тези лица изглеждаха още по - мургави .	По него - по тројца во ред - летнаа и коњаниците в облак прав , заиграа врвовите на лесните бамбусови копја , покрај прокураторот пројурија , со весело светната соголени заби , лица кои во своите бели турбани изгледаа особено црномурести .	Za njim su u ob trojica u redu – konjanici , posk laganih bambus mimo prokurato lica koja su se p turbanima činila tamnoputa , s v blistavim zubim
17076 Пропустив мимо себя все три вагона , кот вскочил на заднюю дугу последнего , лапой вцепился в какую - то кишку , выходящую из стенки , и укатил , сэкономив , таким образом , гривенник .	Pustil je mimo sebe vse tri voze , skočil na zadnji konec zadnjega , zasadil kremplje v neko cev , ki je štrlela iz stene , in se odpeljal , desetico pa si je prihranil .	Propustivši najpre sva tri vagona , on skoči na papučicu poslednjeg , uhvativši se šapom za nekakvo crevo koje je virilo iz tramvaja , i odveze se , uštedevši tako deset kopejki .	Той изчака да отминат и трите вагона и скочи на задния буфер на последния , вкопчи се в някакъв маркуч , който се подаваше от трамвая , и офейка , спестил по този начин десет копейки .	Пролуштајќи ги крај себе трите вагони , мачорот скокна на задната платформа на последниот , се закачи за некое си црево што се подаваше од ѕидот и си замина , заштедувајќи си го на тој начин гривеникот .	Propustivši kraj vagona , mačak stražnju željezn posljednjem vag uhvatio za neku njega virila , i o uštedjevši na ta kopjejaka .
20173 Дач всего двадцать две , и строится еще только семь , а нас в " МАССОЛИТе " три тысячи .	Vil je samo dvaindvajset in postavljajo jih le še sedem , nas v Masolitu pa je tri tisoč .	Vikendica ima samo dvadeset i dve , gradi se još sedam , a nas je u MASSOLIT - u tri hiljade .	Вилите са всичко на всичко двайсет и две , строят се само седем нови , а в МАССОЛИТ сме три хиляди .	Има се на се дваесет и две вили и се градат уште само седум , а во МАССОЛИТ не има три илјади .	Ljetnikovaca im i dva , a gradi s sedam , a nas j tri tisuće .
23478 - Мне двадцать три года , - возбужденно заговорил Иван , - и я подам жалобу на вас всех .	« » Triindvajset let imam , « je razburjeno spregovoril Ivan , » in tožil vas bom vse skupaj .	- Imam dvadeset i tri godine - uzbuđeno poče Ivan - i uložiću žalbu protivu svih vas .	— На двайсет и три години съм — възбудено заговори Иван — и ще напиша оплакване срещу всички вас .	- Имам дваесет и три години ,- возбудено почна Иван ,- и јас ќе ве тужам сите вас .	– Meni su dvade uzbuđeno je pro žalit ću se na sv
26699 Анна Францевна де Фужере , пятидесятилетняя почтенная и очень деловая дама , три комнаты из пяти сдавала жильцам :	- Ana Francevna de Fougere , petdesetletna spoštovana in zelo podjetna dama , je tri sobe od petih oddajala najemnikom : nekomu , ki se je menda pisal Belomut , in nekomu drugemu ,	Ana Francevna de Fužere , pedesetogodišnja , poštovanja dostojna i uza sve to veoma poslovna dama , tri od pet soba iznajmljivala je stanarima :	Ана Францевна дьо Фужере , петдесетгодишна почтена и много делова дама , даваше три от общо петте си стаи под наем на някакъв човек , който се казваше , струва ми се ,	Ана Францевна де Фужере , педесетгодишна дама , почитувана и многу деловна , трите соби од петте ги издаваше на потстанари :	Ana Francevna pedesetogodišn poslovna dama tri od pet soba

State of corpus

LNG	in full	tokens	lemmas	tags
BG	Bulgarian	2 002 697	43 280	y
BY	Belarusian	482 467	24 131	y
CZ	Czech	1 629 868	47 166	y
DA	Danish	100 448	0	n
DE	German	2 006 781	64 112	y
EE	Estonian	287 948	23 031	y
EL	Greek	600 594	40 141	y
EN	English	814 289	19 886	y
EO	Esperanto	152 660	0	n
ES	Spanish	476 301	21 414	y
FI	Finnish	174 204	0	n
FR	French	448 612	10 870	y
HR	Croatian	899 466	43 263	y
HU	Hungarian	146 505	0	n
HY	Armenian	240 815	7 873	y
IT	Italian	478 315	19 472	y
LT	Lithuanian	280 675	8 001	y
LV	Latvian	147 906	0	n
MK	Macedonian	1 045 873	43 646	y
NL	Dutch	728 061	4 836	y
NO	Norwegian	334 948	13 788	y
PL	Polish	3 396 673	66 492	y
PT	Portuguese	380 659	10 961	y
RO	Romanian	398 088	15 164	y
RU	Russian	3 637 357	78 997	y
SK	Slovak	1 457 925	51 010	y
SL	Slovene	1 132 839	36 229	y
SR	Serbian	1 324 929	42 602	y
SV	Swedish	314 759	0	n
UK	Ukrainian	1 017 054	33 562	y
US	Upper Sorbian	73 266	0	n

32 languages, >400 pairs, 25 mio token
In all major Slavic languages:

- Michail Bulgakov: Master i Margarita
- Stanisław Lem: Solaris
- Umberto Eco: Il nome della rosa
- Patrick Süskind: Das Parfüm
- Nikolaj Ostrovskij: Kak zakaljalos' stal'
- Joanne K. Rowling: Harry Potter and the Sorcerer's Stone
- Ivo Andrić: Na Drini ćuprija
- Milan Kundera: Nesnesitelná lehkost bytí

`http://www.parasolcorpus.org`

Outline

Outline

Prepositions: K 'to', DO 'to', V 'in(to)'

(1) 'Ivan [. . .] could not help himself and, on seeing the water gush **into the tub** in a wide stream from the gleaming faucet, said ironically:'

ru	Иван [. . .] не удержался и, видя, как вода хлещет в ванну широкой струей из сияющего крана, сказал с иронией:
by	. . . шырокім струменем прэ ў ванну вада . . .
uk	. . . як вода рине у ванну широким струменем . . .
uka	. . . як вода плеще до ванни широким струменем . . .
pl	. . . obserwując wodę [. . .] lejącą się [. . .] do wanny . . .
pla	. . . woda szeroko chlusta ze lśniącego kranu do wanny . . .
cz	. . . který se řinul z nablýskaného kohoutku do vany . . .
sk	. . . širokým prúdom tečie do vane voda . . .
sl	. . . kako se voda v širokem curku zliva [. . .] v kad . . .
hr	. . . iz blistave pipe voda u širokoj struji teče u kadu . . .
sr	. . . voda juri u kadu širokim mlazom iz blistave slavine . . .
sra	. . . kako se voda iz blistave slavine [. . .] izliva u kadu . . .
bg	. . . как водата плющи във ваната на широка струя . . .

Prepositions: K 'to', DO 'to', V 'in(to)'

(2) 'Aphranius bowed, moved the chair closer **to the bed**, and sat down'

ru	Афраний поклонился, пододвинул кресло поближе к кровати и сел
by	Афраній пакланіўся, падсунуў крэсла бліжэй да ложа і сеў
uka	Афраній вклонився, підсунув крісло ближче до ліжка й сів
uk	Афраній уклонився, підсунув крісло ближче до ложа і сів
pla	Afraniusz skłonił się, przysunął fotel ku posłaniu i usiadł [. . .].
pl	Afraniusz skłonił się, przysunął tron **bliżej łoża** i usiadł [. . .].
cz	Afranius se uklonil, přisunul křeslo blíž k lůžku a usedl [. . .].
sk	Afránius sa uklonil, pritiahol si kreslo bližšie k lôžku a sadol si
sl	Afranij se je priklonil, pomaknil naslanjač bliže k postelji in sedel
hr	Afranije se poklonio, pritegao naslonjač bliže **krevetu** i sjeo [. . .].
sr	Afranije se pokloni, primače fotelju bliže **krevetu** i sede [. . .].
mk	Афраниј се поклони, ја намести фотељата поблиску до креветот и седна
bg	Афраний се поклони, премести креслото си по-близо до ложето и седна

Prepositions: K 'to', DO 'to', V 'in(to)'

(3) 'Styopa brought the glass **to his lips**'

ru	**K**	Прыгающей рукой поднес Степа стопку к устам, [. . .].
by	**D**	Дрыжачай рукою паднёс Сцёпа гарэлку да вуснаў, [. . .].
uk	**D**	Непевною рукою підніс Стьопа до вуст чарку, [. . .].
uka	**D**	Тремтячою рукою підніс Стьопа стопку до вуст, [. . .].
pl	**D**	Drżącą ręką Stiopa podniósł kieliszek do ust, [. . .].
pla	**D**	Stiopa podniósł rozdygotaną ręką szklaneczkę do ust, [. . .].
cz	**K**	Lotrov rozechvělou rukou pozvedl sklenici k ústům [. . .].
sk	**K**	Šťopa roztrasenou rukou zdvihol k ústam pohárik [. . .].
sl	**K**	S tresočo se roko je Stjopa ponesel kozarec k ustom, [. . .].
hr	-	Stjopa je drhtavom rukom prinio čašicu **ustima**, [. . .].
sr	-	Drhtavom rukom Stjopa prinese čašicu **ustima**, [. . .].
sra	**D**	Со растреперена рака Стјопа ја доближи чашата до уста
bg	**K**	Стьопа поднесе с трепереща ръка чашата към устата си

Outline

Word aligned corpus

```
<w id="7.7" lem="жаркий" pos="Afpmsg">жаркого<waligns>
      <walign lng="bg">
         <w id="8.9" lem="горещ" pos="A---:sm0">горещ</w>
      </walign>
      <walign lng="by">
         <w id="7.5" lem="надзвычай" pos="ADV,norm">надзвычай</w>
         <w id="7.6" lem="душны" pos="A,norm,nom,sg,m:A,norm,acc,sg,m,inan">душны</w>
      </walign>
      <walign lng="cz">
         <w id="4.2" lem="dusný" pos="AAIS6----1A----">dusném</w>
         <w id="4.3" lem="jarní" pos="AAIS6----1A----">jarním</w>
      </walign>
      <walign lng="hr">
         <w id="7.7" lem="neobičan" pos="Rnp">neobično</w>
         <w id="7.8" lem="topal" pos="Afpmsgy">toplog</w>
      </walign>
      <walign lng="mk">
         <w id="2.5" lem="пролетен" pos="A">пролетно</w>
      </walign>
      <walign lng="pl">
         <w id="3.5" lem="wiosenny" pos="adj:sg:nom:n:pos">wiosenne</w>
      </walign>
      <walign lng="pla">
         <w id="8.8" lem="upalny" pos="adj:sg:gen:m3:pos">upalnego</w>
      </walign>
      <walign lng="sk">
         <w id="6.1" lem="teplý" pos="Afpnsn">Teplé</w>
      </walign>
```

