The Regensburg Parallel Corpus of Slavonic

Ruprecht von Waldenfels

Institut für Slavistik
Universität Regensburg

IPI PAN Warsaw, 19.4.06

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

Characteristics of the proposed corpus

- Purpose: Contrastive linguistic studies

Characteristics of the proposed corpus

- Purpose: Contrastive linguistic studies

Design objectives

- *Easy* to use and augment, flexible to demands of researchers (lack of human resources)

Characteristics of the proposed corpus

- Purpose: Contrastive linguistic studies

Design objectives

- *Easy* to use and augment, flexible to demands of researchers (lack of human resources)
- Minimize manual preprocessing

Characteristics of the proposed corpus

- Purpose: Contrastive linguistic studies

Design objectives

- *Easy* to use and augment, flexible to demands of researchers (lack of human resources)
- Minimize manual preprocessing
- Depend as little as possible on language specific resources: for many Slavonic languages, they are not easy to come by

Main strategies

- Concentrate on 20th century prose

Main strategies

- Concentrate on 20th century prose
- Slavonic languages, as well as German and whatever is in need...
- Use what's easily available

Main strategies

- Concentrate on 20th century prose
- Slavonic languages, as well as German and whatever is in need...
- Use what's easily available
- Use what's translated into *many* languages in order to take advantage of distribution effects

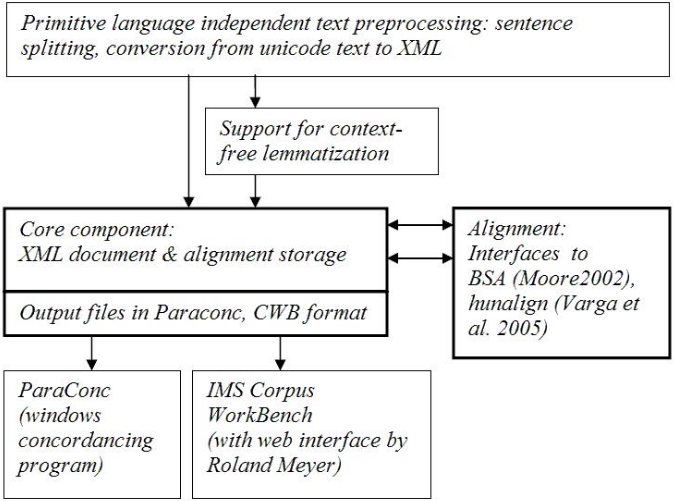
Main strategies

- Concentrate on 20th century prose
- Slavonic languages, as well as German and whatever is in need...
- Use what's easily available
- Use what's translated into *many* languages in order to take advantage of distribution effects
- Try to get a *balanced* corpus in the sense that we have similar distribution of original texts /translations

Main strategies

- Concentrate on 20th century prose
- Slavonic languages, as well as German and whatever is in need...
- Use what's easily available
- Use what's translated into *many* languages in order to take advantage of distribution effects
- Try to get a *balanced* corpus in the sense that we have similar distribution of original texts /translations
  (NOT only from English)

Overview – architecture



Primitive language independent text preprocessing: sentence splitting, conversion from unicode text to XML

Support for context-free lemmatization

Core component: XML document & alignment storage

Output files in Paraconc, CWB format

Alignment: Interfaces to BSA (Moore2002), hunalign (Varga et al. 2005)

ParaConc (windows concordancing program)

IMS Corpus WorkBench (with web interface by Roland Meyer)

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

Input module

Input format: Plain UTF-8 text files, annotated with information about the document and chapter divisions

Input module

Input format: Plain UTF-8 text files, annotated with information about the document and chapter divisions, for example BulgakovMaster_PL.txt

```
author=Michaił Bułhakow
origauthor=Михаил Булгаков
title=Mistrz i Małgorzata
origtitle=Мастер и Маргарита
translator=Irena Lewandowska i Witold Dąbrowski
...
endheader

I

<BLOCK> 1. Nigdy nie rozmawiaj z nieznajomymi

Kiedy zachodziło właśnie gorące wiosenne słońce, na
Patriarszych Prudach zjawiło się dwu obywateli....
```

Input module

- The system
  - performs tokenization and sentence splitting

Input module

- The system
  - performs tokenization and sentence splitting
  - constructs word lists as possible input to a lemmatizer

Input module

- The system
  - performs tokenization and sentence splitting
  - constructs word lists as possible input to a lemmatizer (this is simpler than integration of various lemmatizers)

Input module

- The system
  - performs tokenization and sentence splitting
  - constructs word lists as possible input to a lemmatizer (this is simpler than integration of various lemmatizers)
  - converts the text file and possible word-lemma lists to XML

Input module

- The system
  - performs tokenization and sentence splitting
  - constructs word lists as possible input to a lemmatizer (this is simpler than integration of various lemmatizers)
  - converts the text file and possible word-lemma lists to XML
- The component ensures quick and easy augmentation of the corpus, with the possibility to include primitive lemmatization. Richer annotation can be done done directly on the XML document files. Multimodal information, translation comments and the like are not provided for.

