Annotation of Lexical Cohesion in English and German: Automatic and Manual Procedures

Jose Manuel Martinez Martinez
Universität des Saarlandes

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes

Kerstin Kunz
Universität Heidelberg

kerstin.kunz@iued.uni-heidelberg.de
{e.lapshinova,j.martinez}@mx.uni-saarland.de

Abstract

The present paper describes procedures to annotate lexical cohesion in GECCo, a corpus of English and German texts that includes both written and spoken data. Lexical cohesion is an important linguistic component of meaningful discourse and contributes to the overall coherence and thematic continuity of a text. Aiming at a highly precise, fine-grained annotation and avoiding time-consuming procedures, we combine automatic and manual annotation procedures. In this paper, we present the main concepts underlying the annotation and outline the encoding scheme that we apply. We describe the annotation principles and the classification of the sense relations included in our scheme. We also present both automatic and manual procedures and evaluate them in terms of their performance and inter-annotator agreement.

1 Aims and Motivation

This paper describes the annotation of lexical cohesion in GECCo, a corpus of English and German texts that includes both written and spoken data. Lexical cohesion is one of the major types of cohesion contributing to the overall coherence and thematic continuity of a text. It therefore is an important linguistic component of effectively organised and meaningful discourse.

Our overarching goal is an empirical analysis of the realisation of cohesive strategies in English and German and also in written and spoken registers. For this reason, one of the major challenges is defining fine-grained categories that permit the identification of commonalities and differences in terms of various cohesive aspects across the languages and registers under analysis.

As our interest lies in the linguistic properties of lexical cohesion, another challenge is to obtain a highly precise annotation without wasting to much time and labour. Therefore, we start the annotation process with semi-automatic procedures that help to identify candidates of lexical chains and assign their semantic relations. For the sake of convenience, this annotation step was performed on the English texts only. We then proceed with the manual annotation of the English texts. On the one hand, this provides us with a precise annotation of lexical cohesion, and on the other hand, it allows us to test and evaluate the automatic procedures. As the evaluation results indicate unsatisfactory performance of the automatic procedures, we decide to apply only manual annotations for the German texts. In the final step, we evaluate the manual annotation of both English and German texts by calculating inter-annotator agreement. Both automatic and manual procedures are evaluated at three levels: 1) candidate identification, 2) chain construction, and 3) sense relation assignment.

The paper is structured as follows. We provide the theoretical background and state of the art in Section 2 and describe the principles underlying and the categories included into our annotation scheme in Section 3. The annotation procedures are described in Section 4, and their evaluation is presented in Section 5. In Section 6, we summarise and discuss our results.

2 Theoretical Background

Lexical cohesion is regarded as one major type of cohesion contributing to the overall coherence and thematic continuity of a text. The concept was introduced by Halliday and Hasan (1976), whose main focus was on textual relations between linguistic expressions beyond the level of the clause. Halliday and Hasan posit lexical cohesion alongside four other major types of cohesion: coreference, substitution, ellipsis and conjunction. As

illustrated by example (1), lexical cohesions differs from them in terms of **structure** and **semantics**.

(1) *I live in a town called Reigate. It's between London and the countryside which is quite nice. It takes us about 25 minutes to get to London on the train. I say it's a town, it's more of a village. It's quite small. It's very nice actually, it's a nice place to live. And I grew up in a place called Banstead which is fairly close to Reigate.*

2.1 Structure

Contrary to Halliday and Hasan's other four types, the cohesive devices signalling a relation to other expressions in the text are not grammatical items such as proforms, determiners or conjunctions. As the term suggests, the cohesive relation is triggered by lexis, as between *village*, *town* and *place* in example (1). The focus of our project is on the extraction and annotation of nominal elements, although Halliday and Hasan also include relations between verbs, adjectives and adverbs.

