Semantic relations in a bilingual corpus of different registers

Oliver Čulo¹, Kerstin Kunz² & Torsten Zesch³
¹Johannes Gutenberg University, Mainz
²Saarland University, Saarbrücken
³UKP Lab, TU Darmstadt

05/03/09
Overview

- Motivation
- Research questions
- Operationalisation of indicators for register variation with respect to semantic relations
- Analysis design
  - Corpus design
  - Annotation
- Problems = conclusion + outlook
Motivation (1)

Research into register variation on semantic level of language

Research into register variation on linguistic level of cohesion

⇒ Insight into texture of text in different registers
Motivation (2)

- evaluate and enhance lexical chaining module in DKPro

- CL perspective: computational insight into one aspect of textuality
Research questions

■ Intralingual perspective: How do registers vary in one language?

■ Crosslinguistic perspective:
  ■ How do registers vary across languages?
  ■ How does crosslinguistic variation differ from intralingual variation?

■ Translation perspective:
  ■ Which shifts between translation and original are due to register differences?
Lexical cohesion as indicator of properties of register

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Subdimension</th>
<th>Operationalisation</th>
<th>Textual indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tenor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

05/03/09

DGFS09
Lexical cohesion as indicator of properties of register

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Subdimension</th>
<th>Operationalisation</th>
<th>Textual indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td>Experiential domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Goal orientation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tenor</td>
<td>Social hierarchy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>Language role</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Lexical cohesion as indicator of properties of register

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Subdimension</th>
<th>Operationalisation</th>
<th>Textual indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td>Experiential domain</td>
<td>Domain type</td>
<td>Referent types in lexical chains</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Domain continuity</td>
<td>Chain interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Domain progression</td>
<td>Textual distance between elements in one chain</td>
</tr>
<tr>
<td></td>
<td>Goal orientation</td>
<td>Narration ⇔ argumentation</td>
<td>Frequency of chains/ elements in one chain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Syntactic function and position of nouns</td>
<td>Recurrence ⇔ semantic relations</td>
</tr>
</tbody>
</table>

05/03/09
Lexical cohesion as indicator of properties of register

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Subdimension</th>
<th>Operationalisation</th>
<th>Textual indicators</th>
</tr>
</thead>
</table>
| Tenor     | Social hierarchy   | level of expertise | specific ↔ general nouns
|           |                    |                    | Hyperonymy, Hyponymy ↔ recurrence, synonymy |
# Lexical cohesion as indicator of register variation

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Subdimension</th>
<th>Operationalisation</th>
<th>Textual indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>Medium</td>
<td>Spoken ⇔ written</td>
<td>Ellipsis, substitution ⇔ lexical cohesion</td>
</tr>
<tr>
<td></td>
<td>Language role</td>
<td>Ancillary ⇔ constitutive</td>
<td>Demonstrative reference ⇔ lexical cohesion</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time/ space nouns ⇔</td>
</tr>
</tbody>
</table>
Annotation data

Corpus design
Analysis Design
Segmentation & Alignment
The CroCo Corpus

Annotation

- Translations
  - morphology
  - POS
  - semantic relations
  - grammatical functions
  - phrase structure
  - chunks

- Originals
  - morphology
  - POS
  - semantic relations
  - grammatical functions
  - phrase structure
  - chunks

05/03/09

DGFS09
Annotation of semantic relations

I Automatic annotation

II Manual annotation
## GLexi or DKPro?

<table>
<thead>
<tr>
<th>GLexi</th>
<th>DKPro lexical chainer</th>
</tr>
</thead>
<tbody>
<tr>
<td>various algorithms ready at hand</td>
<td>mainly two algorithms, still (but rapidly) evolving</td>
</tr>
<tr>
<td>already evaluated (Cramer &amp; Finthammer 2008)</td>
<td>evaluation ongoing</td>
</tr>
<tr>
<td>German only (English planned)</td>
<td>English &amp; German</td>
</tr>
</tbody>
</table>
DKPro annotation pipeline

Web

PDF

Knowledge source

Low-level Preprocessing

PDF reader, Wikipedia reader, Forum reader, Language Identification, Data cleansing

Low-level Preprocessing

Tokenizer, Sentence splitter, Stopword tagger

Morphological Analysis

Stemmer, Lemmatizer, Compound splitter

Syntactic Analysis

PoS-Tagger, Parser, Phrase chunker

Semantic Analysis

Named Entity tagger, Sentiment detector, Word sense disambiguation, Lexical chain annotator

Project-specific Analysis

Email analysis, Named Entity disambiguation

Data export

Database export, XML, HTML, Indexer (Lucene, Terrier), ARFF export
CroCo-adapted DKPro pipeline

CroCo

Knowledge source

Low-level Preprocessing

Low-level Preprocessing

Morphological Analysis

Syntactic Analysis

Semantic Analysis

Project-specific Analysis

Data export

PDF reader, Wikipedia reader, Forum reader, Language identification, Data cleansing

Tokenizer, Sentence splitter, Stopword tagger

Stemmer, Lemmatizer, Compound splitter

PoS-Tagger, Parser, Phrase chunker

Named Entity tagger, Sentiment detector, Word sense disambiguation, Lexical chain annotator

Email analysis, Named Entity disambiguation

Database export, XML, HTML, Indexer (Lucene, Terrier), ARFF export

05/03/09
Lexical Chaining Architecture

Lexical Chaining Algorithms
- Silber & McCoy (2002)
- ...