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

Core component: Document and alignment repository

XML Formats

- XML documents encoding: header (in extended TEI),(chapter) divisions,

Core component: Document and alignment repository

XML Formats

- XML documents encoding: header (in extended TEI),(chapter) divisions,sentence segments,

Core component: Document and alignment repository

XML Formats

- XML documents encoding: header (in extended TEI),(chapter) divisions,sentence segments,tokens, <lemma>, <tag1>..<tag5>

Core component: Document and alignment repository

XML Formats

- XML documents encoding: header (in extended TEI),(chapter) divisions,sentence segments,tokens, <lemma>, <tag1>..<tag5>
- Stand-off alignment files encode begin and end of corresponding segments

Core component: Document and alignment repository

XML Formats

- XML documents encoding: header (in extended TEI),(chapter) divisions,sentence segments,tokens, <lemma>, <tag1>..<tag5>
- Stand-off alignment files encode begin and end of corresponding segments e.g.
  <alig Ln1Strt="0" Ln2Strt="0" Ln1End="1" Ln2End="1"/>

Core component: Document and alignment repository

XML Formats

- XML documents encoding: header (in extended TEI),(chapter) divisions,sentence segments,tokens, <lemma>, <tag1>..<tag5>
- Stand-off alignment files encode begin and end of corresponding segments e.g. <alig Ln1Strt="0" Ln2Strt="0" Ln1End="1" Ln2End="1"/>
- Linking by filename and language shorts: LemKongres_DE.rpc, LemKongres_RU.rpc, LemKongres_DE-RU.alg

Core component

Annotation

Generic content:
According to the resources available, files in different languages will contain data of different content and quality

Core component

Annotation

Generic content:
According to the resources available, files in different languages will contain data of different content and quality, e.g.:

- Ukrainian: No annotation.
- Russian: Text and lemmas (context-free lemmatizer RMORPH by Grigori Sidorov)
- Slovak: Tagged text (Garabik 2005)

Core component

Annotation

Generic content:
According to the resources available, files in different languages will contain data of different content and quality, e.g.:

- Ukrainian: No annotation.
- Russian: Text and lemmas (context-free lemmatizer RMORPH by Grigori Sidorov)
- Slovak: Tagged text (Garabik 2005)

Tag sets are not uniform across languages, that is, <tag1> is a cover term for whatever information has been annotated for a given language

- Russian: Text and lemmas (context-free lemmatizer RMORPH by Grigori Sidorov)
  <s id="0"><tok>Ты<lemma>ты</lemma></tok>
  <tok>должна<lemma>должный</lemma></tok>
  <tok>сделать<lemma>сделать</lemma></tok>
- Slovak: Tagged text (tagging thanks to Gabarek, Bratislava)
  <s id="0"><tok>Musíš<lemma>musieť</lemma>
  <tag1>VB-S—2P-AA—</tag1>
  <tag2>VKesb+</tag2></tok>
  <tok>robiť<lemma>robiť</lemma>
  <tag1>Vf———A—-</tag1> <tag2>VIe+</tag2></tok>

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
    - Input module
    - Core component: Document and alignment repository
    - Alignment
    - Output module
    - State of the corpus
- Does lemmatization help alignment?
    - The alignment tools used in RPC
    - Questions
    - Experiment setup
    - Experiment: Polish-Russian and German-Russian
- Conclusions

Alignment

- Each language version is pairwise aligned to every other

LemSolaris_PL
LemSolaris_DE
LemSolaris_RU
LemSolaris_SB

→

LemSolaris_DE-PL
LemSolaris_DE-RU
LemSolaris_DE-SB
LemSolaris_PL-RU
LemSolaris_PL-SB
LemSolaris_RU-SB

- Every n-th language adds n-1 alignments

Alignment

- Each language version is pairwise aligned to every other

LemSolaris_PL		LemSolaris_DE-PL
LemSolaris_DE		LemSolaris_DE-RU
LemSolaris_RU	→	LemSolaris_DE-SB
LemSolaris_SB		LemSolaris_PL-RU
		LemSolaris_PL-SB
		LemSolaris_RU-SB

- Every n-th language adds n-1 alignments
- Any query on more than two languages will rely on these pairwise alignations

Alignment

- Each language version is pairwise aligned to every other

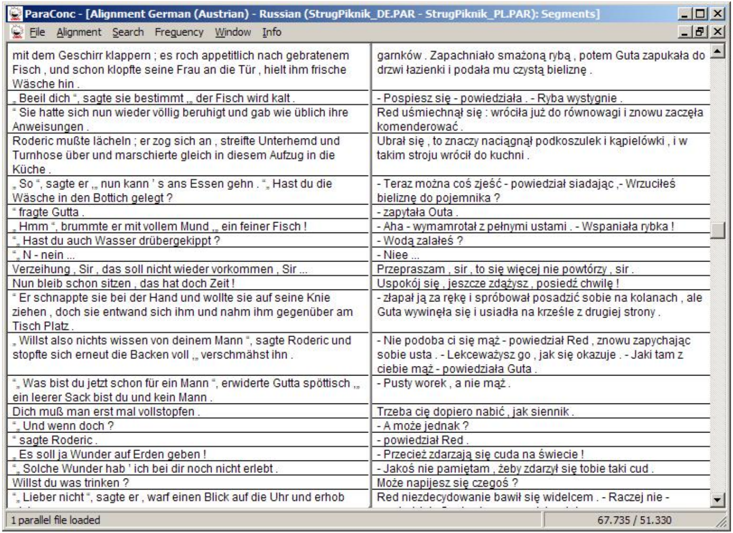
LemSolaris_PL		LemSolaris_DE-PL
LemSolaris_DE		LemSolaris_DE-RU
LemSolaris_RU	→	LemSolaris_DE-SB
LemSolaris_SB		LemSolaris_PL-RU
		LemSolaris_PL-SB
		LemSolaris_RU-SB

- Every n-th language adds n-1 alignments
- Any query on more than two languages will rely on these pairwise alignations
- Automatic construction of alignments via scripts; two aligners supported:
  - hunalign
  - BSA

