# German-English Contrasts in Cohesion - Towards an empirically-based comparison (GECCo)

Project report
Project number STE 840/6-1 / 6-2 and KU 3129/1-2

# 1 General data

## 1.1 DFGReference number

STE 840/6-1 und 6-2 und KU 3129/1-2

## 1.2 Applicants

1) Univ.-Prof. Dr. Erich Steiner

*Institut/Lehrstuhl*:
Professur Übersetzungswissenschaft Englisch,
Fachrichtung Sprachwissenschaft und Sprachtechnologie
Universität des Saarlandes
Campus, Geb. A.2.2, Raum 1.05
66123 Saarbrücken
Tel. 0681 302-4482
Fax 0681 302-64375
E-Mail: E.Steiner@mx.uni-saarland.de

2) Univ.-Prof. Dr. Kerstin Kunz

*Institut/Lehrstuhl*:
Professur Übersetzungswissenschaft: Englisch
Institut für Übersetzen und Dolmetschen
Universität Heidelberg
Raum: SUED 007a
Plöck 57a, 69117 Heidelberg
Tel. 06221-547227
Fax: 06221-547559
E-Mail: kerstin.kunz@iued.uni-heidelberg.de

Das Projekt wurde 2013 für die Antragsteller, damals noch beide an der Universität des Saarlandes, wie oben genannt für 3 Jahre als Fortsetzungsprojekt bewilligt. Mit Antritt einer Professur durch Kerstin Kunz an der Universität Heidelberg wurde das Projekt ab Anfang 2016 zu gleichen Teilen aufgeteilt.

## 1.3 Theme

English - German contrasts in cohesion: empirically-based comparison of the written-spoken continuum

## 1.4 Funding period

Period under review: July 2013 - January 2017

Funding periods:
First phase: April 2011 - March 2013
Second phase: Mai 2013 - January 2017

## 1.5 Selected Publications

Fuller lists of publicatins are referred to below

a)

Kunz, Kerstin, Ekaterina Lapshinova-Koltunski, José Manuel Martínez-Martínez, Katrin Menzel & Erich Steiner (forthcoming). Shallow features as indicators of English-German contrasts in lexical cohesion. In: Languages in Contrast.18:2.

Kunz, Kerstin, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel & Erich Steiner (forthcoming May 2017). GECCo - an empirically-based comparison of English-German cohesion. In: De Sutter, G., Delaere, I. & Lefer, M.-A. (eds.). Empirical Translation Studies. New Methodological and Theoretical Traditions. Trends in Linguistics. Studies and Monographs [TiLSM] 300. Mouton de Gruyter.

Kunz, Kerstin, Ekaterina Lapshinova-Koltunski & José Manuel Martínez-Martínez (2016). Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In: Proceedings of CORBON at NAACL-HLT2016, San Diego.

Martínez-Martínez, José Manuel, Ekaterina, Lapshinova-Koltunski & Kerstin Kunz (2016). Annotation of Lexical Cohesion in English and German: Automatic and Manual Procedures. In: Proceedings of the Conference on Natural Language Processing (Konferenz zur Verarbeitung natürlicher Sprache) - KONVENS-2016, September, Bochum, Germany.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski (2015). Cross-linguistic analysis of discourse variation across registers. In: K. Ajmer & H. Hasselgård (eds.). Special Issue of the Nordic Journal of English Studies.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski (2014). Cohesive conjunctions in English and German: Systemic contrasts and textual differences. In: Vandelanotte, L., K. Davidse, C. Gentens & D. Kimps (eds.). Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora. Amsterdam/New York: Rodopi. 229-262.

Lapshinova-Koltunski, Ekaterina & Kerstin Kunz (2014). Annotating Cohesion for Multilingual Analysis. In: Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation in conjunction with LREC2014 the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Island, Mai 2014.

Menzel, Katrin (2014). Ellipsen als Stil- und Kohäsionsmittel in deutschen und englischen politischen Reden. In: Germanistische Mitteilungen. Zeitschrift für Deutsche Sprache, Literatur und Kultur, 40.1: Deutsch kontrastiv. Brüssel. 31-50.

Steiner, Erich (2015). Contrastive studies of cohesion and their impact on our knowledge of translation. In: Zhang, Meifang and Munday, Jeremy (eds.). Discourse Analysis and Translation. Special issue of TARGET. International Journal of Translation Studies. 27:3. pp. 351-369. Amsterdam: John Benjamins.

b) /

c) /

# 2 Project report

This report summarizes work performed and results achieved in the second phase of the GECCo project. See `http://www.gecco.uni-saarland.de/GECCo/index.html` for more information.

