CROCO

Ubiquitous
Knowledge
Processing

Project No. STE 840/5-1
sponsored by

Deutsche
Forschungsgemeinschaft
DFG

LINGUISTIC PROPERTIES OF TRANSLATIONS
A CORPUS-BASED INVESTIGATION FOR THE LANGUAGE PAIR ENGLISH-GERMAN

Semantic relations in a bilingual corpus of different registers

Oliver Čulo[1], Kerstin Kunz[2] & Torsten Zesch[3]
[1]Johannes Gutenberg University, Mainz
[2]Saarland University, Saarbrücken
[3]UKP Lab, TU Darmstadt

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

Dimension	**Subdimension**	**Operationalisation**	**Textual indicators**
Field			
Tenor			
Mode			

Lexical cohesion as indicator of properties of register

Dimension	Subdimension	Operationalisation	Textual indicators
Field	Experiential domain		
	Goal orientation		
Tenor	Social hierarchy		
Mode	Language role		

Lexical cohesion as indicator of properties of register

Dimension	Subdimension	Operationalisation	Textual indicators
Field	Experiential domain	Domain type Domain continuity Domain progression	Referent types in lexical chains Chain interaction Textual distance between elements in one chain Frequency of chains/ elements in one chain
	Goal orientation	Narration ⬄ argumentation	Recurrence ⬄ semantic relations Syntactic function and position of nouns

Lexical cohesion as indicator of properties of register

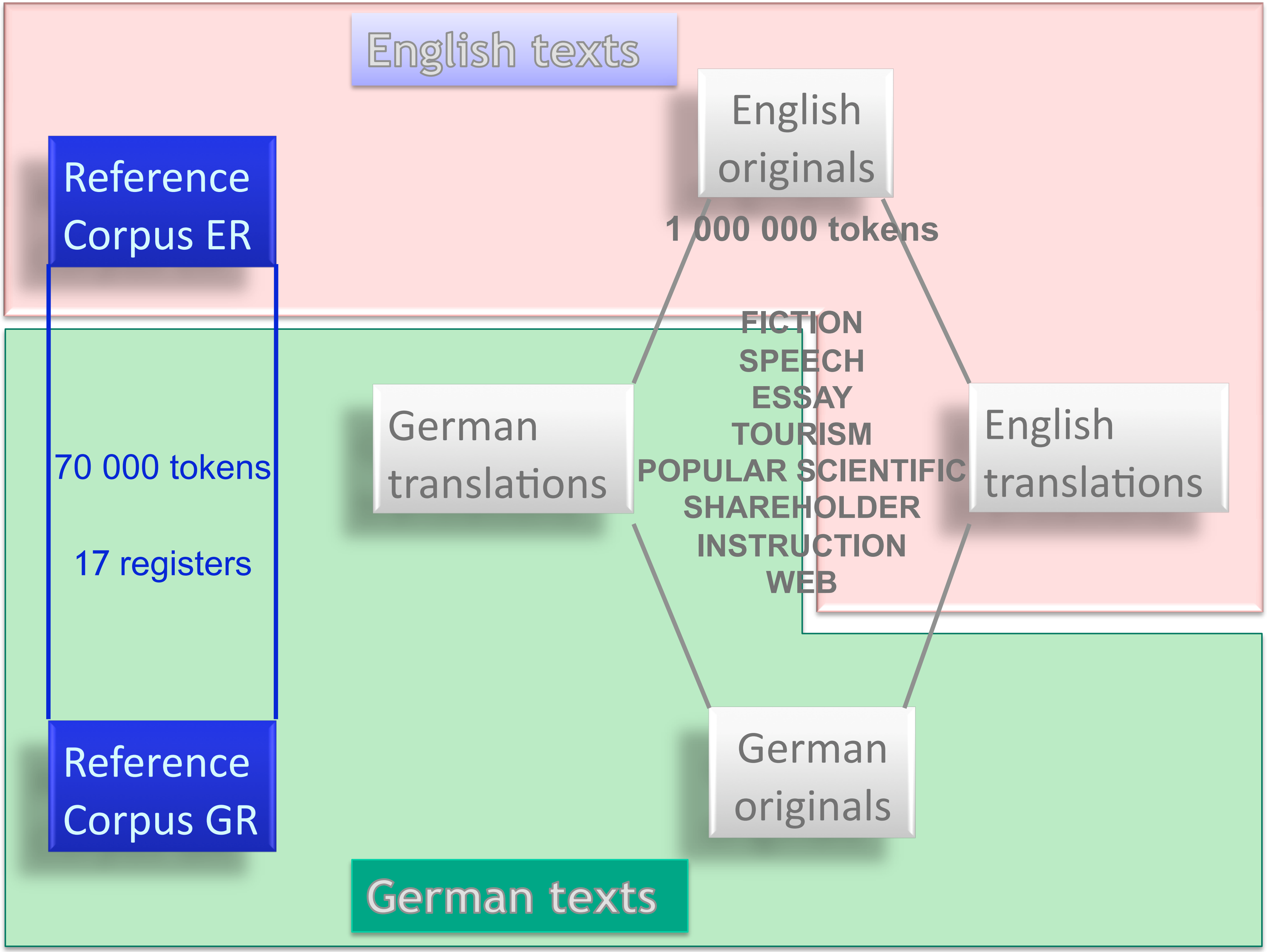
Dimension	Subdimension	Operationalisation	Textual indicators
Tenor	Social hierarchy	level of expertise	specific ⇔ general nouns Hyperonymy, Hyponymy ⇔ recurrence, synonymy

Lexical cohesion as indicator of register variation

Dimension	Subdimension	Operationalisation	Textual indicators
Mode	Medium	Spoken ⇔ written	Ellipsis, substitution ⇔ lexical cohesion
	Language role	Ancillary⇔ constitutive	Demonstrative reference ⇔ lexical cohesion Time/ space nouns ⇔

