GECCo: Corpus to Analyse German-English Contrasts in Cohesion

Kerstin Kunz, Ekaterina Lapshinova-Koltunski*, José Manuel Martínez Martínez*, Katrin Menzel*, Erich Steiner***

Saarland University*, University of Heidelberg**

kerstin.kunz@iued.uni-heidelberg.de, e.lapshinova, j.martinez, k.menzel, e.steiner@mx.uni-saarland.de

www.gecco.uni-saarland.de

THEORETICAL BACKGROUND AND PHENOMENA IN FOCUS

Cohesion

Cohesion refers to the text-internal relationship of linguistic elements that are overtly linked via lexical and grammatical devices across sentence boundaries to be understood as a text. It is an important component of effectively organised and meaningful **discourse**, as the **message** being communicated in discourse is not just a set of clauses, but forms a **unified, coherent whole.**

Cohesive Devices

Cohesive devices are "a set of lexicogrammatical systems that have evolved specifically as resources for making it possible to transcend the boundaries of the clause"

Conjunction logico-semantic relations between propositions (e.g. addition, contrast, cause)

Reference identity between instantiated entities

Lexical cohesion similarity between entities of the same type based on sense relations (e.g. hyperny-my, part-whole relations)

Substitution/ellipsis similarity between different instantiated entities of the same type.

Examples

Conj

syntactic:

connector	subjunct	adverbial

semantic:

additive	adversative	causal	temporal	modal

Reference

personal:

head	modifier	*it/es*

demonstrative:

head	modifier	adverbial	local	temporal

comparative:

general	specific

Lex.cohesion

element type: compounds/multiwords, gen. nouns, etc.

relation: synonym, repetition, etc.

Substitution

nominal:

one	*eine/r*

verbal:

do so	*tun*

clausal:

so	*so*

Ellipsis

nominal:

EO: We have the Dee river on one side of the peninsula and the Mersey on the other ⊗.
GO: Können wir das Mikro bisschen lauter machen? Habe ich's falsche ⊗ genommen?

verbal:

EO: Better go on up while you still can ⊗.
GO: oder die Intervalle sind so gestaltet oder die Präsentationszeiten ⊗ so gestaltet...

clausal:

EO: Who says that? – My parents ⊗.
GO: An diesem Punkt, ist die Hälfte von Ihnen zu meinem Entsetzen, gescheitert. – Warum ⊗?

Cohesive relations: coreference and lexical chains

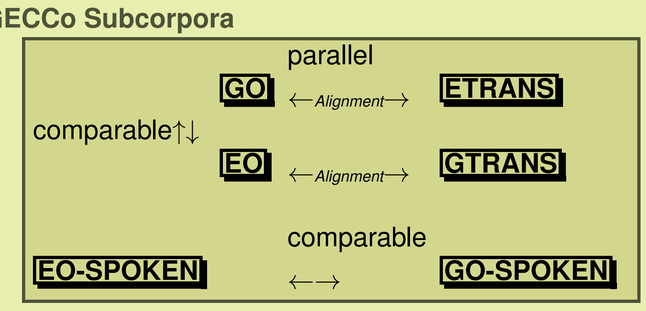
EO-POPSCI:
Three weeks later, on July 1, **Cassini** will approach **Saturn** from below the ring plane, crossing through the wide gap between the F and G rings. To slow **the spacecraft** enough to allow **it** to go into orbit, **it** will fire **its engine** for 97 minutes in the opposite direction of **its** travel. During **the engine** burn, **Cassini** will make **its** closest approach to **Saturn**, coming within 18,000 kilometers of **the gas giant**.

GO-POPSCI:
Ich habe bisher recht pauschal von **Umwelt**, äußeren Realitäten oder von Beziehungssystemen gesprochen. Wie verhält es sich nun, wenn wir nicht nur allgemein von **Umwelt** sprechen, sondern davon ausgehen, dass **diese Umwelt** ein **Du**, ein Partner ist, der mir nicht passiv gegenübersteht, mir nicht einfach Widerstand entgegensetzt oder sich von mir formen lässt, sondern in der Begegnung und Beziehung selbst ebenso aktiv ist wie ich. Inwiefern ist **dieses Du** eine von mir geschaffene **Umwelt**, das **"Werk"** von mir? Inwiefern bin ich die von **diesem Du** geschaffene **Umwelt**, **sein "Werk"**?

cf. (Halliday&Hasan, 1976)

CORPUS ARCHITECTURE

GECCo Subcorpora



Corpus Basic Information (January 2015)

	Nr-of-texts	Nr-of-sent	Nr-of-tokens
EO	110	13.727	286.221
GO	121	15.736	288.490
GTRANS	110	13.970	284.561
ETRANS	121	15.180	322.223
EO-SPOKEN	74	8.380	292.821
GO-SPOKEN	68	11.949	219.070
EO-TOTAL	305	37.287	901.265
GO-TOTAL	299	41.655	792.121
TOTAL	604	78.942	1.693.386