See `http://fedora.clarin-d.uni-saarland.de/gecco/` for a documentation of the GECCo corpus (in cooporation with CLARIN-D).

The corpus and its documentation can be accessed online for queries with CQPweb `http://fedora.clarin-d.uni-saarland.de/gecco/`/`http://corpora.clarin-d.uni-saarland.de/cqpweb/`. Unrestricted access to the corpus texts is not possible due to property-right restrictions. Querying the corpus, however, with and without GECCo's annotations is possible and open to members of the research community and students.

See `http://www.gecco.uni-saarland.de/GECCo/files/GECCo_Korpus_Overview.pdf` for an overview of the GECCO corpus architecture.

## 2.1 Research questions and goals

The second phase of the GECCo-project aimed at the following research goals (see project proposal):

GOAL 1: Extension of empirical base of the GECCo corpus resource by: (1) additional registers of spoken language and borderline registers, (2) multilayer lexicogrammatical annotation of spoken registers, (3) automatic and semi-automatic strategies for the identification of substitution, reference and conjunction, (4) corpus analyses of lexical cohesion and ellipsis.

GOAL 2: Modelling of cohesion in a contrastive perspective English-German: (1) Integrate contrasts between spoken registers, (2) Add systematic studies on ellipsis and lexical cohesion for the full range of registers.

GOAL 3: Generalization of English-German contrasts in cohesion (1) Generalize contrasts in terms of difference in strength (quantitative) and in typical lexicogrammatical expressions (qualitative) of the written-spoken dimension. (2) Explain registerial contrasts within each language in terms of the different requirements of context-dependent and context-independent processing. (3) Explain contrasts across the two languages in terms of a (quantitative and qualitative) weakening of the spoken-written difference in English relative to German.

## 2.2 Work packages

The overall work of the project consisted of 6 work packages (plus this final report): Work Package 1: Expansion of corpus resource (related mainly to goal 1 above) WP1.1 Compilation of spoken registers most typical for spoken language, including multimedia as well as new registers at the borderline between spoken and written discourse. WP1.2 Annotation of new registers with features of spoken language. WP1.3 Lexicogrammatical annotation compatible with existing corpora. Testing of new automatic tools for some linguistic levels. WP1.4 Encoding for corpus exploitation.

The spoken part of the corpus, comprising two registers at the end of the first project phase (ACADEMIC: academic lectures and INTERVIEW: interviews) was extended with the following new spoken registers: internet forum excerpts (FORUM), medical consultation interactions (MEDCONSULT), sermons (SERMON), and TV talkshow excerpts (TALKSHOW).

Systematic sampling of spoken data and its annotation was more difficult than foreseen, but during 2015 reached a stage where it was sufficient for at least most of the empirical questions asked by the project. Automatic and semi-automatic strategies for the identification of substitution, co-reference and conjunction in the spoken corpora were developed and implemented.

The overall corpus and the associated query facilities are available in different corpus releases, documented in `http://fedora.clarin-d.uni-saarland.de/gecco/`. These corpus versions (GECCO-UPOS, GECCOCOH and GECCOCHAIN) are encoded with annotations on different linguistic levels as they are dedicated to different contrastive analysis tasks (see below).

Work Package 2: Systemic analysis of lexical cohesion and cohesive ellipses as major reflections of the written-spoken continuum (related to goal 2 above) WP2.1 Conceptualization. WP2.2 Systemic contrasts. WP2.3 Hypotheses. WP2.4 Operationalization of research parameters for corpus analysis, particularly in spoken language registers.

This was successfully carried out.

Work Package 3: Corpus annotation and exploitation on the level of cohesion WP3.1 Development of strategies/tools for the automatic annotation/extraction of cohesive devices and cohesive relations in spoken registers WP3.2 Elaboration of guidelines for the manual post correction of the automatic annotations of lexical cohesion and cohesive ellipses (devices and relations) for all registers WP3.3 Manual correction of lexical cohesion and cohesive ellipses annotations WP3.4 Extraction of cohesive devices and cohesive relations from corpus data

This work has been carried out, though for lexical chains with only a subset of representative registers. The main reason was that the annotation of lexical chains and sense relationships between chain elements requires substantially more human intervention and interpretation than originally foreseen. The same applies to many subtypes of ellipsis, but as ellipses are not extremely frequent cohesive devices, all corpus registers (including the translations of the written corpus texts) as well as the spoken corpus registers (apart from three that were added to the corpus at a later stage) were annotated for cohesive ellipses and their textual antecedents.