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

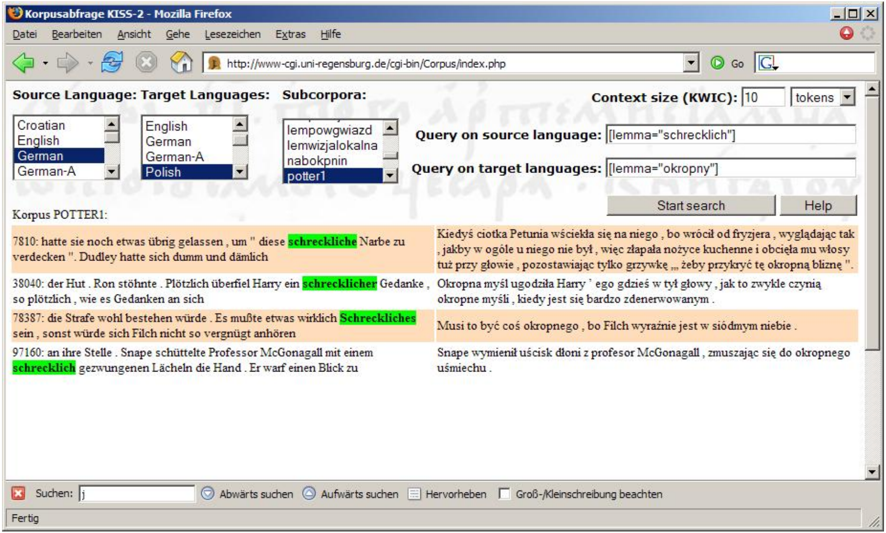
Output in ParaConc format


ParaConc - [Alignment German (Austrian) - Russian (StrugPiknik_DE.PAR - StrugPiknik_PL.PAR): Segments]
File
Alignment
Search
Frequency
Window
Info
1 parallel file loaded
67.735 / 51.330

mit dem Geschirr klappern ; es roch appetitlich nach gebratenem Fisch , und schon klopfte seine Frau an die Tür , hielt ihm frische Wäsche hin .	garnków . Zapachniało smażoną rybą , potem Guta zapukała do drzwi łazienki i podała mu czystą bieliznę .
„ Beeil dich “, sagte sie bestimmt ,„ der Fisch wird kalt .	- Pospiesz się - powiedziała . - Ryba wystygnie .
“ Sie hatte sich nun wieder völlig beruhigt und gab wie üblich ihre Anweisungen .	Red uśmiechnął się : wróciła już do równowagi i znowu zaczęła komenderować .
Roderic mußte lächeln ; er zog sich an , streifte Unterhemd und Turnhose über und marschierte gleich in diesem Aufzug in die Küche .	Ubrał się , to znaczy naciągnął podkoszulek i kąpielówki , i w takim stroju wrócił do kuchni .
„ So “, sagte er ,„ nun kann ' s ans Essen gehn . “„ Hast du die Wäsche in den Bottich gelegt ?	- Teraz można coś zjeść - powiedział siadając ,- Wrzuciłeś bieliznę do pojemnika ?
“ fragte Gutta .	- zapytała Outa .
„ Hmm “, brummte er mit vollem Mund ,„ ein feiner Fisch !	- Aha - wymamrotał z pełnymi ustami . - Wspaniała rybka !
“„ Hast du auch Wasser drübergekippt ?	- Wodą zalałeś ?
“„ N - nein ...	- Niee ...
Verzeihung , Sir , das soll nicht wieder vorkommen , Sir ...	Przepraszam , sir , to się więcej nie powtórzy , sir .
Nun bleib schon sitzen , das hat doch Zeit !	Uspokój się , jeszcze zdążysz , posiedź chwilę !
“ Er schnappte sie bei der Hand und wollte sie auf seine Knie ziehen , doch sie entwand sich ihm und nahm ihm gegenüber am Tisch Platz .	- złapał ją za rękę i spróbował posadzić sobie na kolanach , ale Guta wywinęła się i usiadła na krześle z drugiej strony .
„ Willst also nichts wissen von deinem Mann “, sagte Roderic und stopfte sich erneut die Backen voll ,„ verschmähst ihn .	- Nie podoba ci się mąż - powiedział Red , znowu zapychając sobie usta . - Lekceważysz go , jak się okazuje . - Jaki tam z ciebie mąż - powiedziała Guta .
“„ Was bist du jetzt schon für ein Mann “, erwiderte Gutta spöttisch ,„ ein leerer Sack bist du und kein Mann .	- Pusty worek , a nie mąż .
Dich muß man erst mal vollstopfen .	Trzeba cię dopiero nabić , jak siennik .
“„ Und wenn doch ?	- A może jednak ?
“ sagte Roderic .	- powiedział Red .
„ Es soll ja Wunder auf Erden geben !	- Przecież zdarzają się cuda na świecie !
“„ Solche Wunder hab ' ich bei dir noch nicht erlebt .	- Jakoś nie pamiętam , żeby zdarzył się tobie taki cud .
Willst du was trinken ?	Może napijesz się czegoś ?
“„ Lieber nicht “, sagte er , warf einen Blick auf die Uhr und erhob	Red niezdecydowanie bawił się widelcem . - Raczej nie -

(http://www.athel.com/para.html)

Output to Corpus WorkBench



http://www-cgi.uni-regensburg.de/cgi-bin/Corpus/index.php

Source Language: Croatian, English, German, German-A (selected: German)

Target Languages: English, German, German-A, Polish (selected: Polish)

Subcorpora: lempowgwiazd, lemwizjalokalna, nabokpnin, potter1 (selected: potter1)

Context size (KWIC): 10 tokens

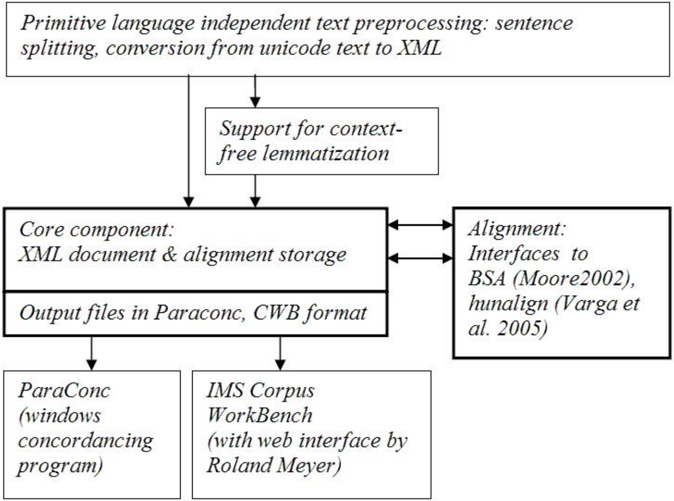
Query on source language: [lemma="schrecklich"]