Annotation data

Corpus design



English texts

Reference Corpus ER

English originals

1 000 000 tokens

FICTION
SPEECH
ESSAY
TOURISM
POPULAR SCIENTIFIC
SHAREHOLDER
INSTRUCTION
WEB

German translations

English translations

70 000 tokens

17 registers

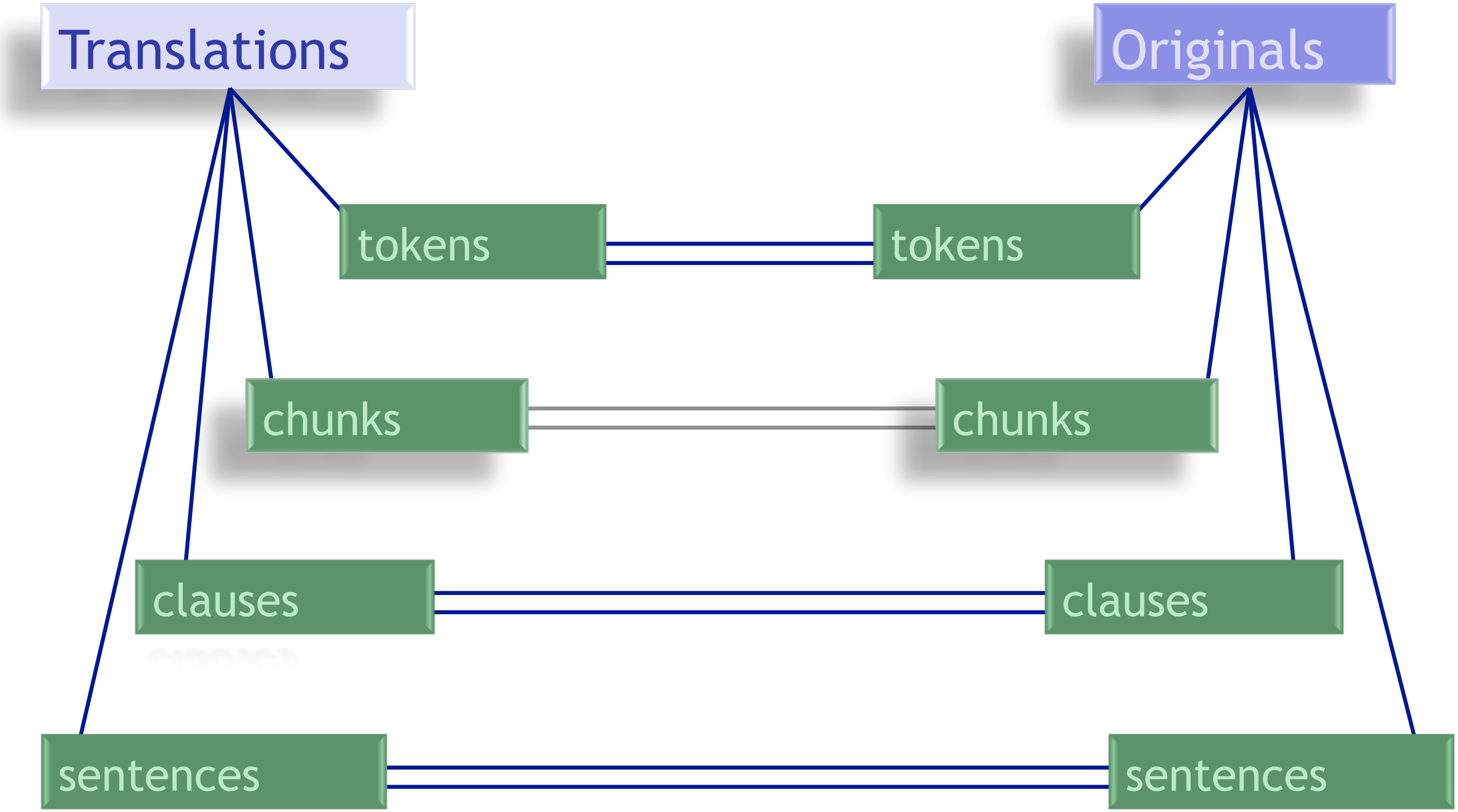
Reference Corpus GR

German originals

German texts

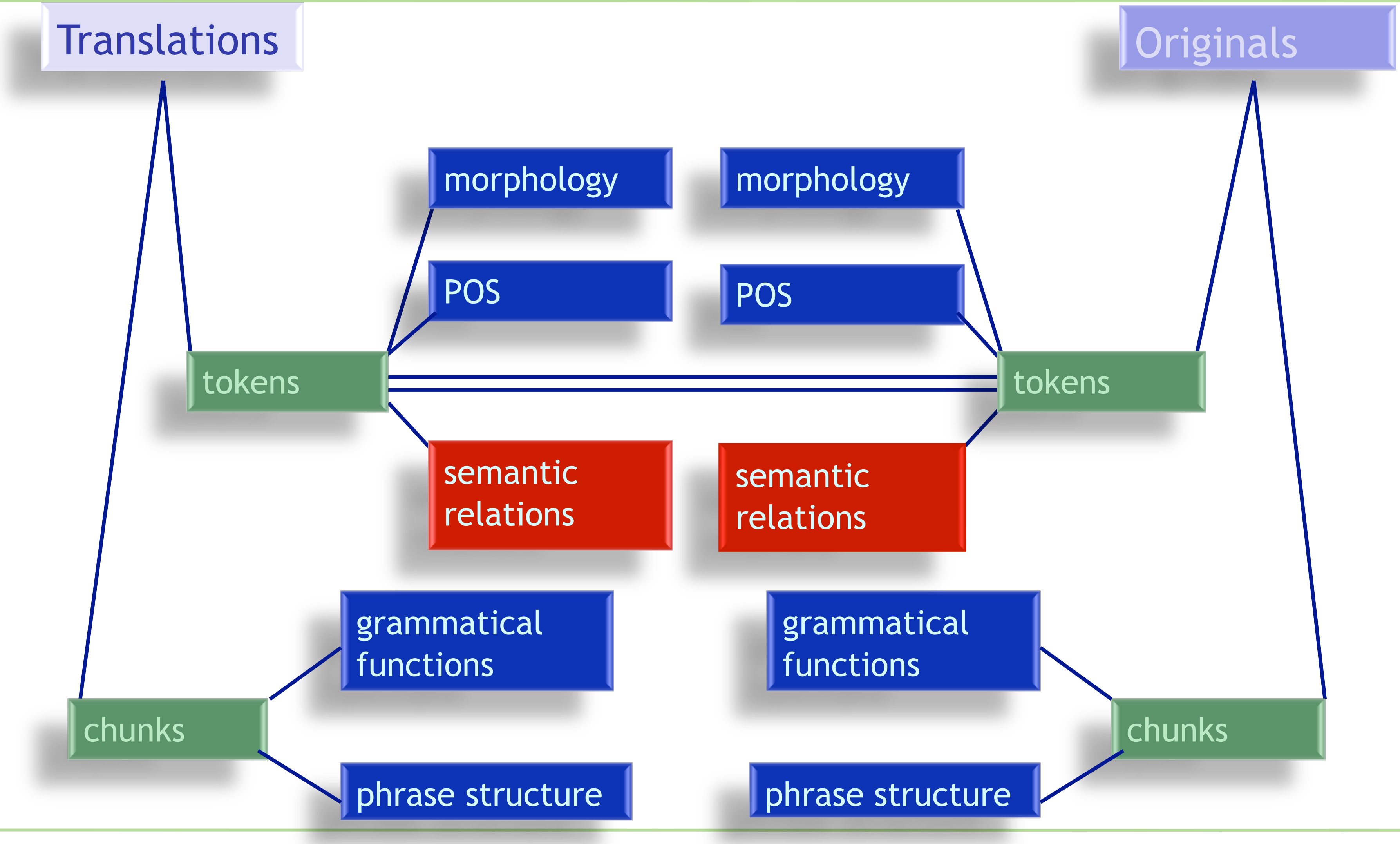
Analysis Design