In order to explore different aspects of cohesion on different levels of granularity, three separate corpus versions were developed. They also permit contrastive comparisons in terms of other linguistic research questions. GECCO-UPOS is the biggest version of the corpus, comprising 14 registers altogether (8 written and 6 spoken registers). It contains annotation that can be obtained with fully automatic procedures on the level of token, lemma and part-of-speech (POS) only. Additionally, we add universal part-of-speech categories which are comparable for the both languages under analysis. These comparable categories were achieved here at the cost of granularity, which is sufficient for the analysis of such shallow features as type-token-ratio (TTR), lexical density (LD), most frequent words and part-of-speech distributions, which are among the most frequently used variables characterising linguistic variation in corpora. GECCO-UPOS was used in our research for contrastive analyses of global lexical features as indicators for lexical cohesion. The corpus version additionally permits an overall comparison of English and German in terms of a broad range of shallow features.

For the contrastive analysis of more complex structures on deeper linguistic levels smaller subcorpora with further annotations are provided. This kind of analysis involves manual annotation procedures which are costly and time-consuming.

The corpus version GECCOCOH contains ten registers: all written registers and two spoken ones. Apart from the shallow features mentioned above, GECCOCOH is encoded with information on syntactic features. Most importantly for our research questions, GECCOCOH is also enriched with information on cohesive devices, signalling different types of cohesion (conjunction, coreference, substitution and ellipsis). For this purpose, annotations on different levels of granularity have been added (e.g. devices of coreference

serving as nominal head or modifier, syntactic and semantic features of conjunctions). In addition, this version provides annotations of coreference chains and elliptical relations.

As mentioned above, a small subset of the corpus (GECCOCHAIN) contains annotations of lexical chains. This includes links between the elements in a lexical chain as well as information on the type of sense relation between chain elements, i.e. relations of synonymy, antonymy, hyponymy, etc.

Work Package 4: Interpretation of findings WP4.1 Statistical evaluation of frequencies, clustering WP4.2 Elaboration of a comprehensive model of contrasts in English and German with a differentiated view on written and spoken registers WP4.3 Elaboration of a model for cohesion in spoken and written language with a view to underlying contexts of communication

We generalized cohesive contrasts in terms of degree, strength, semantic type and variation of cohesive devices and chains with particular consideration of the written-spoken dimension. We attempted explanations of registerial contrasts within each language in terms of the different requirements of context-dependent and context-independent processing. And we looked for traces of a quantitative and qualitative weakening of the spoken-written difference in English relative to German, although here the results did not show an unambiguous confirmation of corresponding assumptions (for example in Leech et al. 2009; Mair 2006; Leisi and Mair 2008). Statistical evaluation was carried out with different combinations of unsupervised and supervised methods to obtain different perspectives on interpretations of our findings.

This has been carried out as planned, although for only a subset of the overall corpus in the areas of lexical cohesion.

Work Package 5: Implications for Research and Teaching responsibility WP5.1 Teaching: Integrating new insights on general contrasts in cohesion English-German into foreign language teaching, and into translator and interpreter training. WP5.2 Research: Integrating new insights for (1) Text linguistics: Contrastive text based-grammars and 2) NLP: Coreference resolution, automatic identification of other strategies of cohesion, machine translation.

WP 5.1. was achieved as planned, leading to separate publications (Steiner 2015, Menzel 2016). WP 5.2. has led to a series of publications on the borderline of linguistics, linguistic engineering and machine translation. (Nedoluzhko and Lapshinova-Koltunski E. (2016); Lapshinova-Koltunski, Kunz and Nedoluzhko (2016); Vela and Lapshinova-Koltunski (2015); Lapshinova-Koltunski, E. and M. Vela (2015)).

Work Package 6: Dissemination of the projects findings took the form of a significant number of project publications and deliverables, as well as the corpus as a resource for automatic querying (see presentations and events under project homepage), a high number of internal project deliverables, partly open to the research community (see deliverables under GECCo's homepage `http://www.gecco.uni-saarland.de/GECCo/deliverables.html` and the PhD-thesis and 2 Habilitation-Theses produced by project members during 2013-2016 (Kunz, Lapshinova-Koltunski, Menzel).

## 2.3 Project outcome

The major deliverable from WP 1 is the GECCo-corpus in various releases as described above. The GECCo-corpus and its documentation can be accessed for queries under `http://fedora.clarin-d.uni-saarland.de/gecco/` and `http://corpora.clarin-d.uni-saarland.de/cqpweb/` and is thus integrated in CLARIN-D. Unrestricted access to the corpus texts is not possible due to well-known property-right restrictions concerning source texts. Querying the corpus, however, with and without GECCo's annotations is possible and has been used by members of the research community and students already. Relevant publications include: Lapshinova-Koltunski and Kunz 2013, 2014; Menzel 2017. The

corpus architecture, shown in Figure 1 below, allows quantitative testing of frequencies of cohesive phenomena (lexical cohesion, conjunction, ellipsis, substitution, reference as devices and/ or chains), generalized properties of these cohesive phenonema (degree, strength, type, variation), and all of these variables dependent on 2 languages (E-G), 14 registers, and two modes (spoken-written).