Query on target languages: [lemma="okropny"]

Korpus POTTER1:

7810: hatte sie noch etwas übrig gelassen , um " diese **schreckliche** Narbe zu verdecken ". Dudley hatte sich dumm und dämlich	Kiedyś ciotka Petunia wściekła się na niego , bo wrócił od fryzjera , wyglądając tak , jakby w ogóle u niego nie był , więc złapała nożyce kuchenne i obcięła mu włosy tuż przy głowie , pozostawiając tylko grzywkę ,,, żeby przykryć tę okropną bliznę ".
38040: der Hut . Ron stöhnte . Plötzlich überfiel Harry ein **schrecklicher** Gedanke , so plötzlich , wie es Gedanken an sich	Okropna myśl ugodziła Harry ' ego gdzieś w tył głowy , jak to zwykle czynią okropne myśli , kiedy jest się bardzo zdenerwowanym .
78387: die Strafe wohl bestehen würde . Es mußte etwas wirklich **Schreckliches** sein , sonst würde sich Filch nicht so vergnügt anhören	Musi to być coś okropnego , bo Filch wyraźnie jest w siódmym niebie .
97160: an ihre Stelle . Snape schüttelte Professor McGonagall mit einem **schrecklich** gezwungenen Lächeln die Hand . Er warf einen Blick zu	Snape wymienił uścisk dłoni z profesor McGonagall , zmuszając się do okropnego uśmiechu .

Overview- architecture



Primitive language independent text preprocessing: sentence splitting, conversion from unicode text to XML

Support for context-free lemmatization

Core component:
XML document & alignment storage

Output files in Paraconc, CWB format

Alignment:
Interfaces to
BSA (Moore2002),
hunalign (Varga et al. 2005)

ParaConc (windows concordancing program)

IMS Corpus WorkBench (with web interface by Roland Meyer)

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

Languages

LNG	in full	tokens	lemmas	tags1	tags2
DE	German	1 154 356	46 971	no	no
EN	English	208 986	0	no	no
HR	Croatian	90 581	0	no	no
PL	Polish	1 861 303	46 132	no	no
RU	Russian	2 352 599	50 126	no	no
SB	Serbian (cyrillic script)	244 277	11 920	no	no
SK	Slovak	620 370	28 794	yes	yes
SX	Serbian (latin script)	74 199	7 801	no	no
UK	Ukrainian	179 630	21 291	no	no

Texts

texts	DE	DEa	EN	HR	PL	RU	RUa	SB	SK	SX	UK
BoellClown	DE					RU			SK		
BoellFrau	DE								SK		
BulgakovMaster					PL	RU		SB			
EUVerf	DE								SK		
EndeMomo	DE					RU					
GralsWelt	DE								SK		
KafkaErz	DE								SK		
LemAstronauci					PL	RU					
LemFiasko					PL	RU					
LemGlosPana					PL	RU					
LemKatar					PL	RU					
LemKongres	DE				PL	RU					
LemPamWannie					PL	RU					
LemPokoj					PL	RU					
LemPowGwiazd					PL	RU					
LemSolaris	DE				PL	RU				SX	
LemWizjaLokalna					PL	RU					
NabokPnin	DE	DEa							SK		
Potter1	DE		EN	HR	PL	RU	RUa	SB	SK		UK
Potter2	DE		EN		PL	RU	RUa				UK
Potter3					PL	RU	RUa				
Potter4					PL	RU					
Potter5					PL	RU					
SloOestHK	DE								SK		
StrugLebedi	DE				PL	RU					
StrugPiknik	DE				PL	RU			SK		

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

Alignment tools

Requirements:

- NO language specific NLP resources such as seed lexica, stop word lists, training sets, etc.

Alignment tools

Requirements:

- NO language specific NLP resources such as seed lexica, stop word lists, training sets, etc.
- As little manual preprocessing as possible (e.g., no paragraph segmentation).

Alignment tools

Requirements:

- NO language specific NLP resources such as seed lexica, stop word lists, training sets, etc.
- As little manual preprocessing as possible (e.g., no paragraph segmentation).

Two choices (to my knowledge):

- BSA: Bilingual-sentence-aligner (Moore 2002)
- hunalign (Varga et al. 2005)

Algorithms

Three stages:

1. sentence-length based algorithm
2. best alignments are used to build a translation model (bsa) / dictionary (hunalign)
3. both used these for final alignment

Algorithms

BSA:

- utilizes all files to be aligned to build the translation model
- outputs only 1-1 beads

Algorithms

BSA:

- utilizes all files to be aligned to build the translation model
- outputs only 1-1 beads
  Use intermediate results to extract 1-1, 1-2, 2-1 beads, discard 0-1, 1-0 beads. Simple heuristic to align the rest.

Algorithms

BSA:

- utilizes all files to be aligned to build the translation model
- outputs only 1-1 beads
  Use intermediate results to extract 1-1, 1-2, 2-1 beads, discard 0-1, 1-0 beads. Simple heuristic to align the rest.

hunalign:

- utilizes single files to build the translation model
- outputs arbitrary non-intersecting beads

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

Questions

Which is more suitable, hunalign or bsa?

Questions

Which is more suitable, hunalign or bsa?

What is the impact of lemmatization on alignment quality?

Questions

Which is more suitable, hunalign or bsa?

What is the impact of lemmatization on alignment quality?

Does this depend on the language pairs chosen?

Questions

Which is more suitable, hunalign or bsa?

What is the impact of lemmatization on alignment quality?

