Chapter 1

Cohesion and coherence in multilingual contexts

Katrin Menzel
Saarland University

Ekaterina Lapshinova-Koltunski
Saarland University

Kerstin Kunz
Heidelberg University

1 Introduction

The volume will investigate textual relations of cohesion and coherence in translation and multilingual text production with a strong focus on innovative methods of empirical analysis as well as technology and computation. Given the amount of multilingual computation that is taking place, this topic is important for both human and machine translation and further multilingual studies.

Coherence and cohesion, the two concepts addressed by the papers in this book, are closely connected and are sometimes even regarded as synonymous (see e.g. Brinker 2010). We draw a distinction concerning the realization by linguistic means.

Coherence first of all is a cognitive phenomenon. Its recognition is rather subjective as it involves text- and reader-based features and refers to the logical flow of interrelated topics (or experiential domains) in a text, thus establishing a mental textual world. Cohesion can be regarded as an explicit indicator of relations between topics in a text. It refers to the text-internal relationship of

Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz. 2017. Cohesion and coherence in multilingual contexts. In Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz (eds.), *New perspectives on cohesion and coherence*, 1–11. Berlin: Language Science Press. DOI:10.5281/zenodo.814458

linguistic elements that are overtly linked via lexical and grammatical devices across sentence boundaries. The main types of cohesion generally stated in the literature are coreference, substitution/ ellipsis, conjunction and lexical cohesion (Halliday & Hasan (1976)). They create relations of identity or comparison, logico-semantic relations or similarity. In the case of coreference and lexical cohesion, COHESIVE CHAINS may contain two or more elements and may span local or global stretches of a text (Halliday & Hasan 1976; Widdowson 1979).

There is another linguistic phenomenon dealt with in several studies of this book, which interacts with cohesion and which also contributes to the overall coherence and topic continuity of a text: INFORMATION STRUCTURE concerns the linguistic marking of textual information as new/ relevant/ salient or old/ less relevant/ less salient (Krifka 2007; Lambrecht 1994). The information in question is presented through linear arrangement of syntactic constituents as either theme or theme, topic or focus or, more generally speaking, in sentence-initial or sentence-final position.

Hence, coherence may or may not be signaled by linguistic markers at the text surface, while cohesion and information structure are explicit linguistic strategies which enhance the recognition of conceptual continuity and the logical flow of topics in texts (Louwerse & Graesser 2007; Halliday & Matthiessen 2004).

One major task involved in the process of translation is to identify the linguistic triggers employed in the source text to develop, relate and change topics. Moreover, the conceptual relations in the mental textual world have to be transferred into the target text by using strategies of cohesion and information structure that conform to target-language conventions. Empirical knowledge about language contrasts in the use of these explicit means and about adequate/ preferred translation strategies is one essential key to systematize the logical flow of topics in human and machine translation. The aim of this volume is to bring together scholars analyzing the cohesion and information structure from different research perspectives that cover translation-relevant topics: language contrast, translationese and machine translation. What these approaches share is that they investigate instantiations of discourse phenomena in multilingual contexts. Moreover, language comparison in the contributions of this volume is based on empirical data. The challenges here can be identified with respect to the following methodological questions:

1. What is the best way to arrive at a cost-effective operationalization of the annotation process when dealing with a broader range of discourse phenomena?

2. Which statistical techniques are needed and are adequate for the analysis? And which methods can be combined for data interpretation?

3. Which applications of the knowledge acquired are possible in multilingual computation, especially in machine translation?

The contributions of different scholars and research groups involved in our volume reflect these questions. All contributions have undergone a rigorous double blind peer reviewing process, each being assessed by two external reviewers. On the one hand, some contributions will concentrate on procedures to analyse cohesion and coherence from a corpus-linguistic perspective (M. Rysová; K. Rysová). On the other hand, our volume will include papers with a particular focus on textual cohesion in parallel corpora that include both originals and translated texts (Kerremans; Kutuzov, Kunilovskaya). Finally, the papers in the volume will also include discussions on the nature of cohesion and coherence with implications for human and machine translation (Lapshinova-Koltunski; Sim Smith, Specia).

Targeting the questions raised above and addressing them together from different research angles, the present volume will contribute to moving empirical translation studies ahead.

2 Phenomena under analysis: Cohesion and coherence

What unifies all of the studies gathered in this volume is that they deal with explicit means of coherence: some works are concerned with particular types of cohesion (M. Rysová; Lapshinova-Koltunski; Sim Smith, Specia), some of them look into the interplay of these different types (Kerremans; Lapshinova-Koltunksi), and some investigate their interaction with information structure (K. Rysová; Kunilovskaya, Kutuzov; Sim Smith, Specia) In most studies, the focus is on the cohesive devices triggering a cohesive relation (M. Rysová; Lapshinova-Koltunski; Kunilovskaya, Kutuzov), others also take account of the relations between cohesive elements (K. Rysová; Kerremans; Sim Smith, Specia).