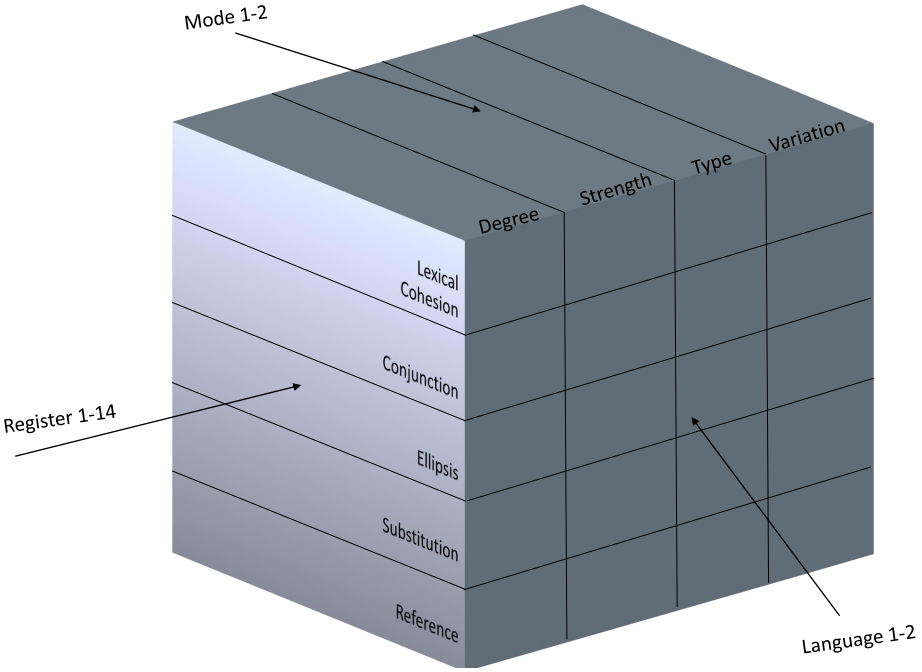


Figure 1: GECCo-Corpus: independent variables, types of cohesive devices, generalizations of cohesive contrasts

WP 2 provided the theoretical input to WPs 3 and 4. Results of this more theoretical work showed that ellipsis is a more heterogeneous notion and more difficult to delimit from other fragments than is often thought (Menzel 2014; 2016 b). As expected, the written-spoken dimension as reflected in cohesive ellipses is more gradual than assumptions of a binary classification assume. Lexical cohesion even on the theoretical level confronts us with the distinction between potential (systemic) and actual (instantial) meaning, as well as with ambiguity in that context. Results as in Kunz (et al. 2016; forthcoming) show that the associated problems can be addressed sufficiently to achieve first satisfactory results.

WP 3 produced the annotated versions of the GECCo-corpus even though for lexical chains with only a subset of representative registers. Fully automatic procedures for the creation of annotations satisfying the linguistically-grounded criteria arising out of WP 2 are not available in sufficient quality. We therefore developed a mixture of automatic pre-coding with human intervention and post-editing to develop subcorpora of sufficient quality (Kunz et al. forthcoming; Kunz et al. 2016).

In contrast to the annotation methods for co-reference, substitution and conjunctive relations, the annotations of lexical cohesion and ellipses result from a largely manual annotation process as this proved more accurate than automatic or semi-automatic methods. The annotations of cohesive devices, antecedents and chains that were created and the extracted patterns can serve as a basis for similar annotations in a larger corpus in the future.

We also profited from collaboration here with the Prague Discourse Tree Bank, leading to interoperable

annotations across theories and corpora (Nedoluzhko and Lapshinova-Koltunski 2016 a,b).

WP 4 produced statistically much refined evaluations of our empirical results (Kunz et al. forthcoming; Kunz et al. forthcoming May 2017; Kunz et al. 2016) from both phases of GECCo's lifetime. Our generalization of cohesive contrasts in terms of degree, strength, semantic type and variation of cohesive devices and chains with particular consideration of the written-spoken dimension is new in the area of contrastive linguistics, at least to our knowledge. Such a generalization appears necessary, though, as one type of tertium comparationis for cross-linguistic comparison. Figure 2 below gives and overview on our findings for contrasts between the languages E-G and the modes spoken-written.