Segmentation & Alignment



The CroCo Corpus

Annotation



Annotation of semantic relations

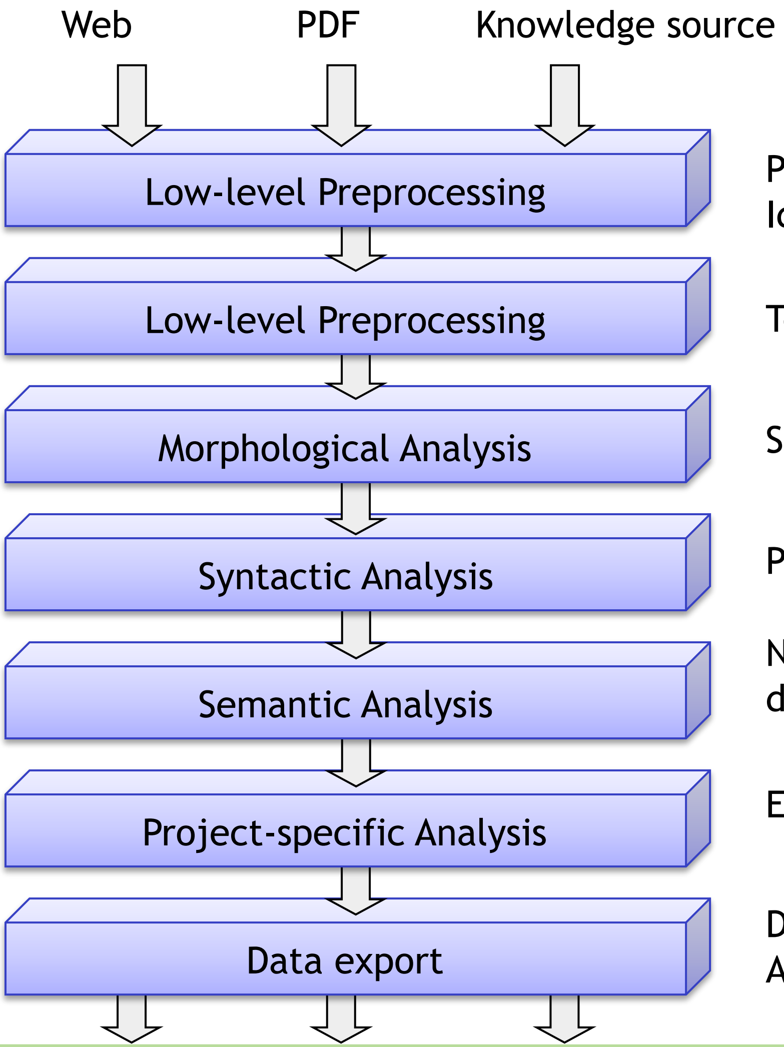
I Automatic annotation

II Manual annotation

GLexi or DKPro?

GLexi	DKPro lexical chainer
various algorithms ready at hand	mainly two algorithms, still (but rapidly) evolving
already evaluated (Cramer & Finthammer 2008)	evaluation ongoing
German only (English planned)	**English & German**