M. Rysová considers discourse connectives from an etymological perspective in order to set up a structural classification of different connective types for her corpus-linguistic analysis of the Prague Discourse Treebank. Taking account of their degree of grammaticalization, she draws a main distinction between primary and secondary discourse connectives. While both types share their textual function of signaling logico-semantic relations between different textual passages (clauses, clause complexes and larger chunks), they differ in terms of their internal structure as well as their syntactic function.

K. Rysová looks into the interplay of coreference and information structure. She analyses whether different types of coreferential expressions occur in the topic or the focus of a sentence. More precisely, coreferential anaphors or antecedents may collide with syntactic elements that are non-contrastive contextually bound (typically given information), contrastive contextually bound (information on some alternative that can be derived from the context but may not be explicitly given), or non-contextually bound (textually new information).

Kerremans focuses on the interaction of coreference and lexical cohesion in order to determine terminological variants of the same conceptual entity. He groups all nominal elements referring to the same entity in coreference chains and merges these chains with corresponding chains in other texts of the same language. Assigning the coreference chains in the English source texts to the corresponding chains in the Dutch and French target texts eventually permits enriching a terminological database.

Kunilovskaya, Kutuzov consider the mapping of given and new information onto syntactic structure. They train machine learning models to compare originals and translations in terms of (a-) typical patterns at sentence boundaries. For this purpose, they analyze a set of cohesive devices (e.g. pronouns and conjunctions) and other features (e.g. parts of speech, word length) in Russian translations from English and in Russian original texts. Contrasts are identified in terms of where and in which linear order these features occur before and after sentence starts.

Lapshinova compares the distribution of various types of cohesion in human and machine translation. Her focus is on cohesive devices indicating identity of reference (coreference) and logico-semantic relations (conjunction). Within coreference, she distinguishes devices serving as nominal heads (e.g. personal and demonstrative pronouns) and those functioning as modifiers (e.g. the definite article, demonstrative determiners). Conjunctions are classified in terms of their syntactic function (e.g. subordinating or coordinating conjunction and the logico-semantic relation they indicate (e.g. additive or temporal). Translations from English into German and original texts of the two languages.

Sim Smith, Specia investigate the textual distribution of lexical cohesion for improving statistical machine translation. They apply two statistical techniques in order to assess the lexical coherence of texts in a multilingual parallel corpus (English, French and German). Contrasts between languages and between translations and originals are identified by analyzing nominal elements contained in lexical chains of one and the same document. The criteria of comparison included in the research are a) in which sentences these elements appear and b) in which syntactic function (subject vs. other).

3 Corpora and languages

This volume has much to offer to the reader interested in electronic corpora as language resources. It provides information on current research into textual characteristics and discourse structures in different types of language corpora and suggests solutions to questions related to annotation procedures, the quantitative analysis and interpretation of data and machine translation for various languages.

Several types of corpora were used for the studies in this volume. Some contributions focus on large-scale monolingual corpora with the purpose of analyzing a particular language and developing methods that can be applied to other languages as well where similar corpora are available. Some researchers demonstrate the pedagogical and scientific value of native and learner corpora that help to reveal differences between native speakers of a given language and non-native speakers in their ways of creating textuality. Finally, some contributions use bi- or multilingual parallel or comparable corpora consisting either of texts in a language and their translations in another language or of original texts in several languages that are similar with regard to their sampling frame, balance and representativeness.

The annotation of discourse relations and the frequency of discourse connectives in large monolingual corpora such as the the Prague Discourse Treebank 2.0 (PDiT) consisting of Czech newspaper texts as a particular type of written texts are discussed in the chapter by M. Rysová. She examines the historical origin of prototypical discourse connectives in Czech, English and German and demonstrates how these findings can help translators to produce more accurate translations of connectives in these languages. Furthermore, her observations are helpful for the annotation of connectives in large corpora of these languages. Discourse connectives arose from various parts of speech in Czech, English and German and display different stages of grammaticalization. In corpus data for modern stages of the languages investigated in this chapter, they can occur, for instance, in the form of conjunctions, particles, prepositional phrases or fixed collocations. Her chapter provides an angle to address such challenges to annotators of discourse connectives as groups of expressions that may not seem straightforward to define in various languages.

K. Rysová's chapter also addresses the analysis of texts from the Prague Dependency Treebank as a large monolingual corpus and focuses on coreferential relations and information structure in Czech. Her chapter demonstrates that the complexity of text coherence demands extensive language resources of authentic

texts from a given language. Large monolingual corpora with multilayer annotation are still relatively rare for many languages. K. Rysová's analysis encourages research into other languages and recommends applying the methodology she used for the annotation and analysis of coreferential relation and information structure to other languages for which similar resources exist.