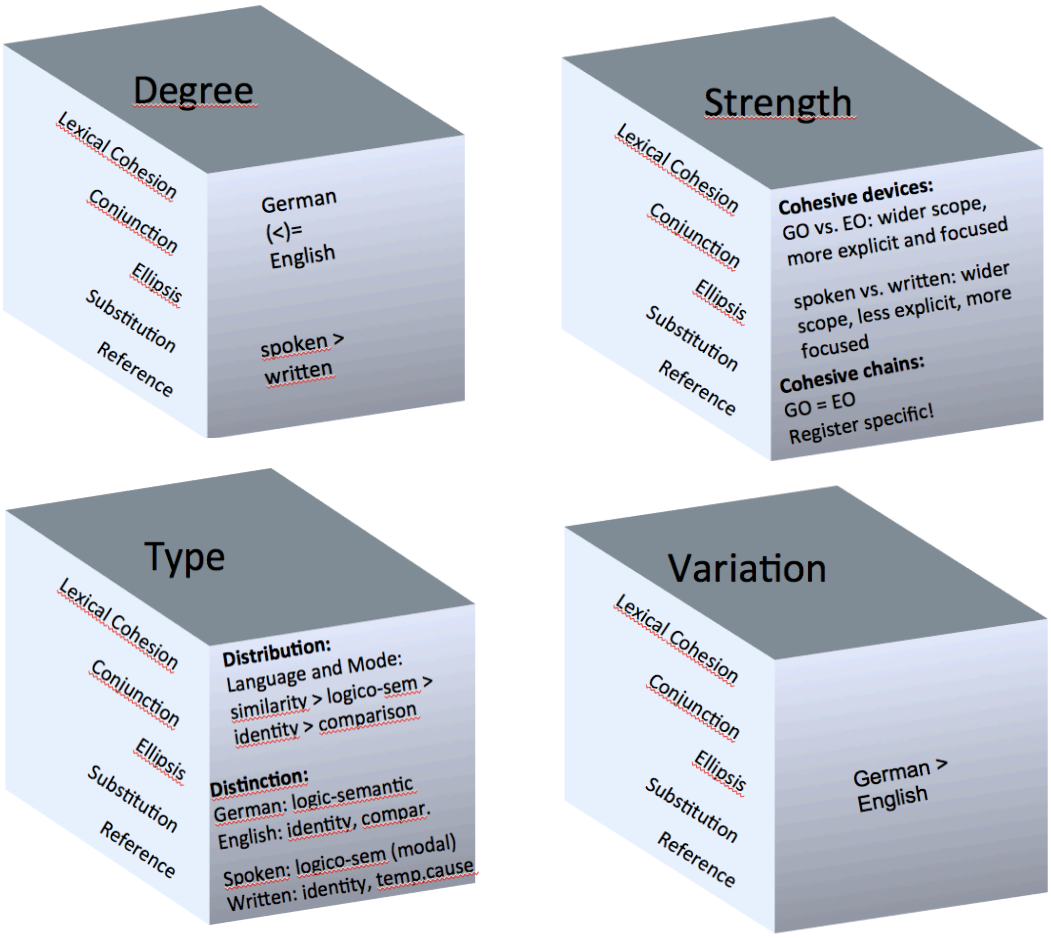


Figure 2: Properties of cohesion along the language and mode distinctions

In summary our analyses show that more contrasts between the two languages exist on lower linguistic levels of classification, such as structural forms and syntactic function, and the degree of semantic specification of cohesive devices. The contrasts between our English and German subcorpora partially stem from systemic differences in the availability of cohesive devices and from differing cultural preferences in terms of explicitness and focus. Fewer language contrasts surface if we consider the features of chains, i.e. the cohesive relations triggered by cohesive devices. More pronounced contrasts, particularly in terms of cohe-

sive relations expressed in chains can be observed when comparing spoken and written registers, from an intralingual as well as interlingual perspective. We note a parallel preference across languages and modes for distributions of general types of meaning relation. Our quantitative analyses, however, reveal significant contrasts with respect to the specific type of meaning relation expressed, the distance between elements in cohesive chains, the number of cohesive chains, chain length, etc. On the one hand, these contrasts are a reflection of register variation, e.g. variation in the function of communication in typical contexts of situation. On the other hand, the contrasts can be attributed to differences in the production mode of communication used in particular contexts of situation. The contrasts in features of cohesive devices and cohesive chains were summarized and interpreted in terms of contrasts on more general dimensions of cohesion such as degree of cohesion, strength of the cohesive relation, type of meaning relation and degree of cohesive variation, as shown in Figure 2.