Does this depend on the language pairs chosen?

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
  - Input module
  - Core component: Document and alignment repository
  - Alignment
  - Output module
  - State of the corpus
- Does lemmatization help alignment?
  - The alignment tools used in RPC
  - Questions
  - Experiment setup
  - Experiment: Polish-Russian and German-Russian
- Conclusions

Experiment

- Extract a set of randomly chosen segments of one language.
- Align them manually to the second language.
- This gives a gold standard of *right alignments* .

Experiment

- Extract a set of randomly chosen segments of one language.
- Align them manually to the second language.
- This gives a gold standard of *right alignments* .

Let the alignment tools align these texts, and compare the output to the gold standard.

Recall What proportion of the gold standard's 'right' alignments were output by the aligner?

Precision What proportion of the alignments output were ‘right’ alignments?

Fmeasure Harmonic mean of precision and recall.

Definitions

- Two texts: Text A and Text B with segments $(\mathrm{a}_i \ldots \mathrm{a}_j)$ and $(\mathrm{b}_i \ldots \mathrm{b}_j)$

Definitions

- Two texts: Text A and Text B with segments $(\mathrm{a}_i \ldots \mathrm{a}_j)$ and $(\mathrm{b}_i \ldots \mathrm{b}_j)$
- *Bead*: correspondence of segments of one text to segments of the other

Definitions

- Two texts: Text A and Text B with segments ($a_i \ldots a_j$) and ($b_i \ldots b_j$)
- *Bead*: correspondence of segments of one text to segments of the other

  1-1 bead: ({a_1},{b_1})

Definitions

- Two texts: Text A and Text B with segments $(a_i \ldots a_j)$ and $(b_i \ldots b_j)$
- *Bead*: correspondence of segments of one text to segments of the other
  1-1 bead: $(\{a_1\},\{b_1\})$
  1-2 bead: $(\{a_2\},\{b_2, b_3\})$

Definitions

- Two texts: Text A and Text B with segments $(a_i \ldots a_j)$ and $(b_i \ldots b_j)$
- *Bead*: correspondence of segments of one text to segments of the other
  1-1 bead: $(\{a_1\},\{b_1\})$
  1-2 bead: $(\{a_2\},\{b_2, b_3\})$
  0-1 bead: $(\{\},\{b_4\})$

Definitions

- Two texts: Text A and Text B with segments $(a_i \ldots a_j)$ and $(b_i \ldots b_j)$
- *Bead*: correspondence of segments of one text to segments of the other
  1-1 bead: $(\{a_1\},\{b_1\})$
  1-2 bead: $(\{a_2\},\{b_2, b_3\})$
  0-1 bead: $(\{\},\{b_4\})$
- *Alignment:* set of beads

Definitions

- Two texts: Text A and Text B with segments $(\mathrm{a}_i \ldots \mathrm{a}_j)$ and $(\mathrm{b}_i \ldots \mathrm{b}_j)$
- *Bead*: correspondence of segments of one text to segments of the other
  1-1 bead: $(\{\mathrm{a}_1\},\{\mathrm{b}_1\})$
  1-2 bead: $(\{\mathrm{a}_2\},\{\mathrm{b}_2, \mathrm{b}_3\})$
  0-1 bead: $(\{\},\{\mathrm{b}_4\})$
- *Alignment:* set of beads
  $A = \{(\{a_i\},\{b_i,b_i\}),\ (\{a_i,a_i\},\{b_i,b_i\}),\ (\{a_i\},\{b_i\}),\ \ldots\}$

What's a right alignment?

An example:

Line:1 Puść , nie chcę , żebyś mnie dotykał !

What's a right alignment?

An example:

Line:1 Puść , nie chcę , żebyś mnie dotykał !

Line:1 Пусти .
Line:2 Не хочу , чтобы ты ко мне прикасался .

What's a right alignment?

An example:

Line:1 Puść , nie chcę , żebyś mnie dotykał !

Line:1 Пусти .
Line:2 Не хочу , чтобы ты ко мне прикасался .

Gold standard: 1-2 bead $(\{1\},\{1,2\})$ would be perfect.

What's a right alignment?

An example:

Line:1 Puść , nie chcę , żebyś mnie dotykał !

Line:1 Пусти .
Line:2 Не хочу , чтобы ты ко мне прикасался .

Gold standard: 1-2 bead ({1},{1,2}) would be perfect.

Alignment 1 ({},{1}) ({1},{2}) would be less then perfect

What’s a right alignment?

An example:

Line:1 Puść , nie chcę , żebyś mnie dotykał !

Line:1 Пусти .
Line:2 Не хочу , чтобы ты ко мне прикасался .

Gold standard: 1-2 bead ({1},{1,2}) would be perfect.

Alignment 1 ({},{1}) ({1},{2}) would be less then perfect

Alignment 2 ({},{1, 2}) ({1},{3}) completely wrong

What's a right alignment?

An example:

Line:1 Puść , nie chcę , żebyś mnie dotykał !

Line:1 Пусти .
Line:2 Не хочу , чтобы ты ко мне прикасался .

Gold standard: 1-2 bead ({1},{1,2}) would be perfect.

Alignment 1 ({},{1}) ({1},{2}) would be less then perfect

Alignment 2 ({},{1, 2}) ({1},{3}) completely wrong

Alignment 3 ({0, 1, 2},{0,1,2,3}) – a very large (3-4) bead

What's a right alignment?

An example:

Line:1 Puść , nie chcę , żebyś mnie dotykał !

Line:1 Пусти .
Line:2 Не хочу , чтобы ты ко мне прикасался .

Gold standard: 1-2 bead ({1},{1,2}) would be perfect.

Alignment 1 ({},{1}) ({1},{2}) would be less then perfect

Alignment 2 ({},{1, 2}) ({1},{3}) completely wrong

Alignment 3 ({0, 1, 2},{0,1,2,3}) – a very large (3-4) bead

All three alignments would be assigned zero recall and precision.
But in fact, they are of very different quality.