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

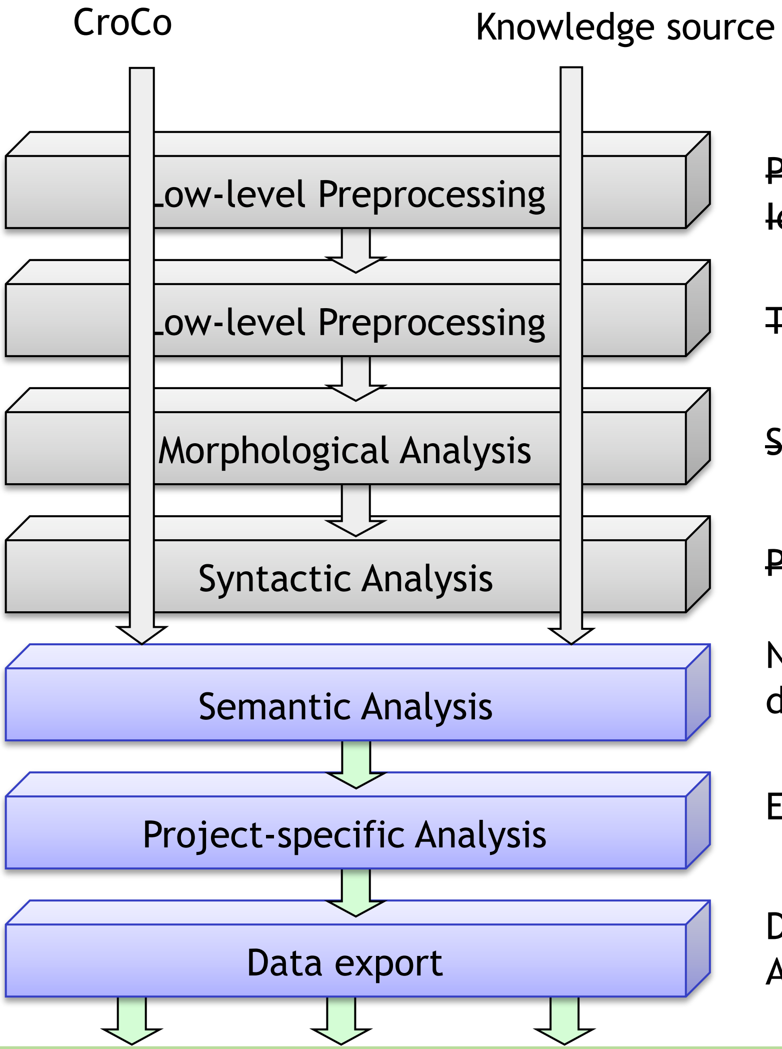
Project-specific Analysis

Email analysis, Named Entity disambiguation

Data export

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

CroCo-adapted DKPro pipeline



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

~~Tokenizer, Sentence splitter, Stopword tagger~~

~~Stemmer, Lemmatizer, Compound splitter~~

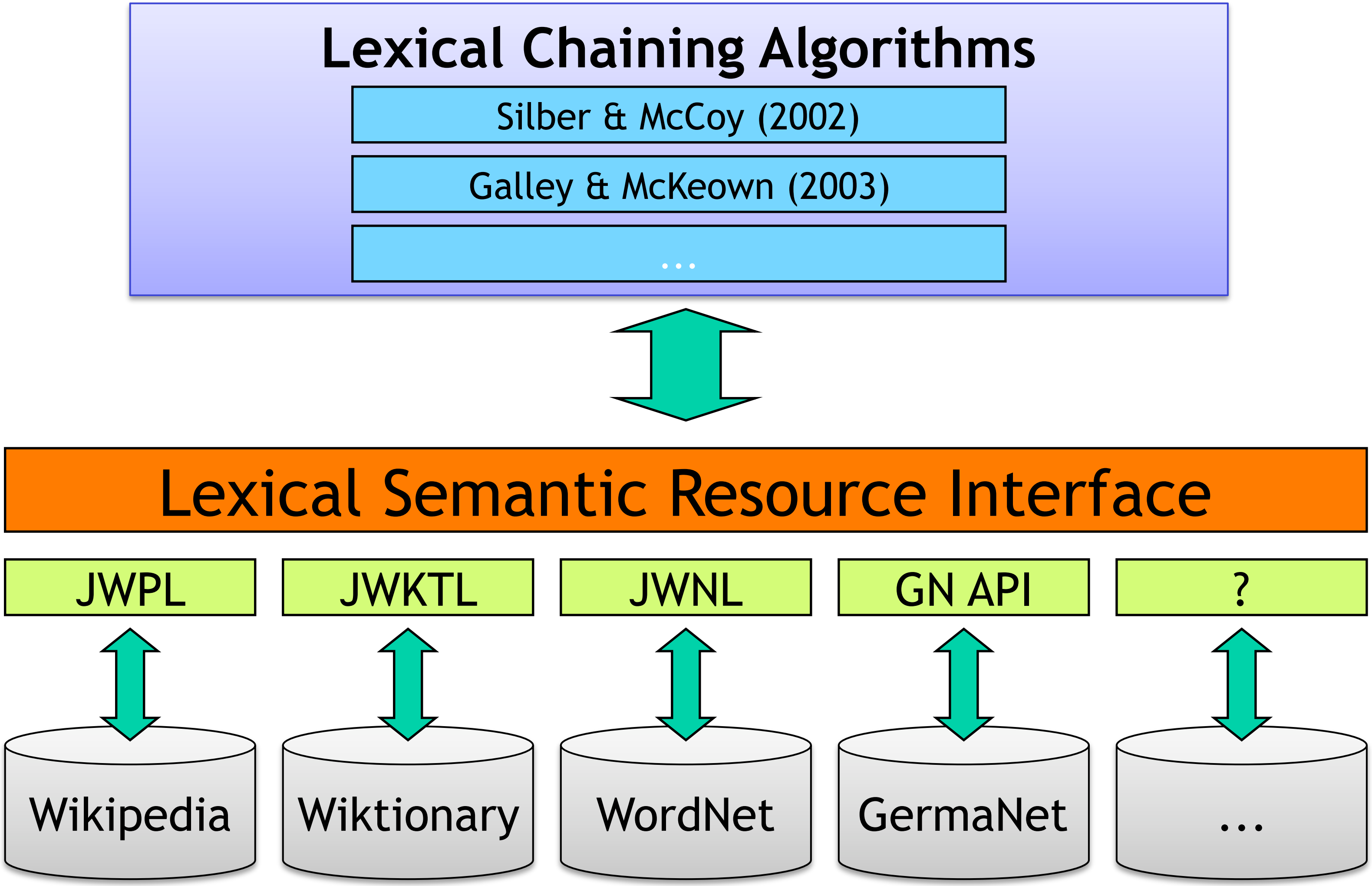
~~PoS-Tagger, Parser, Phrase chunker~~

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

Lexical Chaining Architecture



Lexical Chaining Algorithms

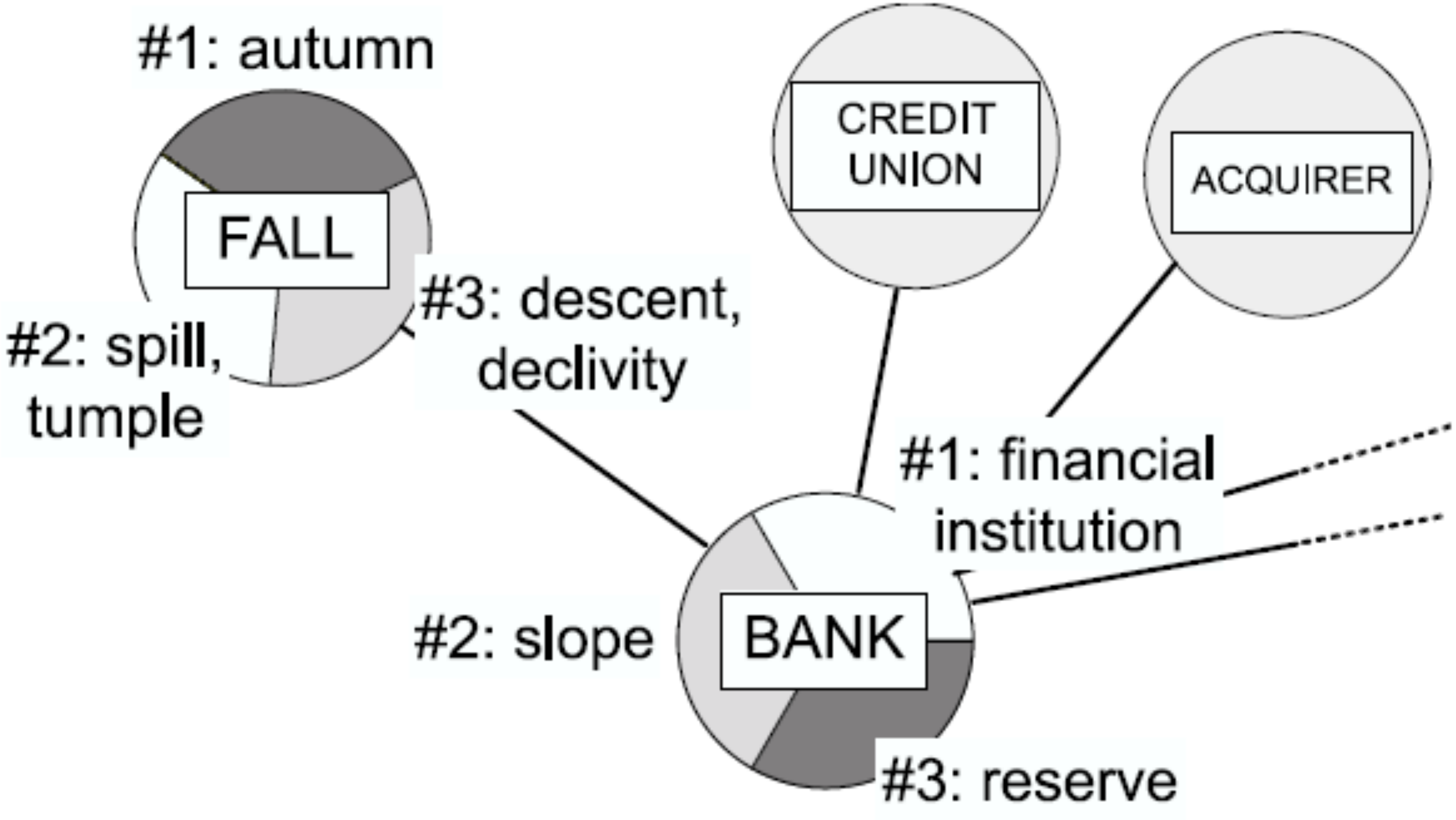
Silber & McCoy (2002)