A comprehensive list of deliverables can be found at: `http://www.gecco.uni-saarland.de/GECCo/deliverables.html`

All project-related presentations and publications are also listed under: `http://www.gecco.uni-saarland.de/GECCo/publications.html` and `www.gecco.uni-saarland.de/GECCo/presentations.html`

Together with annotation guidelines from WP 3, our statistical evaluation techniques constitute important models for a working pipeline in corpus-based empirical work. Results here included wide-ranging overviews on contrasts in cohesion English-German, as well as focused papers on lexical cohesion and on ellipses with particular reference to their function across the spoken-written modes (e.g. Menzel 2017; Kunz et al. forthcoming, Kunz et al. forthcoming May 2017). We hope that these results, together with those from the first project phase on co-reference, substitution and conjunctive relations, will form the basis of a monograph-type account of contrastive cohesion English-German after the lifetime of the project.

As deliverables for WP 5, project members produced publications with detailed outlines of implications of GECCo's findings for language teaching on the one hand, and for modelling and teaching of translation on the other (Menzel 2016; Steiner 2015). Beyond these two areas of application, project members jointly with collaborating co-authors produced several publications integrating findings from GECCo into linguistic engineering and machine translation (Nedoluzhko and Lapshinova-Koltunski E. (2016); Lapshinova-Koltunski, Kunz and Nedoluzhko (2016); Vela and Lapshinova-Koltunski (2015); Lapshinova-Koltunski and Vela (2015)).

WP6 dissemination took the form of a significant number of conference presentations, organization of colloquia and workshops both nationally and internationally (see presentations and events under project homepage), a high number of internal project deliverables, partly open to the research community (see deliverables under GECCo's homepage) and of course the PhD-thesis and 2 Habilitation-Theses produced by project members during 2013-2017 (Kunz, Lapshinova-Koltunski, Menzel).

In a general assessment, we have outlined an overview on contrasts in cohesion between English and German, based on a comparison of the language specific systems, but crucially as well on a corpus-based study of a representative corpus. A summary of our results from GECCo's two phases (2011 - 2013 and 2013 - 2016) should form at least the backbone of a companion volume to, say, König an Gast's 2012 "Understanding English-German Contrasts". Where we hope to have delivered something methodologically new is in the implementation of an empirical corpus-based architecture for investigating cohesion across languages (and we have added at least work on Czech to our bilingual English-German focus). The corpus, its architecture and the empirical validation of frequencies represent at least state-of-the-art methodologies, even if compared to work such as Biber 2014, Biber and Egbert 2016, or in Gast 2015. Another important point of reference is, of course, work in and around DFG-SFB 1102, Information Density and Linguistic Encoding, as in Teich et al. 2015a,b. The latter are inquiring into partly similar phenomena as we are, however so far not in a contrastive perspective and much more heavily relying on data mining approaches than we are. GECCo is using some methodologies shared with data mining, but in the evaluation of findings only.

Comprehensive lists of presentations by project members (and where relevant their co-operating partners) can be found on the GECCo-website.

## 2.4 Project members

- Prof. Dr. Erich Steiner (nicht aus Projektmitteln)
- Prof. Dr. Kerstin Kunz (nicht aus Projektmitteln)
- PD Dr. Ekaterina Lapshinova-Koltunski
- Dr. Katrin Menzel (nicht aus Projektmitteln)
- José Manuel Martínez-Martínez
- Dr. Stefania Degaetano-Ortlieb

*Student assistants*

- Stefanie Bauernfeind
- Nadine Braun
- Tabea Buth
- Natalija Grbavac
- Sarah Justinger
- Pauline Krielke

## 2.5 Project-related theses

**Completed PhD and postdoctoral theses:**

- Kunz, Kerstin (2015). Cohesion in English and German. A corpus-based approach to language contrast, register variation and translation. Habilitationsschrift. Universität des Saarlandes.
- Lapshinova-Koltunski, Ekaterina (2016). Inter- and Intralingual Variation in a Multilingual Context: Dimensions, Interactions, and their Implications. Habilitationsschrift. Universität des Saarlandes.
- Menzel, Katrin (2017). Understanding English-German contrasts - a corpus-based comparative analysis of ellipses as cohesive devices, PhD Dissertation. Saarbrücken: Saarländische Universitäts- und Landesbibliothek

**Master theses (selection):**

See all students' theses (BA & MA) at:

`http://www.gecco.uni-saarland.de/GECCo/publications.html`

- Bauernfeind, Stefanie Anna Franziska (2014). Kohäsive Konjunktionen in populär-wissenschaftlichen Texten: eine korpusbasierte kontrastive Analyse. MA Thesis.