2.2 Semantics

The conceptual relation set up by lexical cohesion differs from co-reference and also from what is called bridging in the literature. **Co-reference** and **bridging** are both based on information **instantiated** in the text, the former evoking a relation of identity and the latter a relation of similarity between individual referents in the same text. Our concept of lexical cohesion concerns context-free **sense relations** such as meronymy, hyponymy, synonymy, as described in Lyons (1977) or Winston and Herrmann (1987). Hence what is created from a semantic or conceptual perspective is a relation of similarity between **types** of referents, see also Tanskannen (2006) and Berzlanovich (2008). We also account for **cohesive chains**, which span all nominal elements belonging to the same semantic field (see below). Quite often, devices of lexical cohesion are preceded by co-referential devices, such as the definite article or demonstrative determiners. The interaction of co-reference and lexical chains is assumed to be a major indicator of coherence (Hasan, 1985a; Hasan, 1985b; Martin, 1992). This interaction is left aside here, although it was demonstrated, for instance, by Kunz et al. (2016).

2.3 State of the art in annotation of lexical cohesion

Lexical chains have often been used in natural language processing to solve tasks like text summarization (Doran et al., 2004), or forum thread linking (Wang et al., 2011). However, fewer proposals have tried to use such chains for the study of lexical cohesion (see Teich and Fankhauser (2005) or Bartsch et al. (2009)).

According to Teich and Fankhauser (2004), an automatic lexical chain builder is desirable to reduce human effort devoted to lexical chain annotation and to obtain more consistent results.

Most automatic algorithms rely on either thesauri (Doran et al., 2004; Wang et al., 2011; Fankhauser and Teich, 2004) or statistical associations between words (Wang et al., 2011). The former approach allows one to create not only chains but also to establish various types of semantic relations at the cost of a recall, which is limited to the coverage of the thesaurus.

The annotation of lexical chains is a complex task and so is the operationalization of its evaluation. Some authors (Wang et al. (2011) and Doran et al. (2004)) assess the quality of their techniques to automatically produce lexical chains using an extrinsic approach. This is to test the performance of the system in terms of improving an extrinsic task (forum thread linking and text summarization, respectively). Teich and Fankhauser (2004) carry out an intrinsic evaluation on the methodology used under a purely linguistic point of view, comparing the output of their system with a manually annotated gold standard. However, their evaluation is qualitative.

Further works on lexical cohesion related to natural language processing include Morris and Hirst (1991) and Barzilay and Elhadad (1999).

There exist some works focusing on the comparison of automatic and manual annotations. For instance, Hollingsworth and Teufel (2005) present an approach to directly evaluate the quality of lexical chains, in comparison to a human gold standard. This approach differs from previous evaluation efforts which adopted extrinsic methods relying on word sense disambiguation or on the final application result (the summary or the text segmentation), rather than the focusing on the properties of the lexical chains themselves. The authors also perform a meta-evaluation to compare the best of the metrics used for the evaluation.

3 Annotation Scheme

In order to guarantee consistency throughout the whole process of annotation, detailed descriptions and disambiguation rules had to be defined. They concern the segmentation of nominal elements, the textual distance allowed between nominal elements, the account of different word senses, the classification of the type of sense relation between two nominal elements, and the grouping of several nominal elements in one lexical chain. We can only provide an overview in this paper, details can be found in our annotation guidelines (Kunz, 2014).

Segmentation A nominal element may consist of one noun only, or it may be a compound such as *private teacher*, or a term pattern, such as *head of faculty*. Annotations are based on entries in standard dictionaries (e.g. Cobuild[1] or Longman[2] for English and DWDS[3] for German).

Distance Sense relations are always analysed in linear order, between the two closest elements in a lexical chain. According to Halliday and Hasan's concept, only relations between nominal elements in different clauses, clause complexes, or larger textual passages are cohesive. This would imply that the sense relation between *town* and *Reigate* in example (1) in the first clause is not cohesive but the relation between *town* and *London*. We however decide to annotate all adjacent elements within and outside clause boundaries in order to enhance research on intra- and inter-clausal relations with the help of additional annotation layers available in the corpus.