Parameter file: V

```
<type id="V" name="V">
        <criteria><lng>ru</lng><regexp level="preclem">^во?$</regexp> <regexp level="ta
        <criteria><lng>uk</lng><regexp level="preclem">^[ву]$</regexp><regexp level="ta
        <criteria><lng>by</lng><regexp level="preclem">^[ыў]$</regexp><regexp level="ta
        <criteria><lng>mk</lng><regexp level="preclem">^во?$</regexp>  <regexp level="t
        <criteria><lng>bg</lng><regexp level="preclem">^вь?в?$</regexp> <regexp level="
        <criteria><lng>pl</lng><regexp level="preclem">^we?$</regexp> <regexp level="ta
        <criteria><lng>cz</lng><regexp level="preclem">^ve?(-\d)?$</regexp> <regexp lev
        <criteria><lng>sk</lng><regexp level="preclem">^vo?$</regexp> <regexp level="ta
        <criteria><lng>sl</lng><regexp level="preclem">^v$</regexp>   <regexp level="ta
        <criteria><lng>hr</lng><regexp level="preclem">^u$</regexp>   <regexp level="ta
        <criteria><lng>sr</lng><regexp level="preclem">^u$</regexp>   <regexp level="ta
        <criteria><lng>de</lng><regexp level="preclem">^(in|im)s?$</regexp><regexp leve
        <criteria><lng>lt</lng><regexp level="preclem">^į$</regexp><regexp level="tag">
        <criteria><lng>us</lng><regexp level="precword">^[Ww]e?$</regexp>   </criteria>
</type>
```

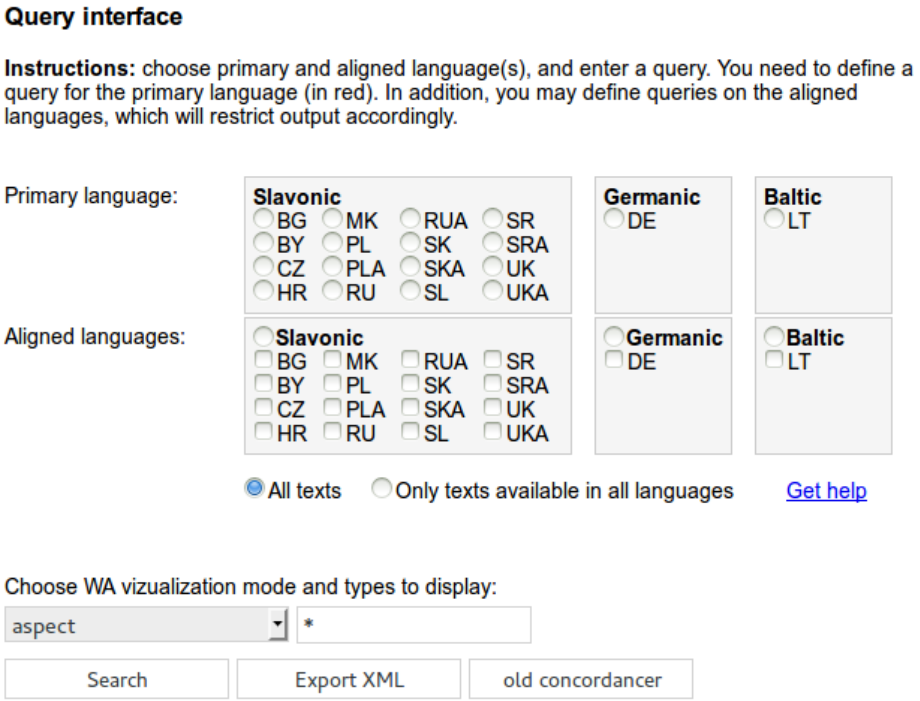
levels: token, lemma, tag; nontoken, nonlemma, nontag
for actual and 2 preceding corpus positions

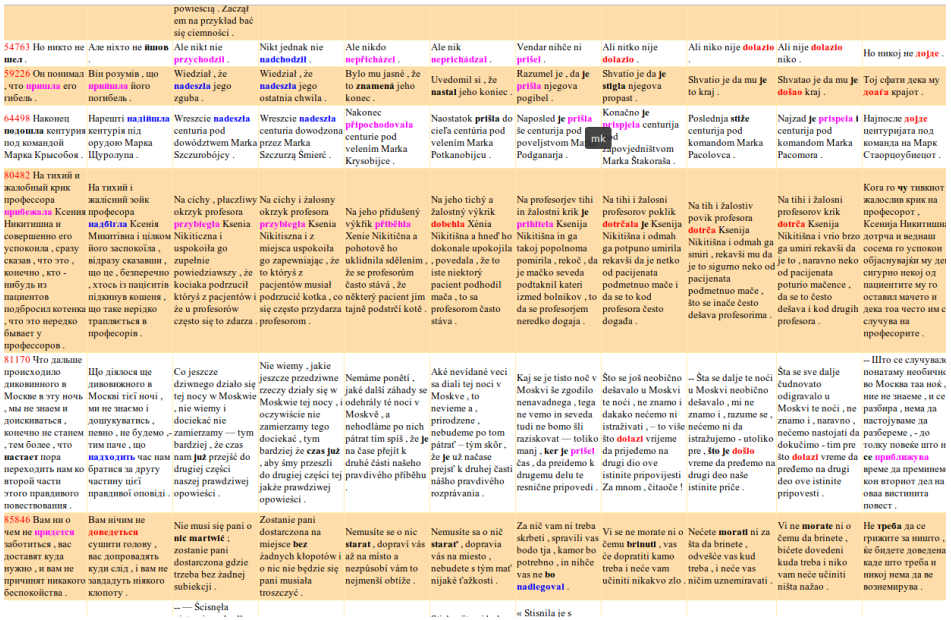
Parameter file: V with case

```
<type id="V" name="V">
  <criteria><lng>mk</lng><regexp level="preclem">^во?$</regexp>  <regexp level="tag">^[Nn].*</regexp>    </criteria>
  <criteria><lng>bg</lng><regexp level="preclem">^вь?в?$</regexp> <regexp level="tag">^N.*</regexp>    </criteria>
  <criteria><lng>de</lng><regexp level="preclem">^(in|im)s?$</regexp><regexp level="tag">^NN$</regexp>   </criteria>
  <criteria><lng>us</lng><regexp level="precword">^[Ww]e?$</regexp>   </criteria>
</type>

<type id="v" name="VACC">
  <criteria><lng>ru</lng><regexp level="preclem">^во?$</regexp> <regexp level="tag">^N...[ga].*</regexp>   </criteria>
  <criteria><lng>uk</lng><regexp level="preclem">^[ву]$</regexp><regexp level="tag">^S.*(acc|gen),.*</regexp>   </crite
  <criteria><lng>by</lng><regexp level="preclem">^[уў]$</regexp><regexp level="tag">^S.*(acc|gen),.*</regexp>   </crite
  <criteria><lng>pl</lng><regexp level="preclem">^we?$</regexp> <regexp level="tag">^subst.*(gen|acc).*</regexp></crite
  <criteria><lng>cz</lng><regexp level="preclem">^ve?(-\d)?$</regexp> <regexp level="tag">^N...(2|4).*</regexp>   </cri
  <criteria><lng>sk</lng><regexp level="preclem">^vo?$</regexp> <regexp level="tag">^N...[ga].*</regexp>   </criteria>
  <criteria><lng>sl</lng><regexp level="preclem">^v$</regexp>   <regexp level="tag">^S...[rt].*</regexp>   </criteria>
  <criteria><lng>hr</lng><regexp level="preclem">^u$</regexp>   <regexp level="tag">^N...[ga].*</regexp>   </criteria>
  <criteria><lng>sr</lng><regexp level="preclem">^u$</regexp>   <regexp level="tag">^N...[ga].*</regexp>   </criteria>
</type>

<type id="w" name="VLOC">
  <criteria><lng>ru</lng><regexp level="preclem">^во?$</regexp> <regexp level="tag">^N...[dl].*</regexp>   </criteria>
  <criteria><lng>uk</lng><regexp level="preclem">^[ву]$</regexp><regexp level="tag">^S.*(dat|loc),.*</regexp>   </crite
  <criteria><lng>by</lng><regexp level="preclem">^[уў]$</regexp><regexp level="tag">^S.*(dat|loc),.*</regexp>   </crite
  <criteria><lng>pl</lng><regexp level="preclem">^we?$</regexp> <regexp level="tag">^subst.*(dat|loc).*</regexp></crite
  <criteria><lng>cz</lng><regexp level="preclem">^ve?(-\d)?$</regexp> <regexp level="tag">^N...(3|6).*</regexp>   </cri
  <criteria><lng>sk</lng><regexp level="preclem">^vo?$</regexp> <regexp level="tag">^N...[dl].*</regexp>   </criteria>
  <criteria><lng>sl</lng><regexp level="preclem">^v$</regexp>   <regexp level="tag">^S...[dm].*</regexp>   </criteria>
  <criteria><lng>hr</lng><regexp level="preclem">^u$</regexp>   <regexp level="tag">^N...[dl].*</regexp>   </criteria>
  <criteria><lng>sr</lng><regexp level="preclem">^u$</regexp>   <regexp level="tag">^N...[dl].*</regexp>   </criteria>
</type>
```

Outline

Color-coded corpus samples





- Based on ParaVoz https://bitbucket.org/rvwfels/paravoz, interface to CWB-encoded parallel corpus,
- CWB output transformed to XML, XSLT for classification of result XML