Available Registers (in terms of SFL)

Sprache	Register
written (imported from CroCo*), texts from 1992 – 2006	
EO	FICTION, ESSAY
GO	INSTR, POPSCI
ETRANS	SHARE, SPEECH
GTRANS	TOU, WEB
spoken (collected**), texts from 2008 – 2012	
EO	INTERVIEW, ACADEMIC, *FORUM*
GO	*TALKSHOW, MEDCONSULT, SERMON*

* (Hansen-Schirra et al., 2012), ** (Lapshinova et al., 2012)

Corpus annotation

1) Word: ⇒ *token, lemma, part-of-speech*
2) Chunk: ⇒ *phrases, clauses and sentences*
cohesive devices and chains
3) Text: ⇒ *text and register borderlines*
4) Extra-linguistic information: ⇒ *meta*

Annotation information

	coreference		conj	
subcorpus	**EO**	**GO**	**EO**	**GO**
ACADEMIC	2.338	2.324	2.317	2.913
INTERVIEW	2.544	2.701	2.696	4.047
ESSAY	1.257	1.506	0.834	1.295
FICTION	3.462	2.527	1.596	1.822
INSTR	0.957	1.264	1.021	1.134
POPSCI	1.490	1.800	1.257	1.844
SHARE	1.268	0.453	0.740	1.047
SPEECH	1.488	1.527	0.808	1.227
TOU	1.087	1.685	0.710	0.748
WEB	1.233	2.128	0.690	0.786
TOTAL	17.124	17.915	12.669	16.863

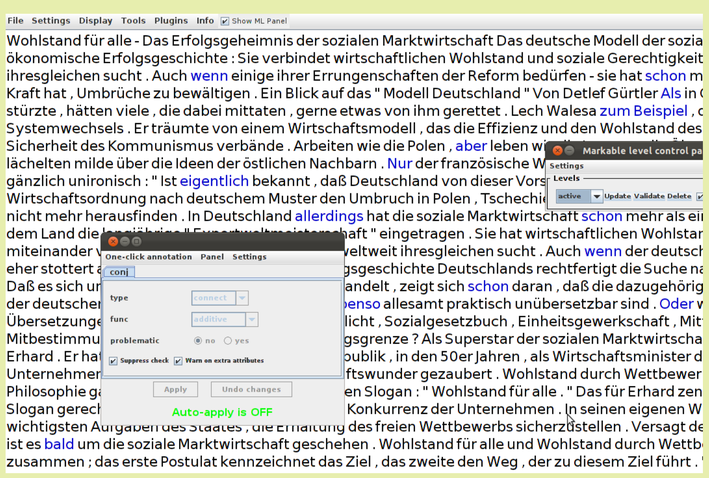
- Releases: GECCO2013, GECCO2014
- Formats: Standoff xml, CWB, MMAX2
- Available for querying via CLARIN-D repository: https://fedora.clarin-d.uni-saarland.de/cqpweb/

ANNOTATION OF COHESION

Annotation procedures (Lapshinova & Kunz, 2014a,b)

	steps	tools
1	automatic pre-annotation	CQP queries+perl scripts
2	manual correction	MMAX2

Example MMAX2 visualisation (Müller & Strube, 2006)



Annotated Structures in XML

<conj func="additive" type="connector"> und < /conj>
<conj func="adversative" type="adverbial"> jedoch < /conj>
<conj func="causal" type="subjunct"> aufgrund dessen < /conj>

Example CQPWeb

Application

- corpus-based analysis of contrasts between English and German, between different registers
- areas of application:
  - linguistics (including comp.linguistics)
  - language teaching
  - translation studies

Acknowledgements

former GECCo team members:
Dr. Marilisa Amoia, Dr. Stefania Degaetano-Ortlieb
student assistants:
Marie-Pauline Krielke, Nadine Braun, Sarah Justinger
funding institution:
Deutsche Forschungsgemeinschaft

References

- Halliday, M.A.K, R. Hasan (1976). Cohesion in English. London, New York: Longman.
- Hansen-Schirra, S., S. Neumann, E. Steiner (2012). Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English - German. Series Text, Translation, Computational Processing. Berlin, New York: Mouton de Gruyter.
- Lapshinova, E., K. Kunz, M. Amoia (2012). Compiling a Multilingual Spoken Corpus. In *Proceedings of GSCP-2012*, BH, Brazil.
- Lapshinova, E., K. Kunz (2014a). Annotating Cohesion for Multilingual Analysis. In Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Reykjavik, May 26, 2014.
- Lapshinova-Koltunski, E. and K. Kunz (2014b). Conjunctions across Languages, Registers and Modes: semi-automatic extraction and annotation. In Diaz-Negrillo, A. and J. Diaz-Perez Francisco (eds). Specialisation and Variation in Language Corpora. Peter Lang.
- Müller, C., M. Strube (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In: S. Braun, K. Kohn, J. Mukherjee (eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang (English Corpus Linguistics, Vol.3).