Word sense If one nominal element occurs more than once in a text and refers two different semantic concepts, and if each of these occurrences enter into a relation to other nominal elements (e.g. *bank* and *financial institution*; *bank* and *building*), two separate lexical chains are established. The assignment of semantic relations follows those defined in WordNet (Fellbaum, 1998) for English, and DWDS for German. The latter integrates GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) and OpenThesaurus (Naber, 2005).

Sense relations We include the following sense relations:

- repetition: orthographical repetition of nominal expressions such as *London* and *London*, or *place* and *place* in example (1) above. In case of compounding, the second element is the determining factor (*stem cell* and *pluripotent cell*, but not *stem cell research* and *stem cell maintenance*).
- antonymy: relation of contrast, as with *inflation* and *deflation*
- synonymy: total synonymy but also near synonymy, such as between technical and common-language terms (e.g. *belly* and *abdomen*).
- hyperonymy: in case the superordinate term follows the more specific term as with *village* and *place*
- hyponymy: in case the specific term follows the superordinate one
- co-hyponymy: between two elements on the same level of specification, such as *town* and *village*
- holonymy: relation, where the whole follows the part (e.g. *quarter* and *town*
- meronymy: part-whole relation, where the part follows the whole (e.g. *town* and *quarter*)
- co-meronymy: succession of two parts that belong to a whole (e.g. *square* and *quarter*).
- type: relation between a common noun and a named entity (e.g. *place* and *Reigate*).
- instance: relation where the named entity follows the common noun (e.g. *Reigate* and *town*).
- co-instance: relation between two named entities (*Reigate* and *London*)

Lexical chains A nominal element can be assigned to a lexical chain if its word sense matches all other elements in the chain, i.e. if one of the types of relations described above could be assigned to the nominal element and each of the other elements in the chain. As a consequence, one nominal element may be a part of several different lexical chains in the same text. See example (2) and the paragraph **Sense relations** above for further discussion.

[1] http://dictionary.reverso.net/english-cobuild
[2] http://www.ldoceonline.com
[3] http://www.dwds.de

4 Annotation Procedures

4.1 Data

The data under analysis includes English and German texts that belong to a variety of registers on a continuum from written to spoken discourse (understood as a sub-dimension of register variation under mode of discourse). The written subcorpus was extracted from the CroCo corpus (Hansen-Schirra et al., 2012), and the spoken subcorpus – from the spoken part of GECCo (Lapshinova-Koltunski et al., 2012).

The registers and the size (in tokens) of annotated subsets are listed in Table 1. ESSAY (political essays) and POPSCI (popular scientific texts) represent written discourse, INTERVIEW (transcribed interviews on various topics) represents spoken discourse, whereas FICTION (fictional texts) contains spoken passages in the form of dialogues and is, in this way, on the borderline between spoken and written registers.

	EO		GO	
register	**texts**	**tokens**	**texts**	**tokens**
ESSAY	23	27171	20	31407
FICTION	10	36996	10	36778
INTERVIEW	9	30057	12	35036
POPSCI	8	27055	9	32639
TOTAL	50	121279	51	135860

Table 1: Information on the corpus size per register

Further annotation layers available in this corpus data include tags on parts of speech, chunks, clause and sentence boundaries, cohesive devices (cohesive reference, conjunction, substitution, ellipsis) triggering coherence in a text, and also chains of relations (for co-reference and ellipsis). The procedures for the annotation of cohesive devices are described by Lapshinova-Koltunski and Kunz (2014).