Color-coded corpus samples

		powieścią . Zaczął em na przykład bać się ciemności .								
54763 Но никто не **шел** .	Але ніхто не **йшов** .	Ale nikt nie **przychodził** .	Nikt jednak nie **nadchodził** .	Ale nikdo **nepřicházel** .	Ale nik **neprichádzal** .	Vendar nihče ni **prišel** .	Ali nitko nije **dolazio** .	Ali niko nije **dolazio** .	Ali nije **dolazio** niko .	Но никој не **дојде** .
59226 Он понимал , что **пришла** его гибель .	Він розумів , що **прийшла** його погибель .	Wiedział , że **nadeszła** jego zguba .	Wiedział , że **nadeszła** jego ostatnia chwila .	Bylo mu jasné , že to **znamená** jeho konec .	Uvedomil si , že **nastal** jeho koniec .	Razumel je , da **je prišla** njegova pogibel .	Shvatio je da **je stigla** njegova propast .	Shvatio je da mu **je** to kraj .	Shvatao je da mu **je došao** kraj .	Тој сфати дека му **доаѓа** крајот .
64498 Наконец **подошла** кентурия под командой Марка Крысобоя .	Нарешті **надійшла** кентурія під орудою Марка Щуролупа .	Wreszcie **nadeszła** centuria pod dowództwem Marka Szczurobójcy .	Wreszcie **nadeszła** centuria dowodzona przez Marka Szczurzą Śmierć .	Nakonec **připochodovala** centurie pod velením Marka Krysobijce .	Naostatok **prišla** do cieľa centúria pod velením Marka Potkanobijcu .	Naposled **je prišla** še centurija pod poveljstvom Ma Podganarja .	Konačno **je prispjela** centurija od zapovjedništvom Marka Štakoraša .	Poslednja **stiže** centurija pod komandom Marka Pacolovca .	Najzad **je prispela** i centurija pod komandom Marka Pacomora .	Најпосле **дојде** центуријата под команда на Марк Стаорцоубиецот .
80482 На тихий и жалобный крик профессора **прибежала** Ксения Никитишна и совершенно его успокоила , сразу сказав , что это , конечно , кто - нибудь из пациентов подбросил котенка , что это нередко бывает у профессоров .	На тихий і жалісний зойк професора **надбігла** Ксенія Микитівна і цілком його заспокоїла , відразу сказавши , що це , безперечно , хтось із пацієнтів підкинув кошеня , що таке нерідко трапляється в професорів .	Na cichy , płaczliwy okrzyk profesora **przybiegła** Ksenia Nikiticzna i uspokoiła go zupełnie powiedziawszy , że kociaka podrzucił któryś z pacjentów i że u profesorów często się to zdarza .	Na cichy i żałosny okrzyk profesora **przybiegła** Ksenia Nikitiszna i z miejsca uspokoiła go zapewniając , że to któryś z pacjentów musiał podrzucić kotka , co się często przydarza profesorom .	Na jeho přidušený výkřik **přiběhla** Xenie Nikitična a pohotově ho uklidnila sdělením , že se profesorům často stává , že některý pacient jim tajně podstrčí kotě .	Na jeho tichý a žalostný výkrik **dobehla** Xénia Nikitišna a hneď ho dokonale upokojila , povedala , že to iste niektorý pacient podhodil mača , to sa profesorom často stáva .	Na profesorjev tihi in žalostni krik **je prihitela** Ksenija Nikitišna in ga takoj popolnoma pomirila , rekoč , da je mačko seveda podtaknil kateri izmed bolnikov , to da se profesorjem neredko dogaja .	Na tihi i žalosni profesorov poklik **dotrčala je** Ksenija Nikitišna i odmah ga potpuno umirila rekavši da je netko od pacijenata podmetnuo mače i da se to kod profesora često događa .	Na tih i žalostiv povik profesora **dotrča** Ksenija Nikitišna i odmah ga smiri , rekavši mu da je to sigurno neko od pacijenata podmetnuo mače , što se inače često dešava profesorima .	Na tihi i žalosni profesorov krik **dotrča** Ksenija Nikitišna i vrio brzo ga umiri rekavši da je to , naravno neko od pacijenata poturio mačence , da se to često dešava i kod drugih profesora .	Кога го **чу** тивкиот жалослив крик на професорот , Ксенија Никитишна дотрча и веднаш сосема го успокои објаснувајќи му дека сигурно некој од пациентите му го оставил мачето и дека тоа често им с случува на професорите .
81170 Что дальше происходило диковинного в Москве в эту ночь , мы не знаем и доискиваться , конечно не станем , тем более , что **настает** пора переходить нам ко второй части этого правдивого повествования .	Що діялося ще дивовижного в Москві тієї ночі , ми не знаємо і дошукуватись , певно , не будемо , тим паче , що **надходить** час нам братися за другу частину цієї правдивої оповіді .	Co jeszcze dziwnego działo się tej nocy w Moskwie , nie wiemy i dociekać nie zamierzamy — tym bardziej , że czas nam **już** przejść do drugiej części naszej prawdziwej opowieści .	Nie wiemy , jakie jeszcze przedziwne rzeczy działy się w Moskwie tej nocy , i oczywiście nie zamierzamy tego dociekać , tym bardziej że **czas już** , aby śmy przeszli do drugiej części tej jakże prawdziwej opowieści .	Nemáme ponětí , jaké další záhady se odehrály té noci v Moskvě , a nehodláme po nich pátrat tím spíš , že **je** na čase přejít k druhé části našeho pravdivého příběhu .	Aké nevídané veci sa diali tej noci v Moskve , to nevieme a , prirodzene , nebudeme po tom pátrať – tým skôr , že **je** už načase prejsť k druhej časti nášho pravdivého rozprávania .	Kaj se je tisto noč v Moskvi še zgodilo nenavadnega , tega ne vemo in seveda tudi ne bomo šli raziskovat — toliko manj , **ker je prišel** čas , da preidemo k drugemu delu te resnične pripovedi .	Što se još neobično dešavalo u Moskvi te noći , ne znamo i dakako nećemo ni istraživati , – to više što **dolazi** vrijeme da prijeđemo na drugi dio ove istinite pripovijesti Za mnom , čitaoče !	-- Šta se dalje te noći u Moskvi neobično dešavalo , mi ne znamo i , razume se , nećemo ni da istražujemo - utoliko pre , **što je došlo** vreme da pređemo na drugi deo naše istinite priče .	Šta se sve dalje čudnovato odigravalo u Moskvi te noći , ne znamo i , naravno , nećemo nastojati da dokučimo - tim pre što **dolazi** vreme da pređemo na drugi deo ove istinite pripovesti .	-- Што се случувал понатаму необично во Москва таа ноќ , ние не знаеме , и се разбира , нема да настојуваме да разбереме , - до толку повеќе што н **се приближува** време да преминем кон вториот дел на оваа вистинита повест .
85846 Вам ни о чем не **придется** заботиться , вас доставят куда нужно , и вам не причинят никакого беспокойства .	Вам нічим не **доведеться** сушити голову , вас допровадять куди слід , і вам не завдадуть ніякого клопоту .	Nie musi się pani o **nic martwić** ; zostanie pani dostarczona gdzie trzeba bez żadnej subiekcji .	Zostanie pani dostarczona na miejsce **bez** żadnych kłopotów i o nic nie będzie się pani musiała troszczyć .	Nemusíte se o nic **starat** , dopraví vás až na místo a nezpůsobí vám to nejmenší obtíže .	Nemusíte sa o nič **starať** , dopravia vás na miesto , nebudete s tým mať nijaké ťažkosti .	Za nič vam ni treba skrbeti , spravili vas bodo tja , kamor bo potrebno , in nihče vas ne **bo nadlegoval** .	Vi se ne morate ni o čemu **brinuti** , vas će dopratiti kamo treba i neće vam učiniti nikakvo zlo .	Nećete **morati** ni za šta da brinete , odvešće vas kud treba , i neće vas ničim uznemiravati .	Vi ne **morate** ni o čemu da brinete , bićete dovedeni kuda treba i niko vam neće učiniti ništa nažao .	Не **треба** да се грижите за ништо , ќе бидете доведена каде што треба и никој нема да ве вознемирува .
		-- — Ścisnęła				« Stisnila je s				

Outline

Aggregating over many examples

```
'Russian'     pipippipipiiiiiiiiipippipiiiiipipiiiipipiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppppippiipipppipiiii
'Czech'       ipiiiipippppiipiippppiiipppppppppipipii?iipipppppppppippppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipipippp?pipppppppippppppppipipippppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippippppipp?iippippppippipippppippppppipippppipipippp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiippiipiipi
```

Aggregating over many examples

```
'Russian'     pipippipipiiiiiiiiipippipiiiiipipiiiipipiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppppippiipipppipiiii
'Czech'       ipiiiipippppiipiippppiiipppppppppipipii?iipipppppppppippppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipipippp?pipppppppippppppppipipippppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippippppipp?iippippppippipippppippppppipippppipipippp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiippiipiipi
```

distances: for each pair, take proportion of differing symbols

Aggregating over many examples

```
'Russian'     pipippipipiiiiiiiiipippipiiiiipipiiiipipiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppppippiipipppipiiii
'Czech'       ipiiiipipppipiipippiiipppppppppipipii?iipipppppppppipppppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipipippp?pipppppppipppppppppipipipppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippipppppipp?iippipppippipipppippppppipipppipipipppp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiippiipiipi
```

distances: for each pair, take proportion of differing symbols

```
'Russian'     pipippip
'Polish'      pipipppi
'Czech'       ipiiiipi
```

Aggregating over many examples

```
'Russian'     pipippipipiiiiiiiiipippipiiiiipipiiiipipiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppppippiipipppipiiii
'Czech'       ipiiiipipppipiipippiiipppppppppipipii?iipipppppppppipppppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipipippp?pipppppppipppppppppipipipppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippipppppipp?iippipppippipipppippppppipipppipipipppp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiippiipiipi
```

distances: for each pair, take proportion of differing symbols

```
'Russian'     pipippip
'Polish'      pipipppi
'Czech'       ipiiiipi
```

RU-PL: 2/9 =0.222

Aggregating over many examples

```
'Russian'     pipippipipiiiiiiiiipippipiiiiipipiiiipipiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppppippiipipppipiiii
'Czech'       ipiiiipipppipiipiippiiipppppppppipipii?iipipppppppppipppppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipiippp?pipppppppipppppppppipipipppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippipppppipp?iippipppippipipppippppppipipppipipipppp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiiippiipiipi
```

distances: for each pair, take proportion of differing symbols

```
'Russian'     pipippip
'Polish'      pipipppi
'Czech'       ipiiiipi
```

RU-PL: 2/9 =0.222 PL-CZ: 5/9 =0.555

Aggregating over many examples

```
'Russian'     pipippipipiiiiiiiiipippipiiiiipipiiiipipiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppppippiipipppipiiii
'Czech'       ipiiiipipppipiipippiiipppppppppipipii?iipipppppppppipppppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipipippp?pipppppppipppppppppipipipppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippipppppipp?iippipppippipipppippppppipipppipipipppp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiippiipiipi
```

distances: for each pair, take proportion of differing symbols

```
'Russian'     pipippip
'Polish'      pipipppi
'Czech'       ipiiiipi
```

RU-PL: 2/9 =0.222 PL-CZ: 5/9 =0.555 CZ-RU: 7/9 =0.777

Matrix with binary values

```
'Russian'     pipippipipiiiiiiiiipippipiiiiipipiiiipipiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppppippiipipppipiiii
'Czech'       ipiiiipipppppiipiippppiiipppppppppipipii?iipipppppppppippppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipiippp?pippppppippppppppppipipippppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippippppipp?iippippppippipipppippppppipippppipipippp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiippiipiipi
```

of distances:

```
[1] 'Russian'      0.0 0.36231884 0.75 0.67741936 0.328125 0.45454547 0.1
[2] 'Polish'       0.36231884 0.0 0.61764705 0.5645161 0.359375 0.4090909
[3] 'Czech'        0.75 0.61764705 0.0 0.45901638 0.52380955 0.41538462 0
[4] 'Slovenian'    0.67741936 0.5645161 0.45901638 0.0 0.5254237 0.254237
[5] 'Bulgarian'    0.328125 0.359375 0.52380955 0.5254237 0.0 0.33333334
[6] 'Croatian'     0.45454547 0.4090909 0.41538462 0.2542373 0.33333334 0
[7] 'Belarusian'   0.14925373 0.26865673 0.7121212 0.55737704 0.3125 0.40
```