Galley & McKeown (2003)

…

Lexical Semantic Resource Interface

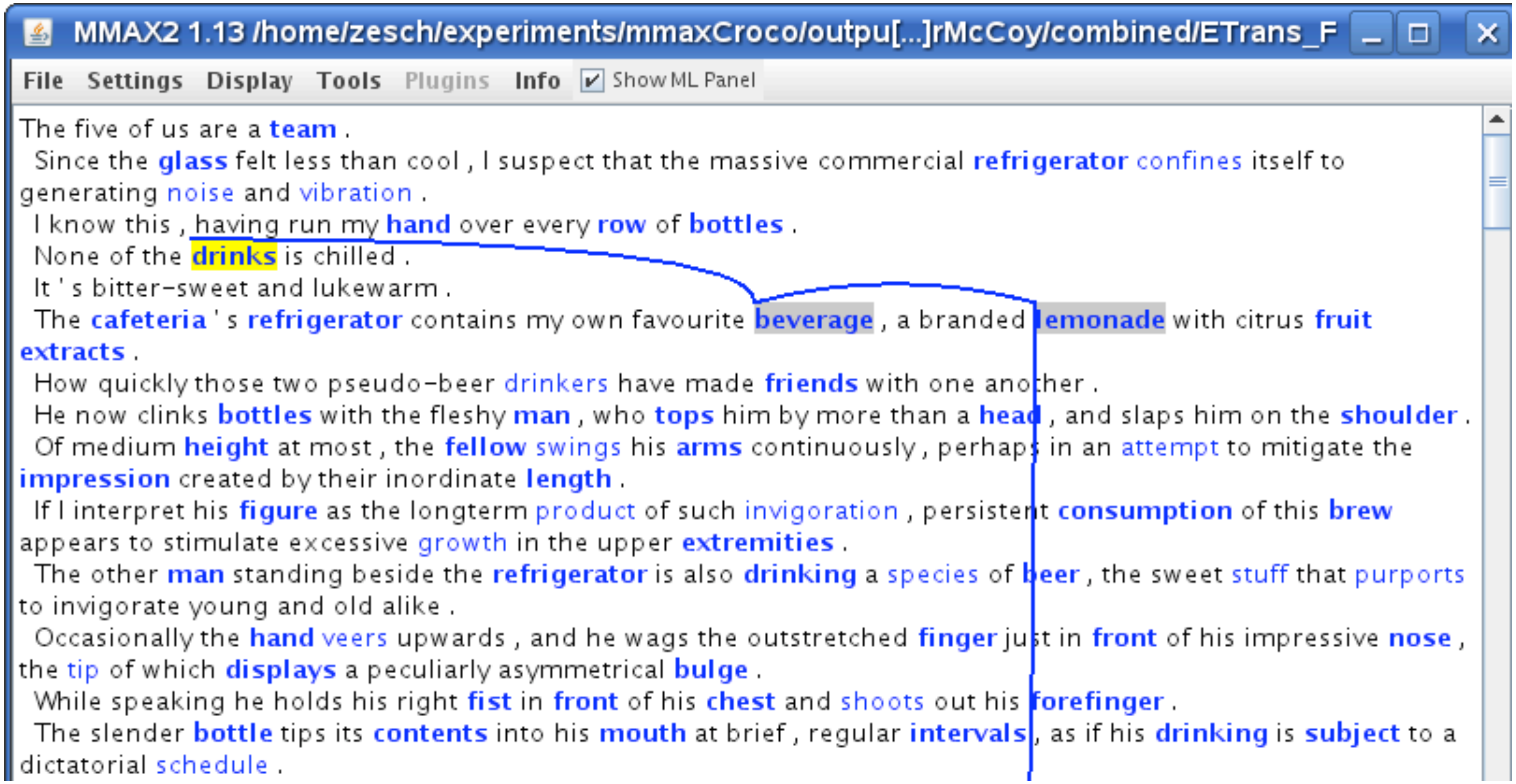
Disambiguation in the Galley & McKeown Algorithm



Annotation configuration

- DKPro components can be configured
- CroCo configuration set to slight overgeneration, followed by manual filtering

Representation of Lexical Chains in MMAX2


MMAX2 1.13 /home/zesch/experiments/mmaxCroco/outpu[...]rMcCoy/combined/ETrans_F
File Settings Display Tools Plugins Info
☑ Show ML Panel