4.2 Automatic annotation procedure

As starting point for our automatic procedures we used the Little Cohesion Helper (LCH)[4], a piece of software written in Python inspired by Fankhauser and Teich (2004). The authors introduced constraints to filter relevant ties related to WordNet (distance of a word from a root in the WordNet, kind of semantic relationship, minimum depth, etc.) and the text (distance between two words in terms of number of intervening sentences and parts of speech), and the chain themselves (maximum length). Similar strategies are reported by Doran et al. (2004) to weight relations (kind of semantic relationship, and type of match for repetitions – exact, partial, fuzzy) and to discard irrelevant chains (length as number of members in the chain, homogeneity –type-token ratio of chain members–, number of repetitions and type of WordNet relation).

A simplified and schematic expression of the typical algorithm to build chains based on thesaurus look-ups is provided in Figure 1.

The original script takes as input a plain text file using NLTK (Bird et al., 2009) to process the text. First, it tokenizes and splits the text into sentences with `Punkt Tokenizer` (Kiss and Strunk, 2006). Second, it adds POS annotation with `Unigram Tagger` trained on the Brown corpus. And third, it performs a semantic analysis with `WordNet` for all nouns which is the basis for building lexical chains. For each noun, all possible relations with other nouns are checked in reverse order of apparition in the text. This yields cohesive tuples for each word pair. Then, any of the two components of the cohesive tuple is checked as to whether it is already in an existing chain. If yes, the tuple is added to the chain, if not, a new chain is created including this tuple. If the tuple has no relation with the direct preceding word, a look up with other previous items is done, until it finds a related term, adding the information about this relationship. Finally, it saves the result as a MMAX2 project for subsequent manual revision (see Section 4.3 for more details).

We modified LCH with the following goals:

- to port it to Python 3
- to circumvent NLTK tokenization and POS tagging (since this information was already encoded and, more importantly, the original token stream had to be preserved to incorporate this new layer of annotation into the corpus)
- to use lemmas instead of word forms to increase recall using WordNet/GermaNet (specially in German)
- to identify WordNet's multi-word expressions in our texts to increase precision and recall
- to improve file handling and character encoding

[4] `http://lch.sourceforge.net`

- to improve generation of well-formed XML (MMAX2 projects)
- to restore the original token stream to integrate the annotation in the corpus

The final workflow is made up of three steps:

1. corpus preprocessing:
   (a) text boundary identification,
   (b) nominal MWE extraction from WordNet,
   (c) extraction of word forms, lemmata and POS tags for each text,
   (d) identification of WordNet's MWEs in texts.
2. annotation with LCH, for each text:
   (a) obtaining lemmas for nouns
   (b) identification of repetitions of unknown nouns (not found in WordNet)
   (c) extraction of all possible pairs representing semantic relations
   (d) building of lexical chains
   (e) generation of chain links (sorting by consecutive elements)
   (f) serialization of results as MMAX2 project.
3. project postprocessing:
   (a) restoring original tokenization
   (b) updating lexical cohesion annotation accordingly
   (c) replacing lemmas by their word forms.

We describe and discuss the evaluation of this annotation procedure in Section 5.1.

4.3 Manual annotation procedure

For the manual annotation of lexical cohesion in our data, we use MMAX2, a tool for manual annotation (Müller and Strube, 2006) facilitating this process. Texts are annotated by four human annotators with linguistic background. The annotation process consists of three main steps: (1) identification of the candidates for lexical chain members, (2) assignment of links between chain members, (3) assignment of sense relations to chain members.

Candidate identification For the texts in English, we partly keep the automatic pre-annotation of candidates for lexical chain members. However, we remove the sense relations to avoid the influence of automatic assignment on the decision of human annotators. The MMAX2 visualisation allows annotators to decide whether the candidates tagged by LCH belongs to a lexical chain.

As our annotation scheme includes nominal cohesion only, all nouns and noun phrases can be considered as candidates for chain members. However, our analyses show that not every nominal element is included into a lexical chain: 60,84% of all nouns in the English texts and 59,56% of all nouns in the German text are members of lexical chains. For this reason, we decide against the automatic annotation of all nouns as candidates.