Operationalization

- Use XSLT scripts on the raw corpus data for classification
- based on word alignment, starting with a primary language
- filter tokens in primary language to modify the question

```
MATRIX
'Bulgarian'   -b---z-aa--aanan----------o-------o-d-----d--z--b--------------b-dbp
'Belarusian'  -z---o-----a-----zz----b-------b----b-bb--zd-----bnn----------------p--
'Czech'       --------yyyyy--n------d-------z---------zz-------bb--r-d------------y-
'Croatian'    -b----ii-rrr---------z----------o----r--------------o---------d--dd-b
'Macedonian'  -------------ynnz-------------------------p--------b----ba-aa-----bdbb
'Polish'      y---nn-------nb----------yobb-------y--z-----y-b-b-----r----------r--
'Russian'     -----o-bbaaa--------b----------------------n----b----b--aaa----y----b
'Slovak'      -o-b----y-aay----------------zz-a-y------z-------b--a---b----a------p-
'Slovenian'   -----o-rrrrr--b----------d---z-r-z--------------b-b-----------------p
'Serbian'     b-----i---rr-------z----z-------zd-------------b-------------i---pd--
'Ukrainian'   --n--z--aaaa---n----z---z-zz---p-------o---------b---nn--------------y
```

Operationalization

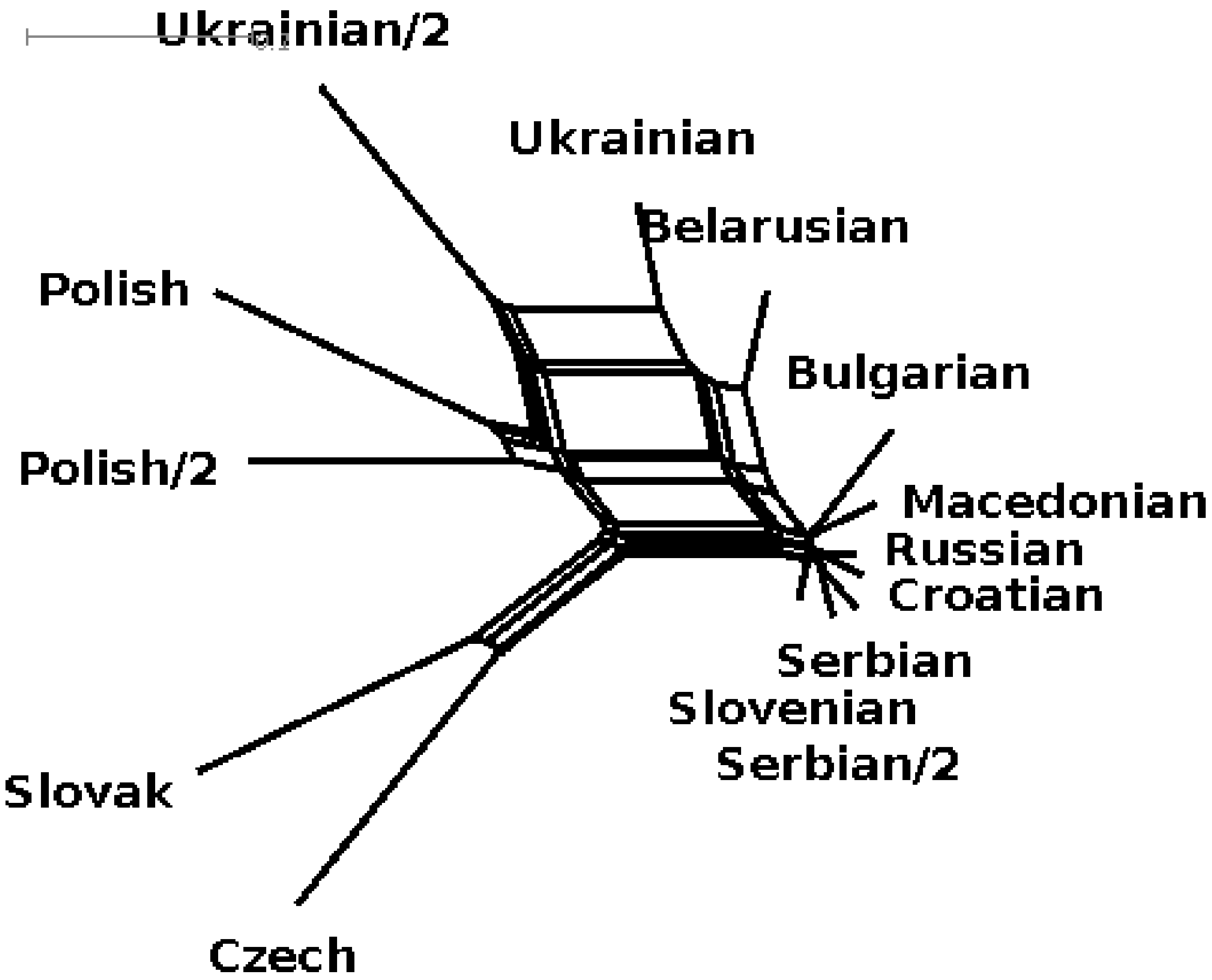
- Use XSLT scripts on the raw corpus data for classification
- based on word alignment, starting with a primary language
- filter tokens in primary language to modify the question

```
MATRIX
'Bulgarian'   -b---z-aa--aanan----------o-------o-d-----d--z--b--------------b-dbp
'Belarusian'  -z---o-----a-----zz----b-------b----b-bb--zd-----bnn----------------p--
'Czech'       --------yyyyy--n------d-------z---------zz-------bb--r-d------------y-
'Croatian'    -b----ii-rrr---------z----------o----r--------------o---------d--dd-b
'Macedonian'  -------------ynnz-------------------------p--------b----ba-aa-----bdbb
'Polish'      y---nn-------nb----------yobb-------y--z-----y-b-b-----r-----------r--
'Russian'     -----o-bbaaa--------b----------------------n----b----b--aaa----y----b
'Slovak'      -o-b----y-aay----------------zz-a-y------z-------b--a---b----a------p-
'Slovenian'   -----o-rrrrr--b----------d---z-r-z--------------b-b-----------------p
'Serbian'     b-----i---rr-------z----z-------zd-------------b-------------i---pd--
'Ukrainian'   --n--z--aaaa---n----z---z-zz---p-------o---------b---nn--------------y
;
```

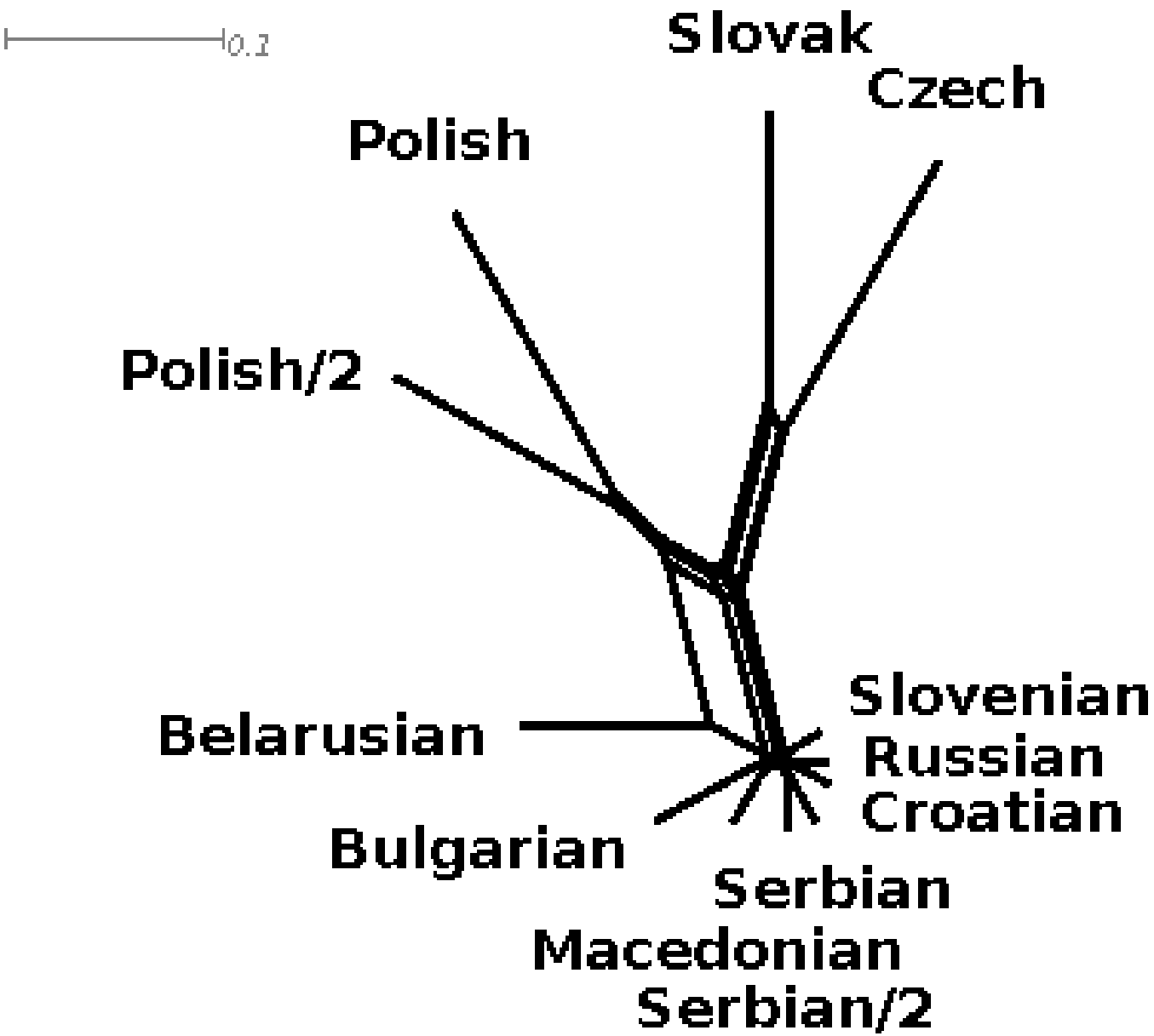
Filter results, calculate Hamming distances, visualize in SplitsTree