Lexical Semantic Resource Interface
- JWPL
- JWKTL
- JWNL
- GN API
- ?

Wikipedia
- Wiktionary
- WordNet
- GermaNet
- ...

05/03/09
Disambiguation in the Galley & McKeown Algorithm

#1: autumn

#2: spill, tumple

#3: descent, declivity

FALL

#1: financial institution

#2: slope

BANK

#3: reserve

CREDIT UNION

ACQUIRER

05/03/09

DGFS09
Annotation configuration

- DKPro components can be configured
- CroCo configuration set to slight overgeneration, followed by manual filtering
The five of us are a team.
Since the glass felt less than cool, I suspect that the massive commercial refrigerator confines itself to generating noise and vibration.
I know this, having run my hand over every row of bottles.
None of the drinks is chilled.
It's bitter-sweet and lukewarm.
The cafeteria's refrigerator contains my own favourite beverage, a branded lemonade with citrus fruit extracts.
How quickly those two pseudo-beer drinkers have made friends with one another.
He now clinks bottles with the fleshy man, who tops him by more than a head, and slaps him on the shoulder.
Of medium height at most, the fellow swings his arms continuously, perhaps in an attempt to mitigate the impression created by their inordinate length.
If I interpret his figure as the longterm product of such invigoration, persistent consumption of this brew appears to stimulate excessive growth in the upper extremities.
The other man standing beside the refrigerator is also drinking a species of beer, the sweet stuff that purports to invigorate young and old alike.
Occasionally the hand veers upwards, and he wags the outstretched finger just in front of his impressive nose, the tip of which displays a peculiarly asymmetrical bulge.
While speaking he holds his right fist in front of his chest and shoots out his forefinger.
The slender bottle tips its contents into his mouth at brief, regular intervals, as if his drinking is subject to a dictatorial schedule.
Manual annotation

(1) Correction: Disambiguation of lexical chains

(2) Annotation: Type of semantic relation
Manual annotation (1)

Correction

⇒ Disambiguation of lexical chains

- POS
- Sense relation
- Lexical chains
Manual annotation (2)

Annotation

⇒ Type of semantic relation:

- Recurrence
- Synonymy
- Hyponymy
- Hyperonymy
- Co-Hyponymy
- Holonomy
- Meronymy
- Co-Meronymy
- Antonymy

05/03/09
transition, one that excluded you from humanity, so bewildered he was at finding himself in this purgatory with half a dozen complete strangers—rocky looking lads sitting round a stove in a roofless house.

Clearly, to be issued with a special pass which allowed him to explore the categories of the Serbian lines were sometimes not much more than a desolate scene, a group of houses with the windows bent arched to the ground. One was a burnt-out car tipped onto its roof, a zigzag line of Pannonian sky.

'That,' I said, 'is a dissolution to the winds,' I said.

'That,' he said, 'is when he saw. He had got Allmayer wrong, he couldn't be sure he would go home again.'

'Otherwise he probably would have been having a good story,' he said.

And now Allmayer had recorded near Vinkovci a few days before Christmas in the first year of the war which always gave me the shivers every time I read it, he could have been talking to the man who was to be his murderer, been asking him what it was like, aiming at a human being, what the feeling was when a head appeared all at once in your sights and you had your finger on the trigger.

'You couldn't anticipate your own death more precisely than he did.'

'If you were superstitious, you might think he brought it down on himself,' I said, 'although I realised it was stupid.

The previous evening I had read about the boy bleeding to death in a ditch with a growing feeling of
woman who had somehow got caught up in the turmoil of the war, best of all an American, who by her mere existence would turn a dreary Balkan report into an exciting story.

It sounded as if he was talking about his novel again and I let him get on with it without interrupting, although I remember how incomprehensible I found his persistence, his compulsive urge to keep coming back to it at the most impossible moments.

"You couldn't ask for a better starting point."

"Perhaps we can build something on it," he said, not sounding very convinced.

It was quiet and I didn't hear Paul laugh when I said that, but I noticed he had to restrain himself, given the quizzical look he suddenly turned on me.

"I don't see where that gets us."

Whatever the reason, it only took up a few lines in Allmayer's article and it ended abruptly with his claim that he didn't see how it finished, just heard three distinct, separate shots from a distance, a rather open ending, true, but it struck me..."
"Das kann ich nicht sagen", erwiderte er.

Doch darüber zu reden, war müssig, Wortgeklirper, das einen nicht weiterbrachte, und ich bat ihn, damit aufzuhoeren und mir lieber zu erklären, nach welchen Kriterien die Stellen in dem Konvolut angestrichen hatte, das vor uns auf dem Tischchen lag.