Link assignment Human annotators not only identify members of lexical chains and assign their sense relations, but also link the chain members. The MMAX2 tool allows visualisation of links between two or more elements. The annotated information is then encoded as `<lexicalcohesion>` for every member. Each member (markable) is automatically provided with an identification number (ID). Every expression which belongs to the same lexical chain is also assigned to the same ID. This information is saved for every text, and then imported into the corpus. The information on the chains can then be extracted with the help of these IDs.

Sense relation assignment As mentioned above, we analyse the sense relations linking two adjacent chain elements. For this purpose, the type of relation is tagged on the second element of each link. For instance, *place* in example (1) is an hyperonym of the preceding nominal expression *village*, and *place* is a repetition of the preceding nominal expression *place*, and so on. The first element in every chain obviously has no sense relation.

The same word may belong to several lexical chains, and therefore may have several markables with different sense relation assignments. This is especially relevant for words within multiword expressions. For an illustration, see *broadcast industry* and *broadcast legislation* in (2-a) and (2-b), which are elements in long lexical chains.

(2) a. *and Ofcom who is the watchdog for the broadcast industry, to, instead of having it 10 per cent over 10 years, we*

reduce that to 10 per cent over 5 years. (...)

b. *I think that is built into broadcast legislation but it is not there for the cinema legislation, for film legislation. There is no film legislation. (...)*

The whole multiword expression *broadcast industry* in (3-a) is a member of the lexical chain *industry – broadcast industry – industry – industry – industry* tagged as a hyponym of *industry*. At the same time, the multiword expression *broadcast legislation* in (3-b) is also a member of another lexical chain with the head *legislation*: *legislation – broadcast legislation – cinema legislation – film legislation – film legislation – legislation.*

In the process of manual assignment of sense relations, human annotators rely on their intuition. However, they are also allowed to consult various resources to solve problematic cases, e.g. WordNet for English and DWDS, GermaNet and OpenThesaurus for German.

The information on the sense relation is also integrated into the structure `<lexicalcohesion>`, see Figure 5. In this example, the items indexed with 'set_49' belong to the "*legislation*" lexical chain mentioned above. The chain contains nine elements and starts with the word *chain* which is, however, outside the text span provided in Figure 5. *Broadcast legislation* is its hyponym, and *cinema legislation* is the co-hyponym of *broadcast legislation*, whereas *film legislation* is the co-hyponym of *cinema legislation*. The second mention of *film legislation* is a repetition. The other set (set_113) in the example in Figure 5 is represented by the lexical chain *UK – Europe – UK*, and is a case of holonymy-meronymy relations.

4.4 Annotation statistics

We summarise the statistics on the structures annotated for lexical cohesion in our data in Table 2. Whereas Table 3 provides statistics on the annotated relations classified per relation.

	EO	GO
nr of chains	2598	1783
nr of relations	11814	11568

Table 2: Manually annotated structures in GECCo

	EO	GO
repetition	6925	6191
hypernym	1046	1104
hyponym	1033	1159
synonym	579	608
co-hyponym	520	570
meronym	436	340
holonym	426	308
antonym	292	465
instance	190	238
type	175	203
co-instance	172	307
co-meronym	75	100
gennoun	1	3

Table 3: Manually annotated sense relations in GECCo

5 Annotation Evaluation

In the evaluation step, we compare automatic and manual annotations (for English texts only), as well as the annotations produced by different annotators (on a sample of English and German texts). The comparison is performed for the following features: 1) markables representing candidate identification, 2) chains representing link assignment, 3) semantic relations representing sense relation assignment.

Markables from both annotation versions are aligned on the basis of their token IDs. Each markable pair containing at least one token in common is considered a markable alignment. We use Jaccard distance to take into account perfect (all tokens in both markables were the same) and spurious (only some tokens were in common) agreement.