Kerremans' chapter demonstrates the invaluable contribution of multilingual parallel corpora including both originals and translated texts as a resource for comparative linguistics and translation studies. The corpus created for Kerremans' study is comprised of written English original texts and their translations into French and Dutch. Terminological variants and coreferential relations from the English source texts have been analyzed from a contrastive perspective. The translation equivalents of these phenomena were retrieved from the French and Dutch target texts in order to create a useful terminological database of translation units and their target-language equivalents for the English-French and the English-Dutch language pairs.

The chapter by Kunilovskaya, Kutuzov deals with the benefits which can be gained from the conjoined use of native and learner corpus data. It compares native and learner varieties of the Russian language with regard to the use of sentence boundaries in a subcorpus of mass media texts from the Russian Learner Translator Corpus. The corpus includes English-Russian learner translations and a genre-comparable subcorpus of the Russian National Corpus, aiming at uncovering differences between native Russian and its learner translated variant.

The chapter by Sim Smith, Specia provides a compelling example of how multilingual corpus data can be used to improve the translation quality in machine-translation models. In this study, original and translated news excerpts in English, French and German from a parallel corpus from the Workshop on Statistical Machine Translation (WMT) were used as well as translations of from French into English from the LIG corpus, which contains news excerpts drawn from various WMT years. The translations that were used for the analysis were provided by human professional translators. They were analyzed with regard to the realisation of lexical coherence, and a multilingual comparative entity-based grid was developed that consists of various types of documents covering the three languages under comparison.

The chapter by Lapshinova-Koltunski describes innovative corpus-based methods to analyze the frequencies and distributions of cohesive devices in multilingual data. Her bilingual corpus contains comparable English and German data for various written text types as well as multiple translations into German which were produced by human translators with different levels of expertise and by

different machine translation systems. This contribution has its focus on the analysis of cohesion in texts from different languages which vary along dimensions such as text-production type, translation method involved and systemic contrasts between source and target language.

4 Methods of investigation

The contributions to this volume cover a wide range of different methods of analysis, starting from manual investigation of previously annotated data, across semi-automatic procedures supporting manual analysis towards fully computational approaches such as entity-grid calculation and automatic sentence segmentation with machine-learning techniques.

Annotation of corpora with information on cohesion- or coherence-related phenomena play a significant role in various descriptive studies based on corpora. They receive particular attention in chapters 2, 3 and 4, in which research design relies to a large extent on annotation. In chapters 5, 6 and partly 7, automatic procedures are used to identify cohesion and coherence phenomena.

Issues of annotation of explicit discourse relations (i.e. relations expressed by concrete language means) in the PDiT are addressed in the study by M. Rysová. She uses the data from PDiT for her analysis to illustrate the difficulty of delineating the boundaries between connectives and non-connectives. For instance, she discusses if frozen lexical forms are a sufficient argument for excluding multiword phrases from discourse connectives and their annotation in the corpus. These phrases clearly signal discourse relations within a text, but they significantly differ from the "prototypical", lexical connectives. The author provides an analysis of historical formation of discourse connectives, justifying their claim that discourse connectives are not a closed class of expressions but rather a scale mapping the grammaticalization of the individual connective expressions. The author believes that this justification may help with the annotation of discourse in large corpora, as was done for PDiT.

The Prague Dependency Treebank was used in the analyses by K. Rysová, who demonstrates how different annotation layers can be used to examine text coherence. The author concentrates on the interplay of two annotation layers: text coreference and sentence information structure. The annotation of sentence information structure is related to contextual boundness, whereas text coreference is understood as the use of different language means for marking the same object of textual reference (the antecedent and the anaphor referents are identical). The author defines all mutual possibilities of coreference relations among con-

textually bound and contextually non-bound sentence items, and analyzes their corpus occurrences. The client-server PML Tree Query (Štěpánek & Pajas 2010) was used to extract the frequency information. The client part is an extension of the tree editor TrEd2 (Pajas & Štěpánek 2008). K. Rysová analyzes the proportion of various mutual possibilities on the basis of corpus occurrences in PDT.

Kerremans uses coreference analysis to study inter- and intralingual terminology variation in a parallel corpus. He proposes a semi-automatic method to annotate terminological patterns that belong to the same coreference chain (called coreferential terminological variants) as an alternative to fully manual labeling, which turns out to be a labour-intensive process. Kerremans method is aimed at supporting manual identification of coreferential terminological variants in the English source texts, annotating these variants according to a common cluster label, extracting them from the text and storing them in a separate database. The automated procedures are implemented in a Perl script ensuring completeness, accuracy and consistency in the data obtained.