Outline

Visualization of single class: DO, Bulgakov

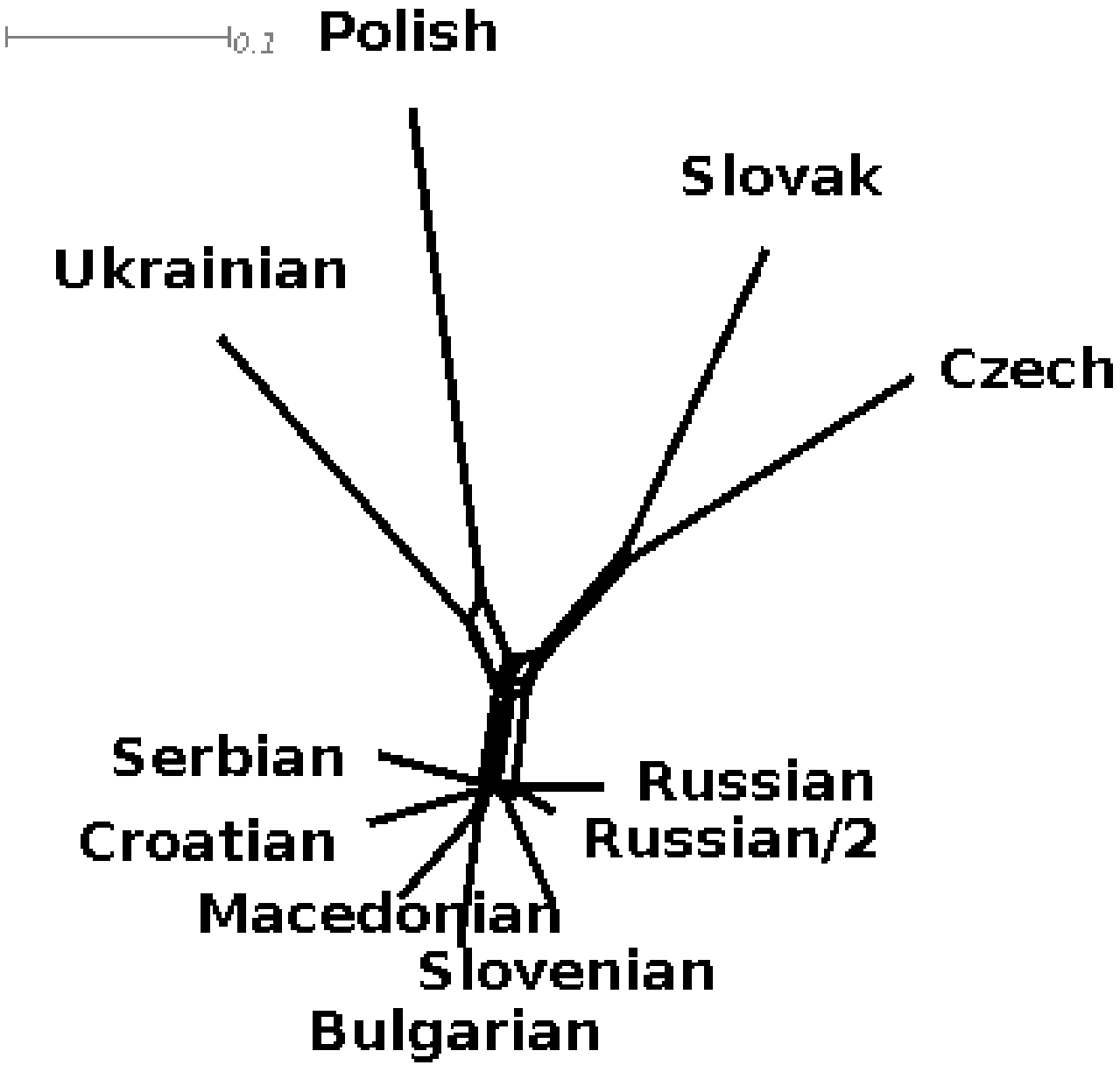


Unmasking: remove UK and UKA from Bulgakov



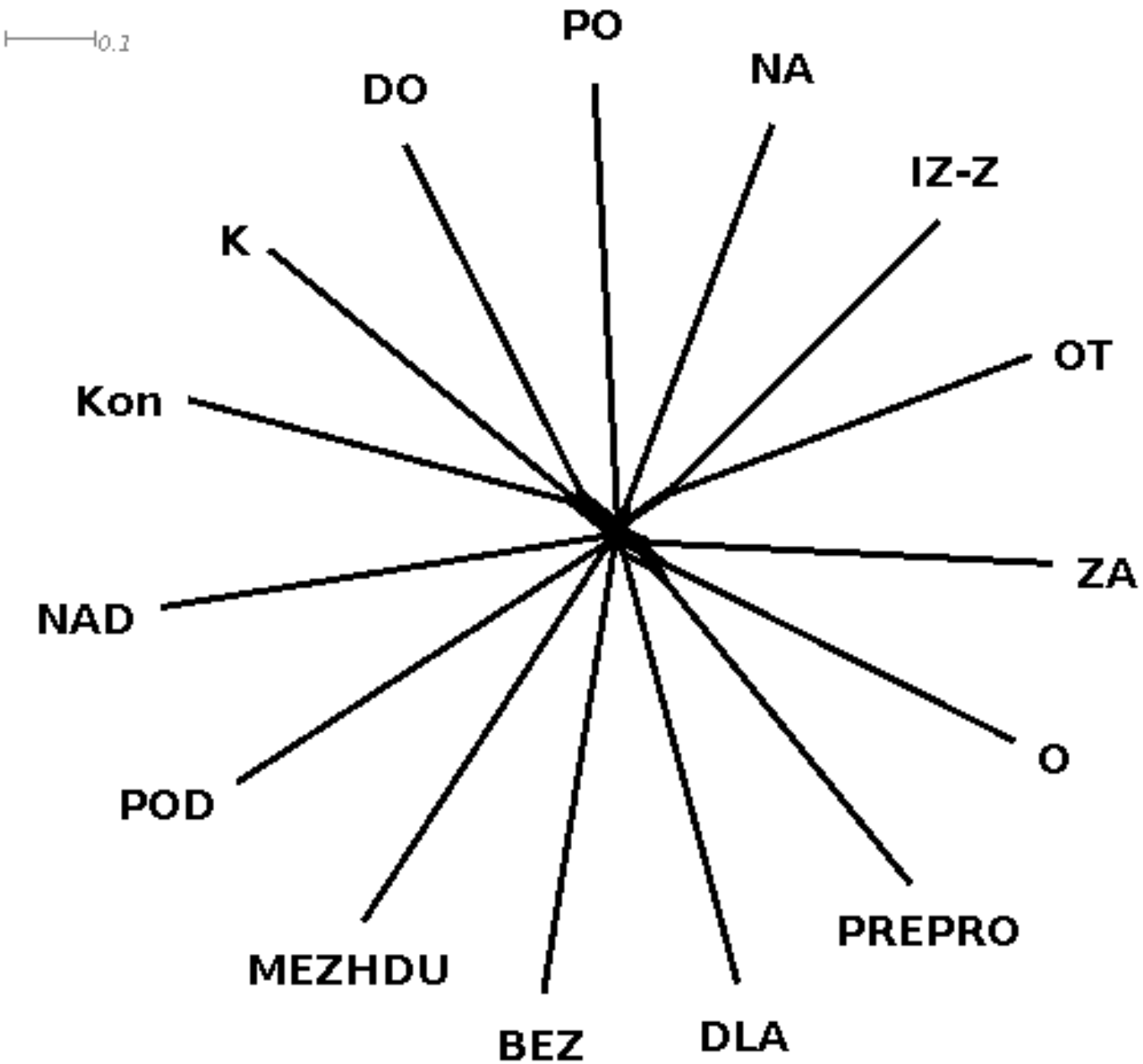
Unmasking: remove BY and UK from Bulgakov