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

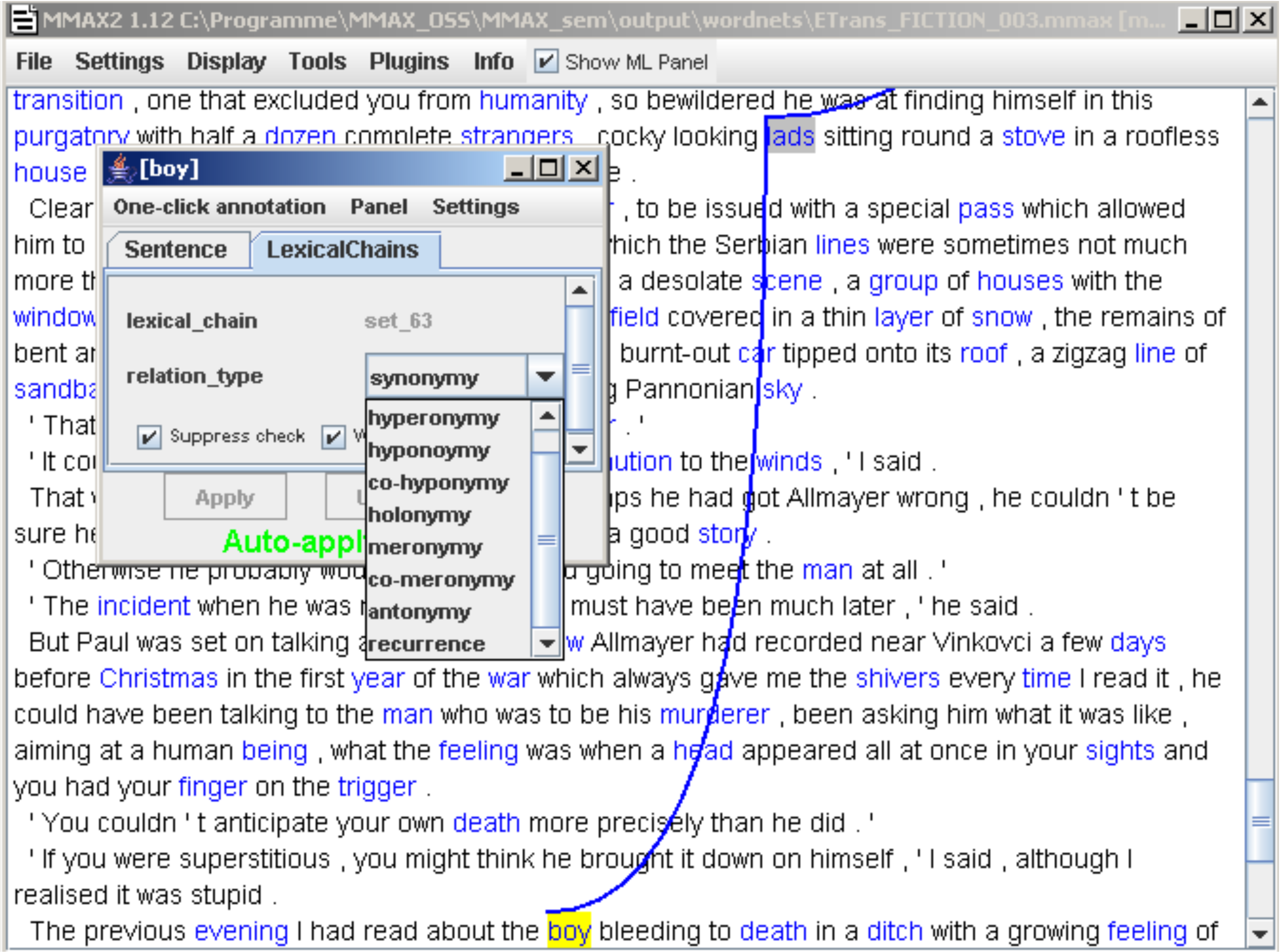
Co-Hyponymy

Holonymy

Meronymy

Co-Meronymy

Antonymy

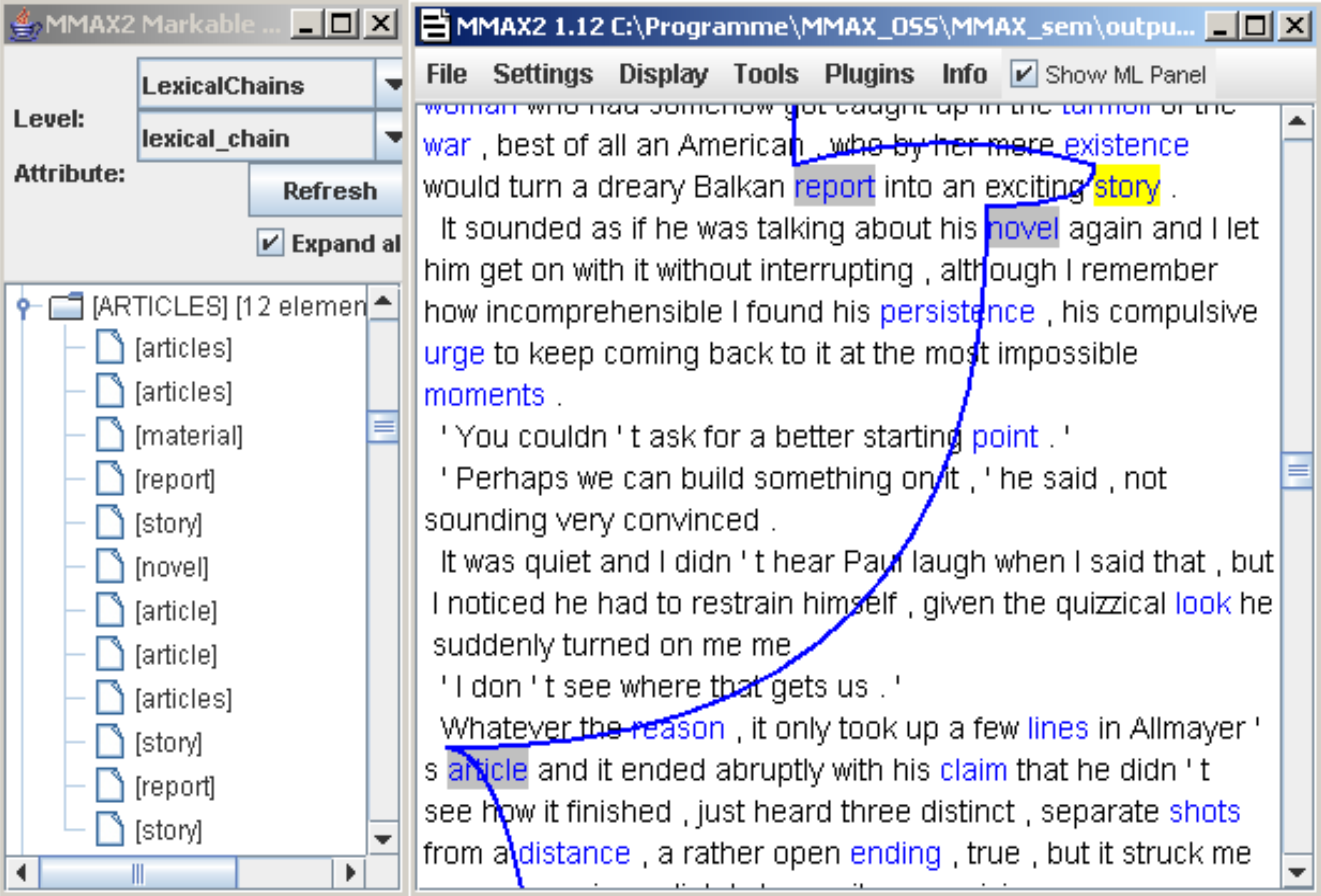


lexical_chain set_63

relation_type synonymy

hyperonymy
hyponymy
co-hyponymy
holonymy
meronymy
co-meronymy
antonymy
recurrence

transition , one that excluded you from humanity , so bewildered he was at finding himself in this
purgatory with half a dozen complete strangers , cocky looking lads sitting round a stove in a roofless
house e .
Clear , to be issued with a special pass which allowed
him to hich the Serbian lines were sometimes not much
more th a desolate scene , a group of houses with the
window field covered in a thin layer of snow , the remains of
bent ar burnt-out car tipped onto its roof , a zigzag line of
sandba g Pannonian sky .
' That . '
' It co ution to the winds , ' I said .
That ps he had got Allmayer wrong , he couldn ' t be