Alignment metrics

For a more realistic picture, we use sentence-level metric (Véronis & Langlais 2000): we evaluate the cartesian product of the aligned segments:

Gold standard ({1},{1,2})
is transformed to ({1},{1}) ,({1},{2})

Alignment metrics

For a more realistic picture, we use sentence-level metric (Véronis & Langlais 2000): we evaluate the cartesian product of the aligned segments:

Gold standard ({1},{1,2})
is transformed to ({1},{1}) ,({1},{2})

Alignment 1 ({},{1}) ({1},{2})
transformed to ({},{1}) ({1},{2})
→recall 0.5, precision 0.5

Alignment metrics

For a more realistic picture, we use sentence-level metric (Véronis & Langlais 2000): we evaluate the cartesian product of the aligned segments:

Gold standard ({1},{1,2})
is transformed to ({1},{1}) ,({1},{2})

Alignment 1 ({},{1}) ({1},{2})
transformed to ({},{1}) ({1},{2})
→recall 0.5, precision 0.5

Alignment 2 ({},{1, 2}) ({1},{3})
transformed to ({},{1}) ({},{2}) ({1},{3})
→recall 0, precision 0

Alignment metrics

For a more realistic picture, we use sentence-level metric (Véronis & Langlais 2000): we evaluate the cartesian product of the aligned segments:

Gold standard ({1},{1,2})
is transformed to ({1},{1}) ,({1},{2})

Alignment 1 ({},{1}) ({1},{2})
transformed to ({},{1}) ({1},{2})
→recall 0.5, precision 0.5

Alignment 2 ({},{1, 2}) ({1},{3})
transformed to ({},{1}) ({},{2}) ({1},{3})
→recall 0, precision 0

Alignment 3 ({0, 1, 2},{0,1,2,3})
transformed ({0}, {0}), ({0}, {1}), ({0}, {2}), ({0}, {3}), ({1}, {0}), ({1}, {1}), ({1}, {2}), ({1}, {3}), ({2}, {0}), ({2}, {1}), ({2}, {2}), ({2}, {3})
→recall 2/2=1, precision 2/12 = 0.167

A look at Alignment 3

Problem: Because of random sampling, Alignment 3 covers sentences not in the gold standard.

Alignment 3 ({0, 1, 2},{0,1,2,3})
transformed ({0}, {0}), ({0}, {1}), ({0}, {2}), ({0}, {3}), ({1}, {0}), ({1}, {1}), ({1}, {2}), ({1}, {3}), ({2}, {0}), ({2}, {1}), ({2}, {2}), ({2}, {3})
→recall 2/2=1, precision 2/12 = 0.167

A look at Alignment 3

Problem: Because of random sampling, Alignment 3 covers sentences not in the gold standard.

Alignment 3 ({0, 1, 2},{0,1,2,3})
transformed ({0}, {0}), ({0}, {1}), ({0}, {2}), ({0}, {3}), ({1}, {0}), ({1}, {1}), ({1}, {2}), ({1}, {3}), ({2}, {0}), ({2}, {1}), ({2}, {2}), ({2}, {3})
→recall 2/2=1, precision 2/12 = 0.167

If, say, ({0},{0}) and ({2},{3}) were also right, just not covered in the gold standard, precision would rise.

→recall 4/4=1, precision 4/12 = 0.25

A look at Alignment 3

Problem: Because of random sampling, Alignment 3 covers sentences not in the gold standard.

Alignment 3 ({0, 1, 2},{0,1,2,3})
transformed ({0}, {0}), ({0}, {1}), ({0}, {2}), ({0}, {3}), ({1}, {0}), ({1}, {1}), ({1}, {2}), ({1}, {3}), ({2}, {0}), ({2}, {1}), ({2}, {2}), ({2}, {3})
→recall 2/2=1, precision 2/12 = 0.167

If, say, ({0},{0}) and ({2},{3}) were also right, just not covered in the gold standard, precision would rise.

→recall 4/4=1, precision 4/12 = 0.25

A look at Alignment 3

Problem: Because of random sampling, Alignment 3 covers sentences not in the gold standard.

Alignment 3 ({0, 1, 2},{0,1,2,3})
transformed ({0}, {0}), ({0}, {1}), ({0}, {2}), ({0}, {3}), ({1}, {0}), ({1}, {1}), ({1}, {2}), ({1}, {3}), ({2}, {0}), ({2}, {1}), ({2}, {2}), ({2}, {3})
→recall 2/2=1, precision 2/12 = 0.167

If, say, ({0},{0}) and ({2},{3}) were also right, just not covered in the gold standard, precision would rise.

→recall 4/4=1, precision 4/12 = 0.25

Random sampling results in low precision of large beads because not all alignments of a file are evaluated.

A look at Alignment 3

Problem: Alignment 3 = ({0, 1, 2},{0,1,2,3}) covers sentences not in the gold standard.

A look at Alignment 3

Problem: Alignment 3 = ({0, 1, 2},{0,1,2,3}) covers sentences not in the gold standard.

Solution: Discard all combinations that do not relate to sentences covered in the gold standard

A look at Alignment 3

Problem: Alignment 3 = ({0, 1, 2},{0,1,2,3}) covers sentences not in the gold standard.

Solution: Discard all combinations that do not relate to sentences covered in the gold standard

Leave out: ({0}, {0}), ({0}, {3}), ({2}, {0}), ({2}, {1}), ({2}, {2}), ({2}, {3})

Keep: ({0}, {1}), ({0}, {2}), ({1}, {0}), ({1}, {1}), ({1}, {2}), ({1}, {3}),

→recall 2/2=1, precision 2/6 = 0.33

A look at Alignment 3

Problem: Alignment 3 = ({0, 1, 2},{0,1,2,3}) covers sentences not in the gold standard.