Visualization: DO, Potter

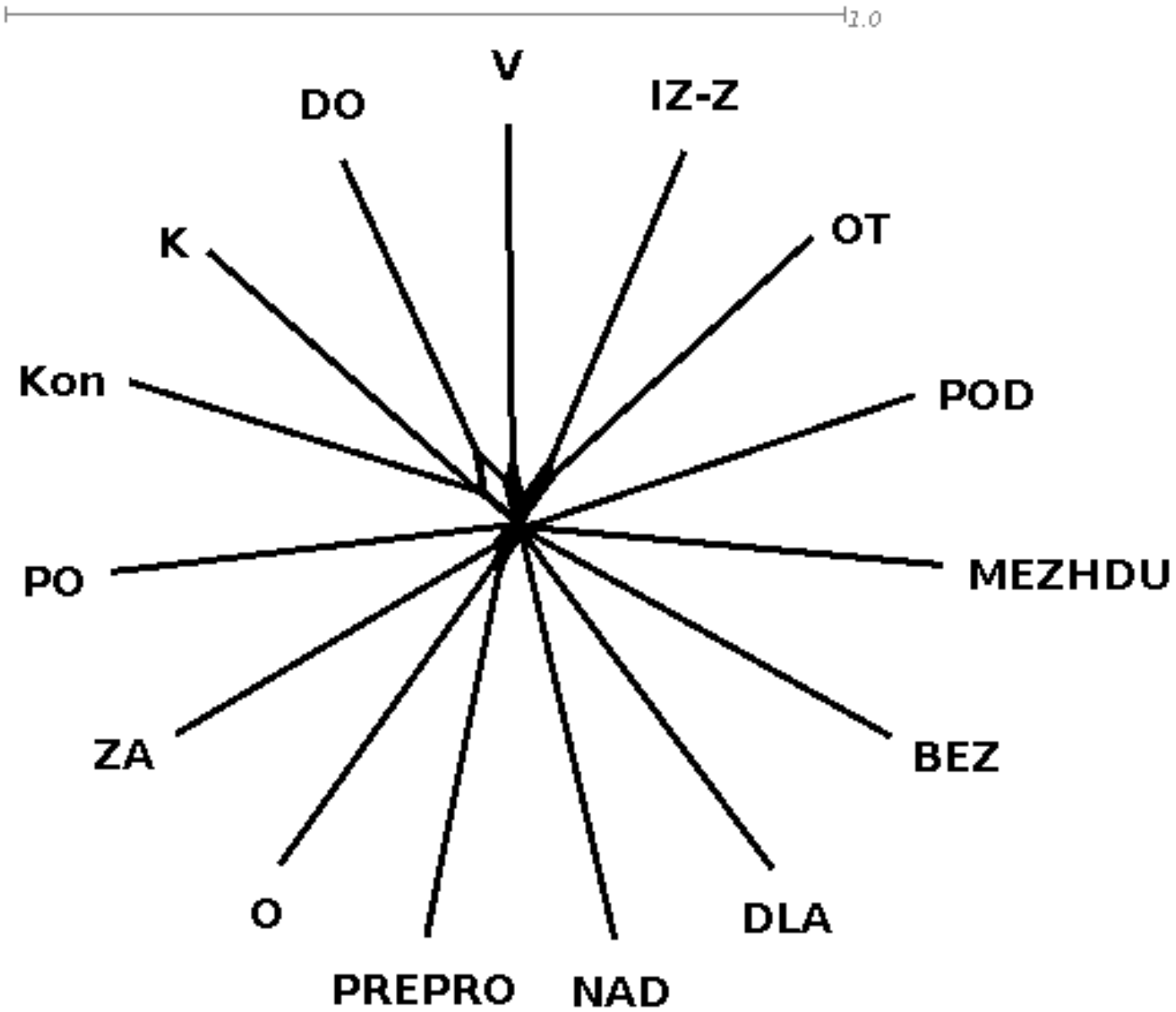


Outline

Preposition types by similarity in context

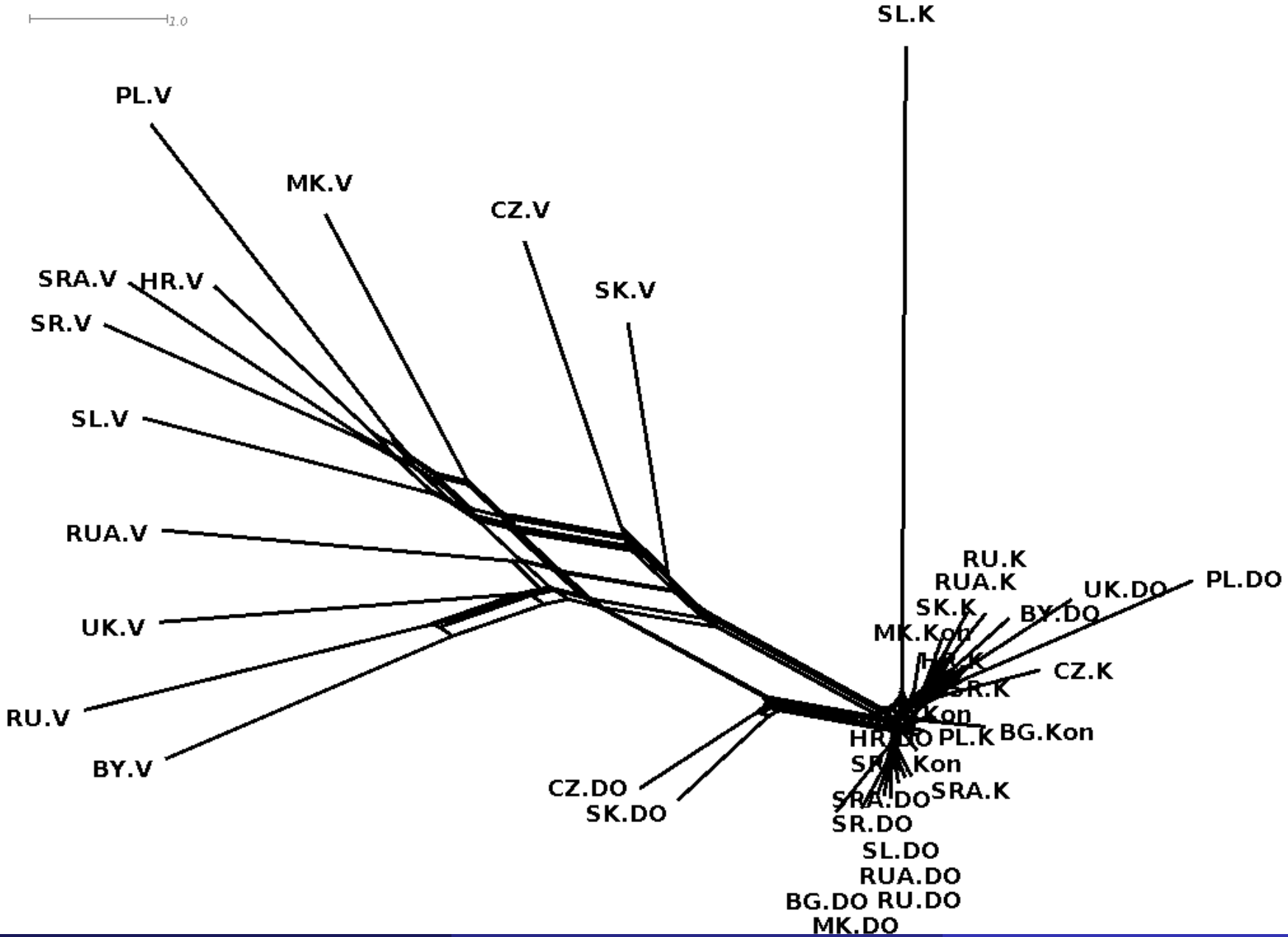


Preposition types by similarity in context

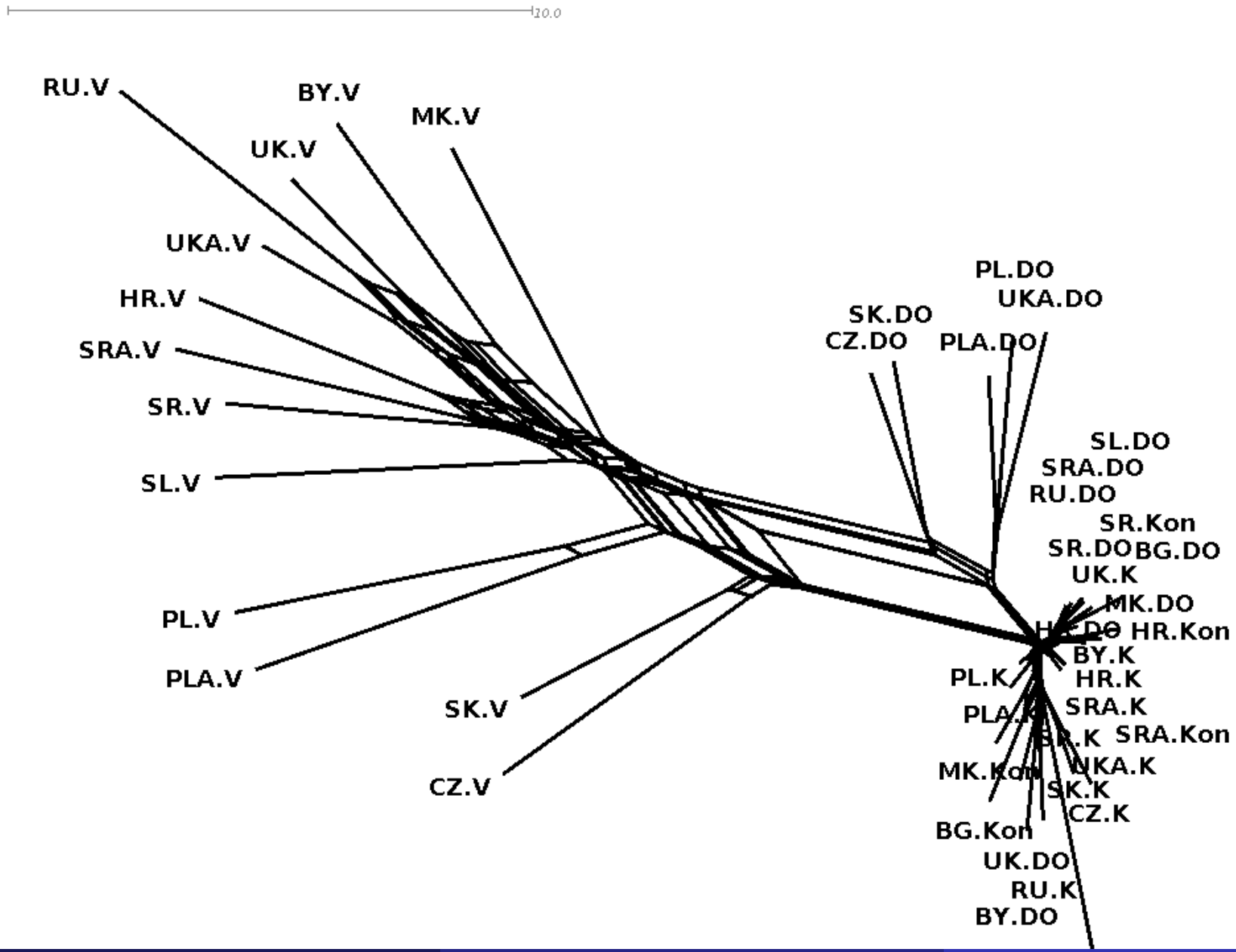


Outline

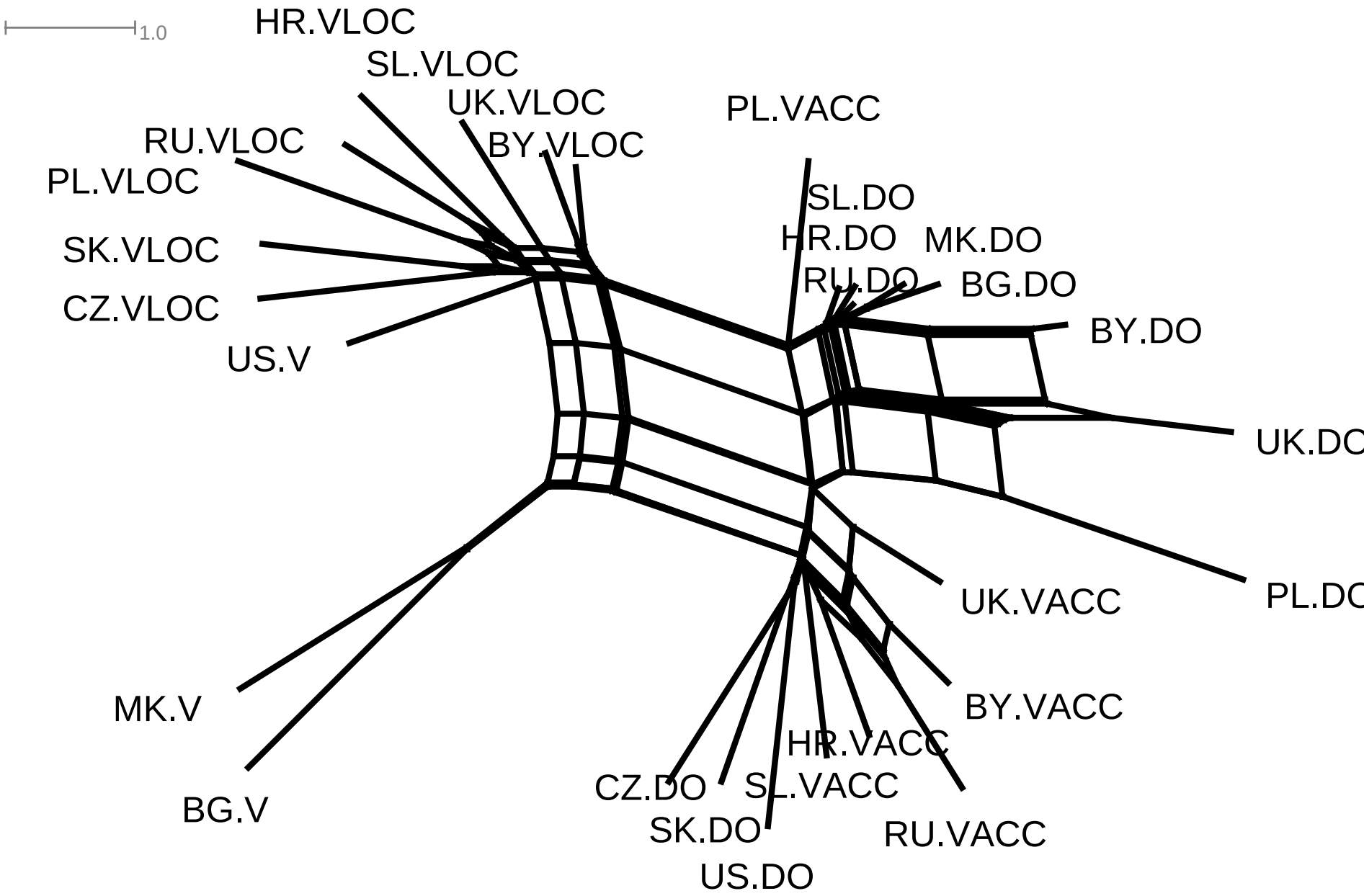
Prepositions by similarity in context: LEM