Kunilovskaya, Kutuzov also apply semi-automatic procedures to a multilingual corpus that contains both parallel and comparable texts. These semi-automatic procedures are applied to detect divergences in sentence structures between translations into Russian and Russian non-translations. The authors deploy statistical techniques from machine learning: they train a decision-tree model to describe the contextual features of sentence boundaries in the reference corpus of Russian texts, which are considered to be an approximation of the standard language variety. The model is then applied to the translation learner corpus, and translated sentences that are different from the standard language variety are identified through the evaluation of predictors and their combinations. Kunilovskaya, Kutuzov use a number of contextual features in sentence-boundary environments for evaluation. The initial set of 82 features was reduced to 48 with the help of feature selection procedures, allowing them to keep only predictive ones. The results of their analysis permit, on the one hand, to manually inspect cases of the model failing to predict sentence boundaries and possibly find the route causes, and on the other hand, to train another model which predicts not sentence boundaries, but inconsistencies between the first-model decisions and what a translator did in a particular context.

Sim Smith, Specia perform an exploratory analysis of lexical coherence in a multilingual context with a view to identifying patterns that could later be used to improve overall translation quality in machine translation models. They use an entity-grid model and an entity-graph metric – two entity-based frameworks that have previously been used for assessing coherence in a monolingual setting.

The authors try to understand how lexical coherence is realized across different languages and apply these techniques in a multilingual setting for the first time. The entity-grid approach is applied to a parallel corpus. Simply tracking the existence or absence of entities allows for direct comparison across languages. However, entity transition patterns may vary from language to language, while retaining an overall degree of coherence. In order to illustrate the differences between the distributions of entity transitions over the different languages, the authors compute divergence scores. They also analyze the reasons for the observed divergence by taking a closer look at their data.

Lapshinova-Koltunski uses a number of visualisation and statistical techniques to investigate the distributional characteristics of subcorpora in terms of occurrences of cohesive devices in human and machine translation. The cohesive features chosen for the comparative analysis were obtained on the basis of automatic linguistic annotation: tokenisation, lemmatisation, part-of-speech tags and segmentation into syntactic chunks and sentences. Cohesive features are operationalized with the Corpus Query Processor (CQP) queries (Evert 2010). This tool allows definition of language patterns in the form of regular expressions that can integrate string, part-of-speech and chunk tags, as well as further constraints, e.g. position in a sentence. With the help of CQP queries, frequencies of various cohesive features are extracted from a corpus containing translation varieties. Then, various descriptive techniques are used to observe and explore differences between groups of texts and subcorpora under analysis.

5 Conclusion

The contributors to this volume are experts on discourse phenomena and textuality who address these issues from an empirical perspective. We hope that this volume provides an innovative and useful contribution to the advancement of linguistic theory and discourse-oriented corpus studies. This volume also aims at addressing the challenges for human and machine translation arising from the interplay of grammatical and lexical indicators of textual cohesion and coherence.

The chapters in this volume are written in an accessible style. They epitomize the latest research, thus making this book useful to both experts of discourse studies and computational linguistics, as well as advanced students with an interest in these disciplines. We hope that this volume will serve as a catalyst to other researchers and will facilitate further advances in the development of cost-effective annotation procedures, in the application of statistical techniques for

the analysis of linguistic phenomena, the elaboration of new methods for data interpretation in multilingual corpus linguistics and machine translation.

References

Brinker, Klaus. 2010. *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*. 7th edn. Berlin: Erich Schmidt Verlag.

Evert, Stefan. 2010. *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*. Version CWB Version 3.0. The OCWB Development Team. http://cwb.sourceforge.net/.

Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman Publishing.

Halliday, Michael A. K. & Christian Matthiessen. 2004. *Introduction to functional grammar*. 3rd edition. London: Arnold.

Krifka, Manfred. 2007. Basic notions of information structure. In Caroline Fery & Manfred Krifka (eds.), *Interdisciplinary studies of information structure 6*, 13–56. Potsdam: Universitätsverlag.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Louwerse, Max M. & Arthur C. Graesser. 2007. Coherence in discourse. In P. Strazny (ed.), *Encyclopedia of linguistics*, 216–218. Chicago: Fitzroy Dearborn.

Pajas, Petr & Jan Štěpánek. 2008. Recent advances in a Feature-Rich framework for treebank annotation. In Donia Scott & Hans Uszkoreit (eds.), *The 22nd international Conference on Computational Linguistics - Proceedings of the Conference*, vol. 2, 673–680. Manchester, UK: The Coling 2008 Organizing Committee.

Štěpánek, Jan & Petr Pajas. 2010. Querying diverse treebanks in a uniform way. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 1828–1835. Valletta, Malta: European Language Resources Association.

Widdowson, H. G. 1979. *Explorations in applied linguistics*. Oxford: Oxford University Press.