Solution: Discard all combinations that do not relate to sentences covered in the gold standard

Leave out: ({0}, {0}), ({0}, {3}), ({2}, {0}), ({2}, {1}), ({2}, {2}), ({2}, {3})

Keep: ({0}, {1}), ({0}, {2}), ({1}, {0}), ({1}, {1}), ({1}, {2}), ({1}, {3}),

→recall 2/2=1, precision 2/6 = 0.33

This still penalizes VERY large beads, but gives the benefit of the doubt for medium sized beads.

Summary: evaluation technique

- Compare aligner otput against manual alignment of randomly chosen segments.
- Do this on a sentence-correspondence level, discarding beads without relation to the gold standard.
- Measure recall, precision, fmeasure.

Summary: evaluation technique

- Compare aligner otput against manual alignment of randomly chosen segments.
- Do this on a sentence-correspondence level, discarding beads without relation to the gold standard.
- Measure recall, precision, fmeasure.

Outline

- Introduction: Aims and architecture of the CWB
- System architecture
    - Input module
    - Core component: Document and alignment repository
    - Alignment
    - Output module
    - State of the corpus
- Does lemmatization help alignment?
    - The alignment tools used in RPC
    - Questions
    - Experiment setup
    - Experiment: Polish-Russian and German-Russian
- Conclusions

Polish-Russian

LemAstronauci
LemFiasko
LemKongres
LemPowGwiazd
LemSolaris
LemWizjaLokalna

BulgakovMaster
StrugLebedi
StrugPiknik

Potter1
Potter2

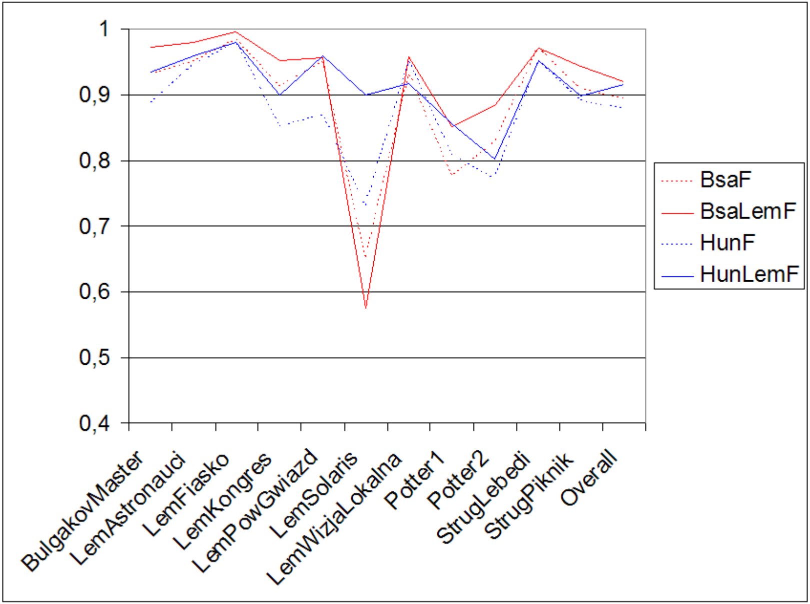
1 million tokens, 77 000 sentences, a sample of 1000 sentences

German-Russian

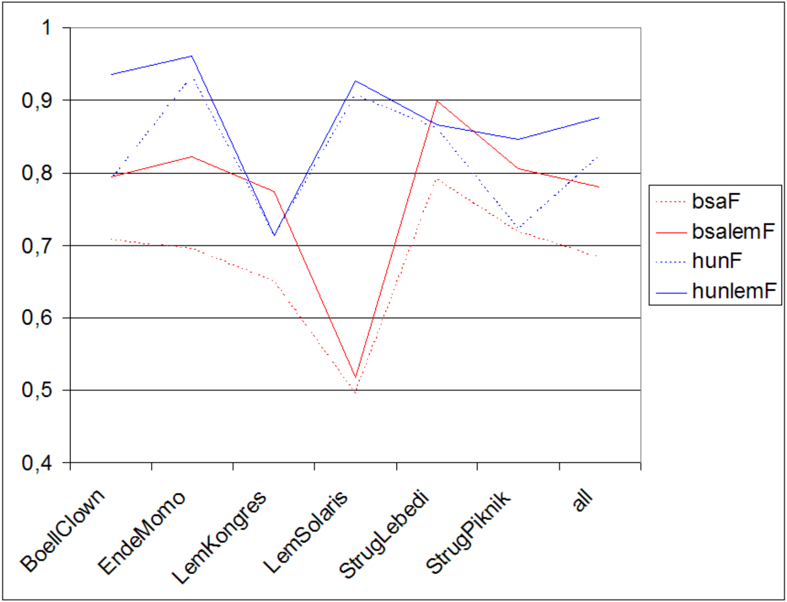
BoellClown
EndeMomo
LemKongres
LemSolaris
StrugLebedi
StrugPiknik