- Braun, Nadine (2015). A Corpus-based Study of Conjunctive Relations in English and German. MA Thesis. Braun, Nadine. 2015. A Corpus-based Study of Conjunctive Relations in English and German. MA Thesis. Busse, Vanessa, 2016. Referenz im politischen Diskurs: eine korpusbasierte Analyse des Deutschen und Englischen. MA Thesis.
- Castañares, Siobhan Briz (2016) Systemische vs. instantiierte Lexis im deutsch-englischen Vergleich anhand verschiedener Registerpaare. MA Thesis.
- Eberhard, Tobias (2017). A contrastive corpus analysis of nominal compounds in English and German. MA Thesis.
- Krielke, Marie-Pauline Anna (2015) Lexikalische Kohäsion und Koreferenz im Deutschen und Englischen. Eine corpusbasierte Studie populärwissenschaftlicher Texte. MA Thesis.
- Küntzer, Nicole (2015). Methods for translating focused elements from English into German using popular-scientific texts as an example. MA Thesis.
- Redeker, Katharina (2017). Die Ellipse als Übersetzungsproblem - eine syntaktische Analyse elliptischer Konstruktionen im Englischen und im Deutschen. MA Thesis.

# References

**See the following website for all references:**

`http://www.gecco.uni-saarland.de/GECCo/publications.html`

Biber, D. & Egbert, J. (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis. Journal of English Linguistics. Sage. 95-137.

Biber, Douglas (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. Languages in Contrast 14.1. (7-34).

Gast, V. (2015). On the use of translation corpora in contrastive linguistics: A case study of impersonalization in English and German. In: Languages in Contrast: 4-33. (Special issue on 'Contrasting contrastive approaches', ed. by B. Defrancq.).

König, E. & V. Gast (2012). Understanding English-German Contrasts. Grundlagen der Anglistik und Amerikanistik. Berlin: Erich Schmidt Verlag. [3rd, extended edition].

Kunz, Kerstin (2015). Cohesion in English and German. A corpus-based approach to language contrast, register variation and translation. Habilitationsschrift. Universität des Saarlandes.

Kunz, K., E. Lapshinova-Koltunski & J.M. Martínez-Martínez (2016). Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In: Proceedings of CORBON at NAACL-HLT2016, San Diego.

Kunz, K., S. Degaetano-Ortlieb, E. Lapshinova-Koltunski, K. Menzel & E. Steiner. (forthcoming Mai 2017). GECCo - an empirically-based comparison of English-German cohesion. In: De Sutter, G., I. Delaere & M.-A. Lefer (eds.). New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies. TILSM series. Mouton de Gruyter.

Kunz, K., E. Lapshinova-Koltunski, J.M. Martínez-Martínez, K. Menzel & E. Steiner (forthcoming). Shallow features as indicators of English-German contrasts in lexical cohesion. In: Languages in Contrast. 18:2

Lapshinova-Koltunski, E. & K. Kunz (2013). Detecting Cohesion: semi-automatic annotation procedures. In: Proceedings of Corpus Linguistics-2013. Lancaster, UK.

Lapshinova-Koltunski, E. & K. Kunz (2014). Annotating Cohesion for Multilingual Analysis. In: Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation in conjunction with LREC2014 the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 2014.

Lapshinova-Koltunski, Ekaterina (2016). Inter- and Intralingual Variation in a Multilingual Context: Dimensions, Interactions, and their Implications. Habilitationsschrift. Universität des Saarlandes.

Lapshinova-Koltunski, E. & M. Vela (2015). Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach. In: Proceedings of EMNLP 2015 Workshop on Discourse in Machine Translation. Lisbon, Portugal, 122-131.

Lapshinova-Koltunski, E., K. Kunz & A. Nedoluzhko (2016). From Interoperable Annotations towards Interoperable Resources: A Multilingual Approach to the Analysis of Discourse. In: Proceedings of LREC-2016. Portoroz, Slovenia.

Leech, G., Marianne Hundt, C. Mair & N. Smith (2009). Change in contemporary English. A grammatical study. Cambridge: CUP.

Leisi, E. & Mair, C. (2008). Das heutige Englisch: Wesenszüge und Probleme. 9th edition Heidelberg: Universitätsverlag Winter.

Mair, C. (2006). Twentieth-century English. History, variation and standardization. Cambridge: CUP.

Menzel, Katrin (2014). Ellipsen als Stil- und Kohäsionsmittel in deutschen und englischen politischen Reden. In: Leuschner, Torsten / Koliopoulou, Maria (ed.): Germanistische Mitteilungen. Zeitschrift für Deutsche Sprache, Germanistische Mitteilungen. Zeitschrift für Deutsche Sprache, Literatur und Kultur, 40.1. Deutsch kontrastiv. Brüssel. 31-50.