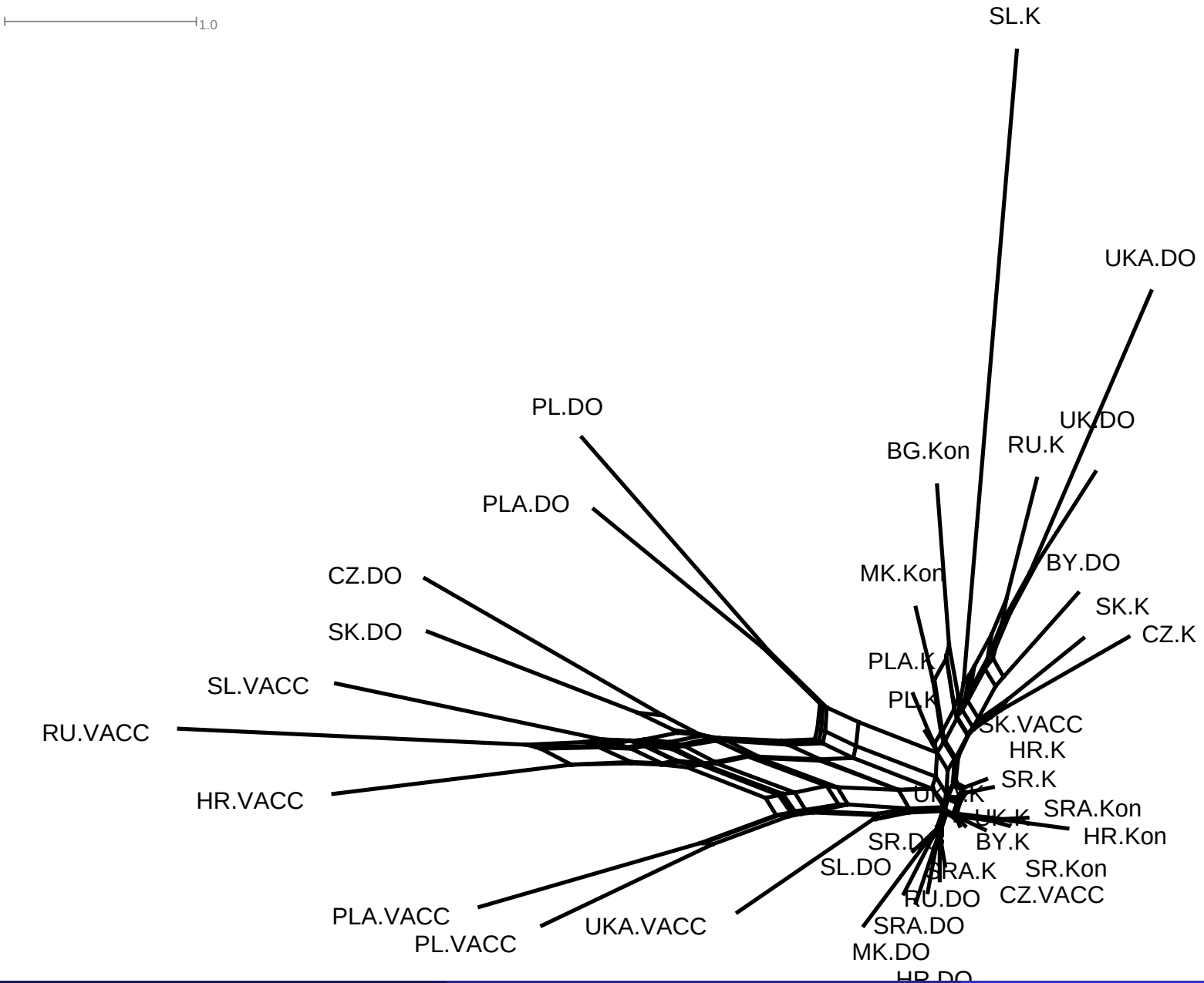
Prepositions by similarity in context: BULGAKOV



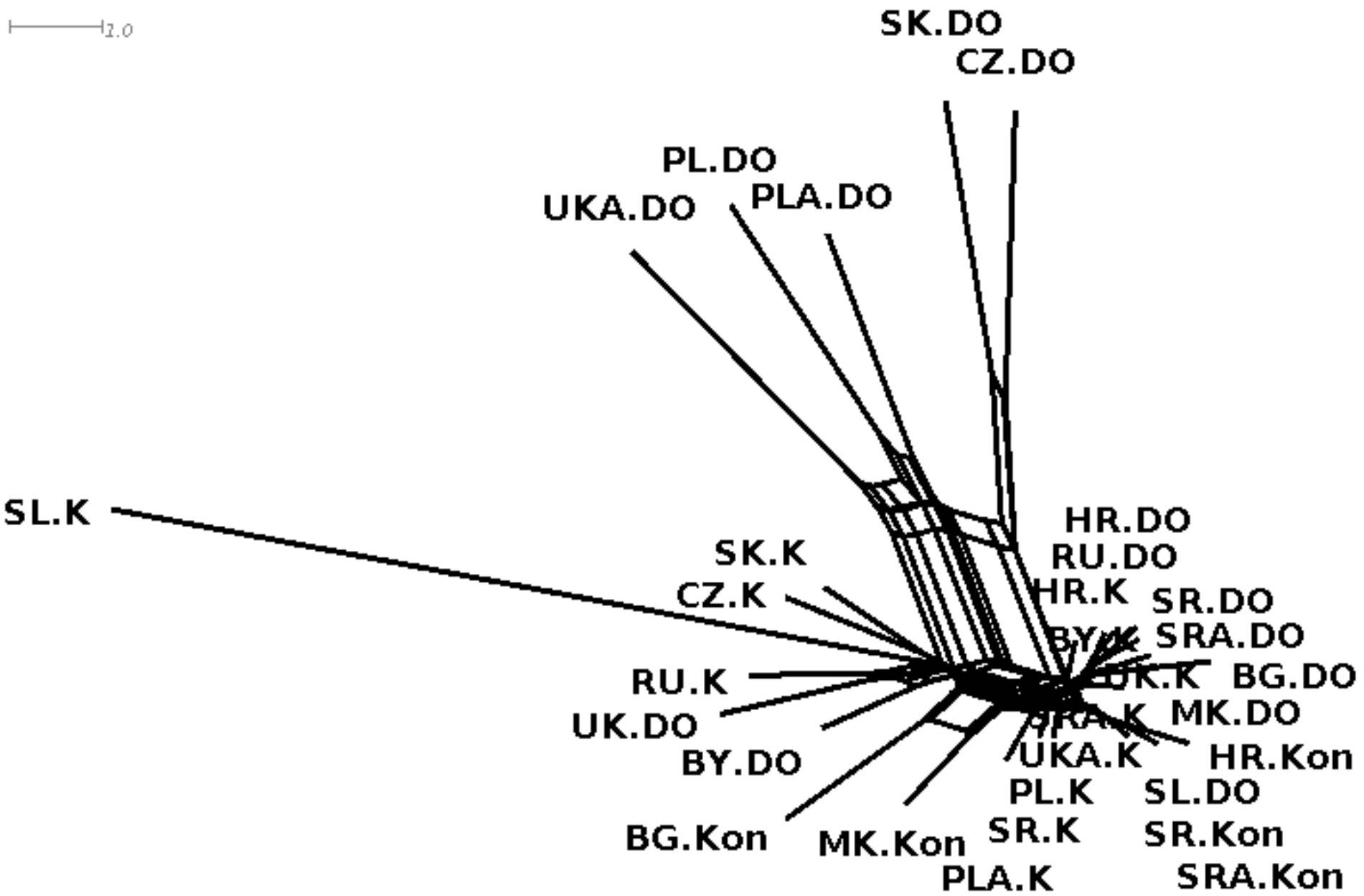
OSTROVSKIJ, with case



Bulgakov, only K, KON, DO and V(acc)



Bulgakov, only K, KON and DO



Outline

verbs equivalent to Russian PRI:

==========PRI:==========

++ru:++ ==>PRI:	++by:++ ==>PRI:	++uk:++ ==>PRI:	++uka:++ ==>PRI:	++pl:++ ==>PRI:	++pla:++ ==>PRI:	++cz:++ ==>PRI:	++sk:++ ==>PRI:	++sl:++ ==>PRI:	++hr:++ ==>PRI:
прийти (66),	прыйсці (53), прыехаць (20),	прийти (43), приїхати	прийти (51), приїхати (16), принести (16),	przyjść (38),	przyjść (43),	přijít (30),	prísť (57),	priti (71),	priznati (1
приходить (52),	прынесці (16), прыходзіць	(17), принести (12),	приходити (11), привести (9), прислати	przyjechać	przybyć (11),	přijet (18),	priznať (11),	priznati (18),	približiti
прийтись (48),	(12), прывезці (9), прызнацца	приходити (9), привезти	(8), привезти (7), прилетіти (7), присісти	(17), przyznać	przynieść (9),	přihnat (7),	prijať (9),	pripeljati	(15),
приказать (29),	(8), прыслухацца (8),	(8), прийти:прийшлий	(6), присідати (6), припасти (6), прийняти	(9), przynieść	przyjąć (9),	přitisknout	pritisnúť (9),	(15),	približavat
принимать (28),	прытуліць (7), прыпасці (7),	(7), привести (7),	(6), притулитися (6), приїжджати (5),	(9), przyjąć (9),	przylgnąć (9),	(7), přiznat	priblížiť (7),	približati	(10), primi
принять (21),	прысесці (6), прытварацца	прислати (7), припасти	прислухатися (5), приснитися (5),	przywieźć (7),	przyznać (5),	(6), přinést	pridať (6),	(15), prinesti	(9), pripad
приходиться (18),	(6), прыязджаць (6),	(6), прикидатися (5),	прилаштувати (5), прикласти (4),	przyśnić (7),	przywieźć (4),	(6), přijmout	priložiť (6),	(14),	(8), prispje
прибавить (18),	прыляцець (6), прывыкнуць	прилетіти (5), прикласти	прикипіти (4), примружитися (3),	przychodzić	przywyknąć	(6), přivézt	prižmúriť (5),	prihajati	(5), privuć
приглашать (16),	(5), прывесці (5), прытуліцца	(5), приїздити (4),	прикидатися (3), приняти:прийняти (3),	(6), przysłać	(4),	(5), přidat	neprísť (4),	(14),	(5),
принести (16),	(5), прыўстаць (4), прыняць	прибрати (4),	прибрати (3), приймати (3), прилягти (3),	(5), przyjeżdżać	przychodzić	(5), připojit	priniesť (4),	pristaviti (7),	pripremiti
привести (13),	(4), прымаць (4), прыслаць	прислухатися (4),	пританцьовувати (3), прикластися (3),	(4),	(4),	(4),	pricestovať	prileteti (7),	(5), priljub
приезжать (12),	(4), прыслухоўвацца (4),	прилучитися (4),	привернути (2), привертати (2),	przygotować	przykładać	přimhouřit	(4), pripojiť	pridružiti	(5),
приехать (12),	прысядаць (4),	присідати (4), прикипіти	причепитися (2), придумати (2),	(4), przypaść	(4), przyłączyć	(3), přivést	(4), priviesť	(7),	prisluškiva
пригласить (11),	прытанцоўваць (4), прызнаць	(4), пританцьовувати (4),	пригадати (2), припадати (2), притискати	(4), przyłożyć	(4),	(3), přiblížit	(3), prihnať	privzdigniti	(4),
признаваться (9),	(3), прымешвацца (3),	причепитися (3),	(2), приєднатися (2), прислухаючисити	(4), przytulić	przyjeżdżać	(3),	(3), nepriniesť	(6),	prisustvov
приближать (9),	прыйсці:прыйсце (3),	приготувати (3),	(2), притихнути (2), приставати (1),	(4),	(3), przyjechać	přitáhnout	(3), prisniť	pripotovati	(4), pričinj
принадлежать (9),	прысутнічаць (3), прылегчы	приєднатися (3),	прищулитися (1), приглушити (1),	przyzwyczaić	(3),	(2),	(3),	(5), pritisniti	(4), prihva
признавать (8),	(3), прыбегчы (3),	прийняти (3),	призвести (1), прищулити (1),	(3), przybyć	przygotować	přiznávat	prit'ahovať	(5),	(4),
приготавливать	прыцягнуць (2), прыцягваць	приснитися (3),	приготувати (1), прикривати (1),	(3), przybrać	(3), przyłożyć	(2),	(2),	prisluškovati	priključiti
(8), приближаться	(2), прыдумаць (2),	припасти:припалий (3),	прикрити (1), пригостити (1),	(3), przyłączyć	(3), przyśnić	přiběhnout	neprichádzať	(4), pripreti	(4),
(8),	прызнацца:прызнавацца (2),	притиснутися (3),	примружувати (1), притримати (1),	(3), przydarzyć	(3),	(2),	(2), pripraviť	(3),	pridružiti
прищуриваться	прыйшоўшы (2), прыкрываць	привертати (2),	пристати (1), прилучитися (1),	(2), przykuć	przypomnieć	přicházet	(2), pristaviť	pripraviti	(4), pripiti
(8),	(2), прыкрыць (2),	примружитися (2),	привласнити (1), пританцювати (1),	(2), przykucnąć	(3), przysłać	(2),	(2), pritískať	(3),	(4), privlač
прислушиваться	прытварыцца (2), прыліць	прискалити (2), присісти	пришити (1), притягти (1), приволочити	(2),	(3), przyciągać	přistoupit	(2), priletieť	prisluhniti	(3),
(8), привезти (7),	(2), прыдзірацца (2),	(2), придумати (2),	(1), пришпилити (1),	przyśpieszyć	(2),	(2), přiletět	(2), privaliť	(3), pripetiti	priznavati
присаживаться	прывалачы (2), прывітаць (2),	прибути:прибулий (2),	пристанути:пристати:пристав:приставити	(2),	przykucnąć	(2), přiložit	(1), priviezť	(2), prijeti	(3),
(7), приводить (7),	прыцэліцца (2), прыкіпець	прикривати (2),	(1), пригадувати (1), пригадатися (1),	przyprowadzić	(2),	(2),	(1), prisadnúť	(2),	prikrivati
притворяться (7),	(2), прыціскацца (2),	приносити (2),	привітати (1), привітатися (1), прибігти	(2), przyznawać	przysiadać	přispěchat	(1), pripliesť	pripravljati	(3), pričini
присниться (7),	прыцягнуты (1), прызнаваць	приняти:прийняти (2),	(1), притуляти (1), прикотити (1),	(2), przymrużać	(2),	(1), přispět	(1), privádzať	(2),	(3), primat
приблизиться (7),	(1), прызнаць:прызнаваць (1),	приймати (2), пригадати	примусити (1), приміряти (1), принадити	(2), przywiązać	przysłaniać	(1), přihodit	(1),	prilegnuti	(3),
присоединиться	прычапіцца (1),	(2), припадати (2),	(1), привабити (1), притупитися (1),	(2), przymrużyć	(2),	(1), přilepit	prichádzať	(2), pripadati	prilegnuti
(7), привлекать	прывяць:прывесці:прывянуць	прибігти (2), притискати	прицмокнути (1), прицмокувати (1),	(2),	przyśpieszać	(1),	(1), pridávať	(2), privleči	(3), prisjet
(6), прикрывать	(1), прызнавацца (1),	(2), привабити (2),	присисати (1), привалитися (1),		(2), przylecieć				

verbs equivalent to Russian PRI: zoom in

++ru:++ ==>PRI:	++cz:++ ==>PRI:	++sk:++ ==>PRI:	++sl:++ ==>PRI:	++hr:++ ==>PRI:
прийти (66),	přijít (30),	prísť (57),	priti (71),	priznati (16),
приходить (52),	přijet (18),	priznať (11),	priznati (18),	približiti
прийтись (48),	přihnat (7),	prijať (9),	pripeljati	(15),
приказать (29),	přitisknout	pritisnúť (9),	(15),	približavati
принимать (28),	(7), přiznat	priblížiť (7),	približati	(10), primiti
принять (21),	(6), přinést	pridať (6),	(15), prinesti	(9), pripadati
приходиться (18),	(6), přijmout	priložiť (6),	(14),	(8), prispjeti
прибавить (18),	(6), přivézt	prižmúriť (5),	prihajati	(5), privući
приглашать (16),	(5), přidat	neprísť (4),	(14),	(5),
принести (16),	(5), připojit	priniesť (4),	pristaviti (7),	pripremiti
привести (13),	(4),	pricestovať	prileteti (7),	(5), priljubiti
приезжать (12),				

Outline

A web based application

Computations (so far) done using R, Perl-scripts, XSLT, OpenCWB, SplitsTree

Projected web interface:

- Using AngularJ, Django, Python, MySQL
- An asynchronous interface: users log on and manage their data and evaluations
- Two types of user data:
  - environment data: parameter files, default options
  - evaluations: individual runs of parameter files with different options and their results
- Results: Colour-coded corpus samples; zip-files with network and distance data, ready visualizations and word lists

A web based application

Computations (so far) done using R, Perl-scripts, XSLT, OpenCWB, SplitsTree

Projected web interface:

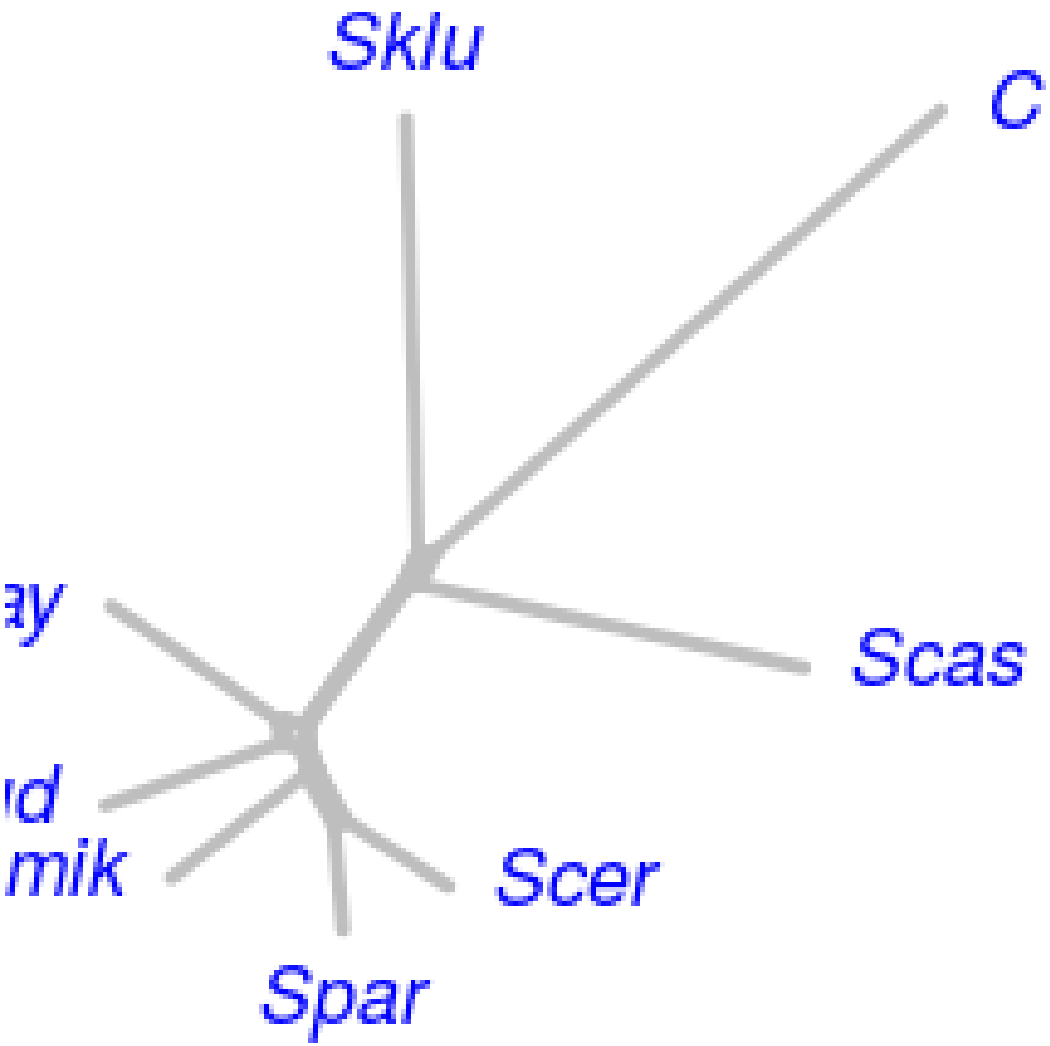
- Using AngularJ, Django, Python, MySQL
- An asynchronous interface: users log on and manage their data and evaluations
- Two types of user data:
  - environment data: parameter files, default options
  - evaluations: individual runs of parameter files with different options and their results
- Results: Colour-coded corpus samples; zip-files with network and distance data, ready visualizations and word lists
- ...comments very much appreciated!

Possible and actual implementation in R

Visualization

- Step one: categorization of the corpus data
  in XSLT so far -> can be replaced by R
  - access corpus data with rcqp (interface to CorpusWorkBench)
  - query primary language to get corpus positions; extract relevant word forms from word alignment/sentence alignment into a table object
  - apply classification to each cell to derive a table of classifications
- Step two: filter the table to restrict it to certain features
  so far done in Perl, elementary in R
- Step three: derive various distance matrices
  built-in in SplitsTree -> basically already implemented in R (see separate slide)
- Step four: visualize as NeighborNet
  -> implemented in R in phangorn package, but not very nice graphics

Phangorn Neighbornet:



Possible and actual implementation in R

Corpus examples

- Step one: query corpus data with rcqp
- Step two: apply xslt stylesheet for color coding

... not quick and not elegant

Summary

- A query system to query a multilingual parallel corpus
- Focus on analyzing functional domains reflected in distributions
- Individual data points and aggregations
- Projected: web-based system for registered user

Thank you

Dziękuję!

www.parasolcorpus.org
ruprecht.waldenfels@gmail.com
This research is supported by the Swiss National Science Foundation