Chain alignment is done by retrieving the chain IDs of the markables aligned in the previous step. We consider a chain alignment any pair of chains having at least one markable in common. Upon identifying the alignments, chain members are retrieved. We use Jaccard distance again to take into account perfect (all markables in both chains are the same) and spurious (only some markables are shared across both chains) agreement.

To evaluate the assigned relations, we collect subsets of aligned markables which share the preceding member in a chain. If the condition is satisfied, the semantic relation assigned to the selected markable is compared. Since only one label is provided, we used binary distance to calculate the agreement. If the relation label is the same on both aligned markables, the agreement is 1, if they are

different agreement is 0.

For each level of analysis we provide the following measures: precision ($P = \frac{|M \cap A|}{|A|}$ where M is the reference dataset –for **M**anual– and A is the test –for **A**utomatic), recall ($R = \frac{|M \cap A|}{|M|}$), and the F-score ($F$, the harmonic mean of P and R, weighted by $\alpha = 0.5$) of the elements annotated by the automatic system, together with the Jaccard coeficient of similarity $J = \frac{I}{U}$, which accounts for the proportion of elements present in both data sets ($I = |M \cap A|$) over the total number of elements being compared ($U = |M \cup A|$). Moreover, we report on the level of agreement for the intersection of elements with the manual reference annotation (I) using Cohen's Kappa (κ) as implemented in NLTK `nltk.metrics.agreement` (Bird et al., 2009) and described by Artstein and Poesio (2008).

We plot a confusion matrix to visualize the quality of sense relation assignments produced by the automatic system (or by human annotators) in relation with a reference annotation. Such a plot depicts the prediction on the X axis (e.g. automatic annotation), and the reference on the Y axis (e.g. manual annotation). The resulting diagonal displays the instances where there is agreement, which means that the predicted label is equal to the true label. The off-diagonal cells represent mislabelling or disagreement. A diagonal with high values is an indicator of many correct predictions or a good agreement.

5.1 Automatic vs. manual procedures

We firstly report on the comparison of the output of the automatic procedures explained in Section 4.2 with its manual annotation. As previously mentioned, the automatic procedures were applied on the English subcorpus only. Table 4 summarises all the measures calculated to evaluate the quality and agreement of the annotation.

	markables	chains	relations
U	23832	5700	11884
I	11884	4089	3262
J	0.50	0.72	0.28
P	0.60	0.65	0.17
R	0.77	0.69	0.22
F	0.67	0.67	0.19
κ	0.90	0.38	0.47

Table 4: Evaluation measures for automatic annotation of lexical cohesion chains.

Markables A total of 23832 markables are compared (U), the intersection of markables present in both annotation sets (I) amounts to 11884 items, what represents a 50 % of them showing some kind of overlap (J). Precision ($P = 0.6$), recall ($R = 0.77$) and the F-score ($F = 0.67$) are low in comparison with the human performance (see Section 5.2). The agreement between both versions at markable level is $\kappa = 0.90$. However, if we extrapolate this measure to the total number of chain members annotated in both versions, the agreement sinks to a mere 45 %.

Chains A total of 5704 chains are compared (U). 72 % of the chains overlap (J). Precision ($P = 0.65$), recall ($R = 0.69$) and the F-score ($F = 0.67$) are lower than human performance. The agreement between both versions regarding the overlapping chains is $\kappa = 0.38$. This clearly indicates that chains share a very low proportion of members in common. If we extrapolate the agreement to the total number of chains, the agreement falls to 27 %.

Relations From the 11884 markables aligned across both versions (U), only 28 % of the markables refer to the same anteceding member in their respective chains (J)). Precision ($P = 0.17$), recall ($R = 0.22$) and the F-score ($F = 0.19$) are very low indicating that the internal arrangement of members within the automatically assigned chain is very different from the human reference. The agreement in the assignment of the relation labels is $\kappa = 0.47$. Nevertheless, this subset of relations represents just 14 % of all the relations annotated. If we extrapolate the agreement to the total number of relations annotated with both methods the agreement drops to 11 %.