Als haette er es geahnt, dachte ich und wusste zugleich, wie unsinnig das war, als haette sein ganzes Hin und Her nur darauf abgezielt, ihn zur falschen Zeit an den falschen Ort zu bringen, wo ihn sein Schicksal ereilte, wie es dann immer hiess, und es stimmt, was Paul sagte ueber das Davor und Danach bei solchen Katastrophen, er hatte recht, man konnte sich dem Zeitpunkt, wo jemand zu Tode kam, beliebig annaehern, konnte das Intervall zwischen dem Augenblick, in dem er noch gelebt hatte, und jenem, in dem er schon tot war, so lange verkleinern, bis man es fast nicht mehr aushielt, zu denken, dass dazwischen ueberhaupt etwas geschehen sein sollte.

"Wenn man es allen recht machen will, stimmt am Ende meistens nicht mehr viel."

"Das liegt an der Vorgabe der Zeitung", sagte er.

Es war daher fuer mich das erste, Paul zu fragen, ob er bemerkt hatte, wie schlecht Teile davon geschrieben waren, als ich ihn am naechsten Morgen traf, aber er winkte bloss ab.

Das waren nur ein paar Beispiele, aber wenn er dann auch noch auf das sogenannte Madchen aus Sarajevo hereinfiel und sein Tagebuch, ein kitschiges Elaborat, das um die Welt gegangen war, wenn er daraus zu Tränen geruehrt zitierte und nicht hoeren wollten, wie falsch der Satz Liebe Mimmy, die politische Lage ist bescheuert von einer Dreizehnjaehrigen war, konnte man nuf, den Kopf schuetteln und ihm resigniert das Verdikt eines Kriegsherren unter die Nase reiben, das er selbst festgehalten hatte, seine Absage an alles billige Moralisieren, mehr noch, seine Belustigung ueber jegliche Bedenken in Zeiten des Krieges, den Ausspruch, sie waeren laecherlich, ein grotesker Luxus, hoechstens etwas fuer Dummkopfe und Amerikaner.

Um so verwunderlicher war es, dass er selbst so oft danebengriff, dass er immer gleich die Ustascha herbeigeschwoeren
'And it doesn't matter, either, since for me there's not much doubt what is the most important thing about the whole collection.'
'I couldn't say,' he replied.
But talking about it was futile, empty _verbiage_ that got us nowhere, so I asked him to stop and explain instead the criteria by which he had marked the _passages_ in the _sheaf_ of _papers_ on the _table_ in front of us.
As if he had had a _premonition_ of what was going to happen, I thought, aware at the same time how meaningless that was, as if the sole _purpose_ of all his toing and froing was to get him to the wrong _place_ at the wrong _time_, when he met his _fate_, and it was true what Paul said about the before and after of such _disasters_, he was right, you could approach the _point_ at which someone died from whichever _way_ you liked, you could keep reducing the _gap_ between the _moment_ when they were still alive and the _moment_ when they were already dead until the _thought_ that anything at all could have happened in between became almost too much to bear.
'If you're trying to please everyone, most _things_ get twisted.'
'It's the _newspaper's_ house _style_ that's to blame,' he said.
So the first _thing_ I did when I met Paul the next morning, was to ask him whether he had noticed how poorly _parts_ of the _articles_ were written, but he just waved my _question_ away.
Those were just a few _examples_, but when, to _cap_ it all, he was even taken in by the supposed _girl_ from Sarajevo and her diary, a kitschy _concoction_ that had gone all over the _world_, when he quoted from it, moved to _tears_ and refusing to accept how false a _sentence_ like _Dear_ Mommy, _The political situation_ is crazy sounded from the _pen_ of a thirteen-year-old, then all you could do was _shake_ your _head_ and quote at him the _verdict_, which he himself had recorded, one of the _warlords_, his rejection of easy moralising, more than that, his _amusement_ that anyone should feel any moral _scruples_ in _wartime_: they were, he had said, _ridiculous_, a _protesque_ _luxury_, at most something for _idiots_ and _Americans_.
It was all the more surprising that he himself struck a wrong note so often, that he kept insisting on bringing in the Ustasa at every possible _opportunity_ and going on at similarly excessive length about the Chetniks, that he _couldn't_ call a _rifle_ a _rifle_, it had to be a Kalashnikov, especially when it was a _woman_ holding it and you could tell between the _lines_ how it both repelled and aroused him, or that he _couldn't_ meet anyone without drinking _slivovitz_ with them, I counted at least two _dozen instances_ in his _articles_.
Whether you thought that was glossing _things_ over or not, arranging _things_ so as not to disturb his _readers_ in their comfortable _armchairs_, that probably expressed something that was always there in his later _interviews_?: _irritation_ at any
Problems

- How to filter lexical cohesion of all lexical relations?
- Technical problem in MMAX2: visualization of chain interaction
- Word senses: gradual shifts ⇔ clear-cut shifts
References


In: Proceedings of the KONVENS Workshop on Lexical-Semantic and Ontological Resources -- Maintenance, Representation, and Standards


References (2)


http://fr46.uni-saarland.de/croco/