sure he a good story .
' Otherwise he probably wou u going to meet the man at all . '
' The incident when he was must have been much later , ' he said .
But Paul was set on talking a w 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 .

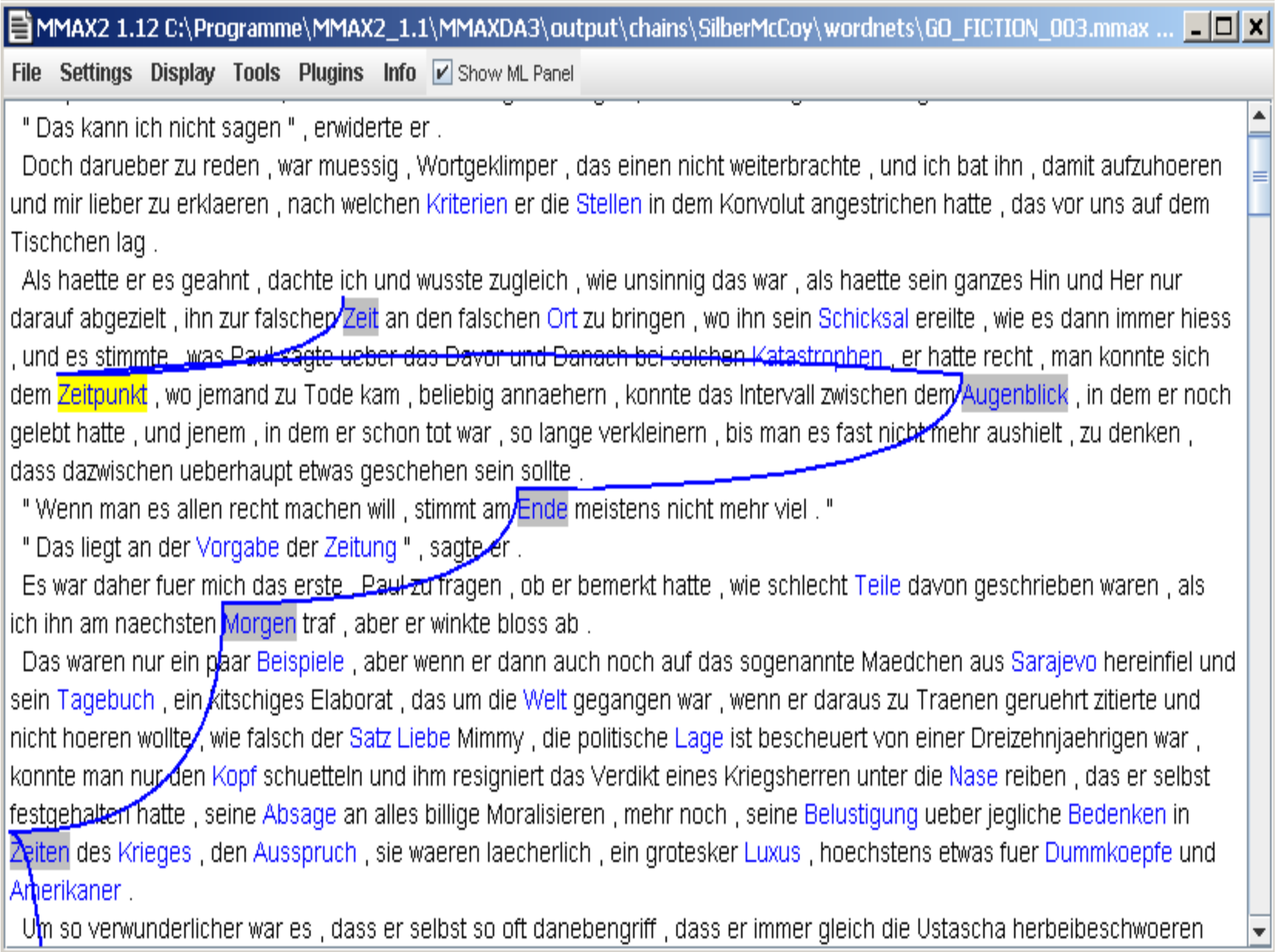
' You couldn ' t ask for a better starting point . '

' 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 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


MMAX2 1.12 C:\Programme\MMAX2_1.1\MMAXDA3\output\chains\SilberMcCoy\wordnets\GO_FICTION_003.mmax ...
File
Settings
Display
Tools
Plugins
Info
Show ML Panel

☑ Show ML Panel

" Das kann ich nicht sagen " , erwiderte er .

Doch darueber zu reden , war muessig , Wortgeklimper , das einen nicht weiterbrachte , und ich bat ihn , damit aufzuhoeren und mir lieber zu erklaeren , nach welchen Kriterien er 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 stimmte , 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 .