0.4 million tokens, 33 000 sentences, a sample of 500 sentences

F-measure: PL-RU



F-measure: DE-RU



More detailed

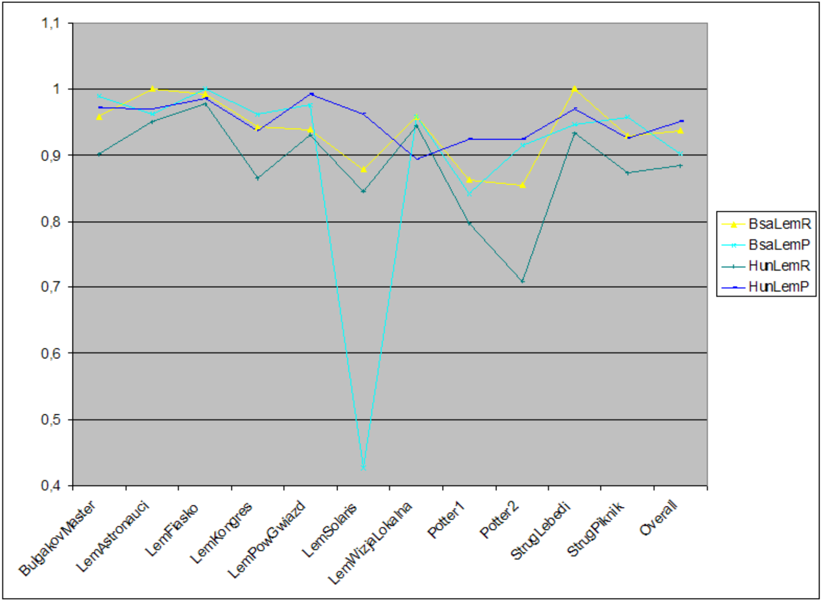
PL-RU

	Recall		Precision		f-measure
hun	0.84		0.91		0.88
hunlem	0.88	-25%	0.95	-45%	0.92
bsa	0.91		0.88		0.89
bsalem	0.94	-33%	0.90	-16%	0.92

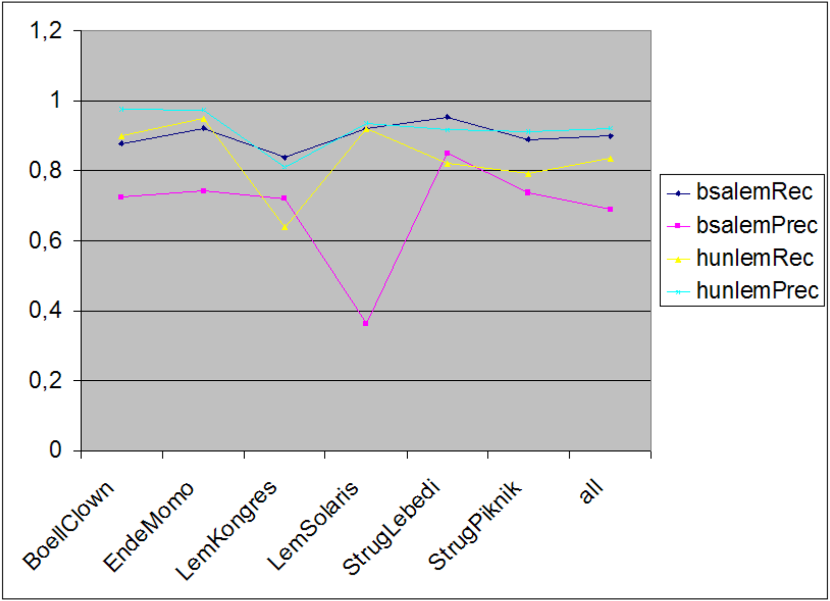
DE-RU

	Recall		Precision		f-measure
hun	0.78		0.88		0.82
hunlem	0.84	-27%	0.92	-33%	0.88
bsa	0.89		0.56		0.68
bsalem	0.90	-10%	0.69	-30%	0.78

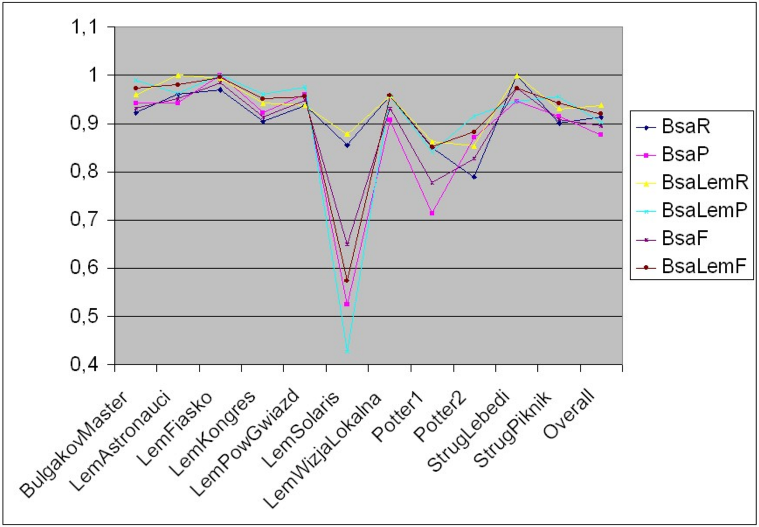
Recall/Precision: PL-RU



Recall/Precision: DE-RU



Experiment: Polish-Russian



Conclusions

- Lemmatization DOES lead to better alignment

Conclusions

- Lemmatization DOES lead to better alignment
- Alignment is more or less difficult depending on the language pair

Conclusions

- Lemmatization DOES lead to better alignment
- Alignment is more or less difficult depending on the language pair
- Alignment quality VERY dependent on text (see Rosen 2005)

Conclusions

- Lemmatization DOES lead to better alignment
- Alignment is more or less difficult depending on the language pair
- Alignment quality VERY dependent on text (see Rosen 2005)
- Sometimes Hunalign, sometimes BSA better (see Rosen 2005)

Conclusions

- Lemmatization DOES lead to better alignment
- Alignment is more or less difficult depending on the language pair
- Alignment quality VERY dependent on text (see Rosen 2005)
- Sometimes Hunalign, sometimes BSA better (see Rosen 2005)
- Influence of evaluation method?

Thank you!

References

Garabik, Radovan (2005): Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. Proceedings of Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005, p. 50 – 58.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, pages 135–144, London, UK. Springer-Verlag.

Rosen, A. (2005): In Search of the Best Method for Sentence Alignment in Parallel Texts. In: Computer Treatment of Slavic and East European Languages. Proceedings of Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005.

Varga Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh & Viktor Trón (2005). Parallel corpora for medium density languages. Proceedings of RANLP'2005. Borovets, Bulgaria, pp. 590-596.