Menzel, Katrin (2016). Textkompetenz in der Fremdsprachenvermittlung und Übersetzerausbildung - ein korpusbasierter Sprach- und Registervergleich zu Kohäsionsmitteln im Englischen und Deutschen In: Thomas Tinnefeld et al. (eds.). SaarbrÃ¼cker Schriften zu Linguistik und Fremdsprachendidaktik. Saarbrücken: HTW.

Menzel, K. (2017). Understanding English-German contrasts - a corpus-based comparative analysis of ellipses as cohesive devices, PhD Dissertation 2016. Universität des Saarlandes.

Nedoluzhko, A. & E. Lapshinova-Koltunski (2016a). Abstract Coreference in a Multilingual Perspective: a View on Czech and German. In Proceedings of CORBON at NAACL-HLT2016, San Diego.

Nedoluzhko A. & E. Lapshinova-Koltunski (2016b). Contrasting Coreference in Czech and German: from Different Frameworks to Joint Results. In: Proceedings of the 22nd INTERNATIONAL CONFERENCE on Computational Linguistics and Intellectual Technologies (Dialogue-2016), 2016, Moscow, RSUH.

Steiner, E. (2015). Contrastive studies of cohesion and their impact on our knowledge of translation. In: Zhang, Meifang and Munday, Jeremy (eds.). Discourse Analysis in Translation Studies Special issue of Target 27:3 (2015).International Journal of Translation Studies. Amsterdam: John Benjamins.

Teich, E., Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H. and Lapshinova-Koltunski, E. (2015a). The Linguistic Construal of Disciplinarity: A Data Mining Approach Using Register Features. Journal of the Association for Information Science and Technology.

Teich, E., S. Degaetano-Ortlieb, H. Kermes, H. & E. Lapshinova-Koltunski (2015b). Register contact: an exploration of recent linguistic trends in the scientific domain. In Gippert, J. and Gehrke, R., eds. Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP, Vol. 5: Historical Corpora: Challenges and Perspectives. Proceedings of the conference Historical Corpora 2012. Narr: Tübingen.

Vela, M. & E. Lapshinova-Koltunski (2015). Register-Based Machine Translation Evaluation with Text Classification Techniques. Proceedings of the 15th MT Summit. Association for Machine Translations in the Americas. Miami, Florida. Oct 30 - Nov 3.

# 3 Summary

The GECCo-project produced a corpus for contrastive linguistic work in the area of textual cohesion. The corpus covers English and German texts in a range of registers and exists in various releases. Its written registers and their lexicogrammatical annotations were imported in a re-organized form from the earlier CroCo-project. The corpus and its documentation can be accessed online for queries with CQPweb `http://fedora.clarin-d.uni-saarland.de/gecco/` / `http://corpora.clarin-d.uni-saarland.de/cqpweb/`. Unrestricted access to the corpus texts is not possible due to property-right restrictions. Querying the corpus, however, with and without GECCo's annotations is possible and open to members of the research community and students.

The linguistic basis of corpus annotations lies in system-based comparisons of cohesive devices in English and German. The annotations allow empirical tests of relevant frequency distributions of cohesive configurations between the two languages, between 14 registers and between spoken and written language use.

The GECCo project produced versions of the GECCo-corpus, though for lexical chains with only a subset of representative registers. The project developed and documented a mixture of automatic pre-coding with human intervention and post-editing to develop sub-corpora of sufficient quality. Collaboration with the Prague Discourse Tree Bank has led to interoperable annotations across theories and corpora.

As far as empirical studies go, statistically refined evaluations of empirical results were produced from both phases of GECCo's lifetime. The project's generalization of cohesive contrasts in terms of degree, strength, semantic type and variation of cohesive devices and chains with particular consideration of the written-spoken dimension is new in the area of contrastive linguistics. This generalization appears necessary as one type of tertium comparationis for cross-linguistic comparison. Together with GECCo's documented annotation guidelines, its statistical evaluation techniques constitute models for a working pipeline in corpus-based empirical work. Empirical results includ wide-ranging overviews on contrasts in cohesion English-German, as well as focused accounts of lexical cohesion and of ellipsis with particular reference to their function across the spoken-written modes.

GECCo has explored and documented three areas of application of its results: its findings feed into language teaching methodologies to allow more discourse-oriented methodologies of teaching, increasing communicative competence. They furthermore provide substantial input for the modelling and teaching of translation, where again an orientation towards improved skills in the creation of target-culture-adapted text production seems highly desirable. And finally, project members jointly with collaborating co-authors produced several publications and prototypes integrating findings from GECCo into linguistic engineering and machine translation, showing how improved control of text cohesion improves linguistic engineering in various aspects.