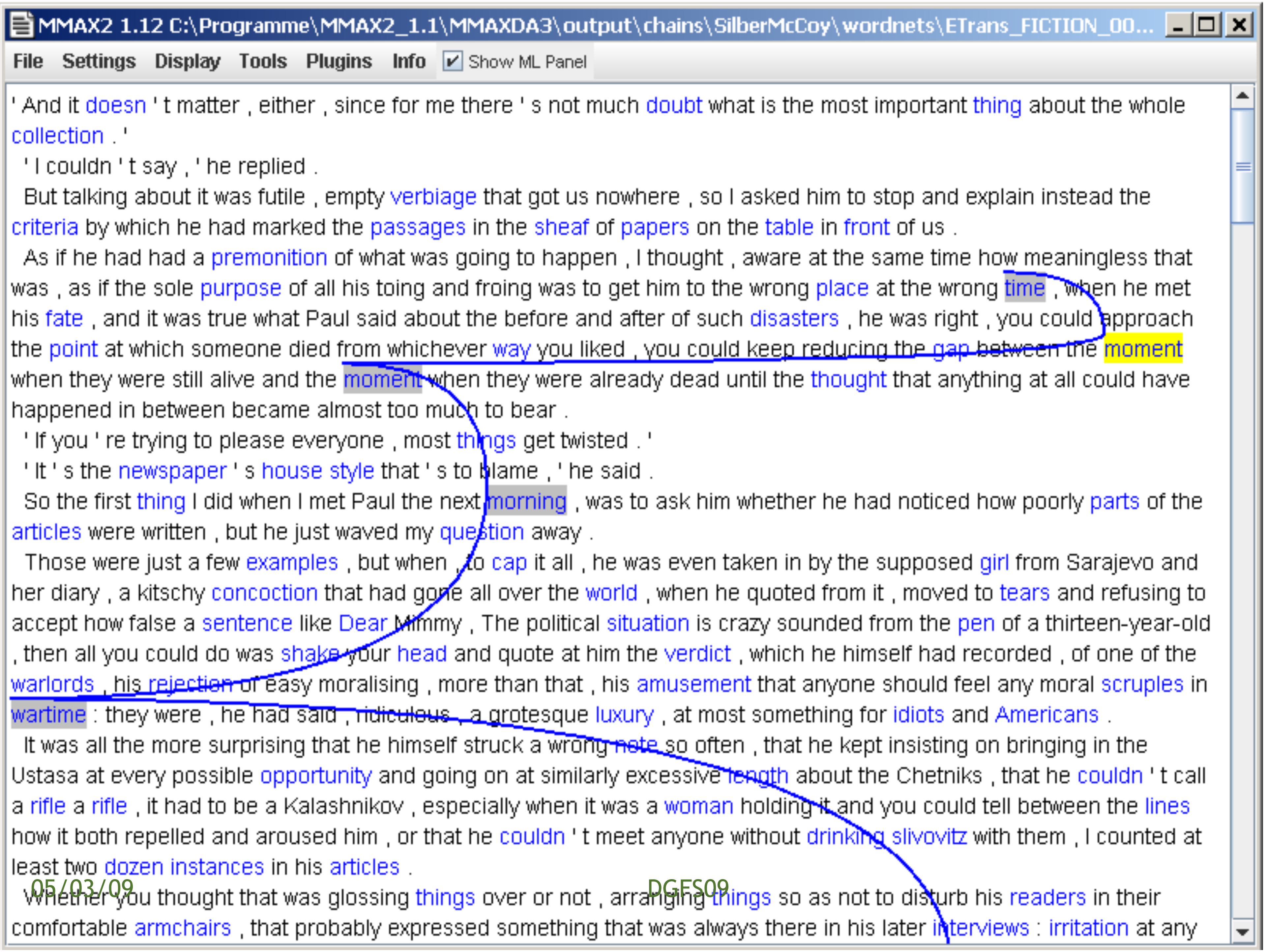
" Wenn man es allen recht machen will , stimmt am Ende meistens nicht mehr viel . "

" 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 Maedchen aus Sarajevo hereinfiel und sein Tagebuch , ein kitschiges Elaborat , das um die Welt gegangen war , wenn er daraus zu Traenen geruehrt zitierte und nicht hoeren wollte , wie falsch der Satz Liebe Mimmy , die politische Lage ist bescheuert von einer Dreizehnjaehrigen war , konnte man nur 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 Dummkoepfe und Amerikaner .

Um so verwunderlicher war es , dass er selbst so oft danebengriff , dass er immer gleich die Ustascha herbeibeschwoeren



MMAX2 1.12 C:\Programme\MMAX2_1.1\MMAXDA3\output\chains\SilberMcCoy\wordnets\ETrans_FICTION_00...
File
Settings
Display
Tools
Plugins
Info
☑ Show ML Panel

' 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 Mimmy , 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 , of 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 grotesque 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

Cramer, I. & M. Finthammer. 2008. An Evaluation Procedure for Word Net Based Lexical Chaining: Methods and Issues. In: Proc. of Global WordNet Conference 2008, Szeged, Ungarn.

Halliday, M.A.K. & R. Hasan. 1976. Cohesion in English. London: Longman

Halliday, M.A.K. & R. Hasan. 1995. Language, context, and text: aspects of language in a social-semiotic perspective. Oxford: Oxford University Press

Galley, M. & K. McKeown. 2003. Improving word sense disambiguation in lexical chaining. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco, Mexico, August.

Garoufi, K., Zesch, T. & I. Gurevych. 2008. Representational Interoperability of Linguistic and Collaborative Knowledge Bases
In: Proceedings of the KONVENS Workshop on Lexical-Semantic and Ontological Resources -- Maintenance, Representation, and Standards

Hoey, M. 1991. Patterns of lexis in text. Oxford: Oxford University Press

Neumann, S., 2008. Quantitative register analysis across languages. In: Swain, Elizabeth (ed.), Thresholds and Potentialities of Systemic Functional Linguistics: Applications to other disciplines, specialised discourses and languages other than English. Trieste: Edizioni Universitarie.

References (2)

Teich, E. & P. Fankhauser. 2004. Exploring Lexical Patterns in Text: Lexical Cohesion Analysis with Word-Net. Proceedings of the 2nd International Wordnet Conference, Brno, Czech Republic. pages 326-331

Steiner, E. 2004. *Translated Texts: Properties, Variants, Evaluations. Frankfurt: Lang.*

Zesch, T.; Müller, C. & I. Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation (LREC), electronic proceedings.

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