The confusion matrix displayed in Figure 2 shows a very low precision of the automatic system (light shadowed diagonal). The automatic system seems to assign many repetitions to instances where humans chose other categories. This can be explained by the nature of the automatic procedures assigning repetitions to the nouns that are not covered by WordNet. Another influential factor is the difference in the subset of the relations used by the automatic system (antonym, holonym, hypernym, hyponym, meronym, synonym, repetition) and the one used by humans who had six additional relations at their disposal (co-hyponym, co-instance, co-meronym, instance, type).

5.2 Inter-annotator agreement in manual procedures

The inter-annotator agreement is calculated on a subset of texts containing 5925 tokens in English and 7102 in German roughly representing a 5 % of the manually annotated subcorpus. The proportion of lexical chains revised for both English (see Table 5) and German (see Table 6) also reaches a 5 %, what in turn amounts to around a 7 % of all sense relations.

	markables	chains	relations
U	1123	175	903
I	903	146	465
J	0.80	0.83	0.52
P	0.92	0.95	0.49
R	0.86	0.82	0.45
F	0.89	0.88	0.47
κ	0.94	0.59	0.62

Table 5: Evaluation measures for IAA in manual annotation of lexical cohesion chains for English.

	markables	chains	relations
U	1169	215	821
I	821	169	350
J	0.70	0.79	0.43
P	0.89	0.98	0.39
R	0.76	0.88	0.33
F	0.82	0.93	0.36
κ	0.97	0.55	0.63

Table 6: Evaluation measures for IAA in manual annotation of lexical cohesion chains for German.

Markables We compare a total of 1123 markables in English (U) of which 80% showed an overlap (J). Precision ($P = 0.92$), recall ($R = 0.86$) and the F-score ($F = 0.89$) are higher than the values for the automatic system. The agreement between both annotators at the markable level is $\kappa = 0.95$. If we extrapolate this measure to the total number of markables annotated in both versions, the agreement remains close to 75 %, which is much better than the IAA achieved with the automatic procedure.

As for German, a total of 1169 are compared (U) showing an overlap of 70 % (see J in the table). The agreement is $\kappa = 0.97$, but extrapolated to the total number of markables it goes down to about 68 %.

Chains A total of 175 chains are compared in English (U) showing an overlap of 83 % (J). Both annotators reached an agreement of $\kappa = 0.59$ for the overlapping chains. If the number is extrapolated for all chains, the agreement still reaches 48 %.

78 % of the 214 chains in German are overlapping. Their agreement amounts to $\kappa = 0.55$ for these subset of chains, what represents only 43 % of the agreement for the total number of chains.

Precision, recall and the F-score are very similar for both languages at this level.

Relations From the 903 markables aligned across both of annotators in English (U), 52 % refer to the same preceding member across chains (J). The agreement regarding the assignment of a semantic relation for these pairs is $\kappa = 0.62$. This subset of relations represents in turn 41 % of the total number of markables annotated. If we extrapolate this agreement to the total number of relations annotated by both annotators, the proportion of relations showing agreement amounts for 30 %.

The confusion matrix plotted in Figure 3 shows a fairly good agreement for most categories, except co-hyponyms competing with co-instances, and co-meronyms with co-hyponyms, as well as synonymy.

43 % of the 821 markables in the German sample aligned across both annotators refers to the same preceding member across chains. The agreement for these pairs is $\kappa = 0.63$. If we extrapolate this agreement to all the markables analyzed it decreases to 21 %.

The confusion matrix in Figure 4, enables the examination of the results broken down by semantic relations. We observe that the relations of antonym, co-hyponym, hyperonym, hyponym, repetition and synonym display a fairly good agreement. However, it is weaker for the rest of the relations as indicated by the lighter grey tones of the cells in the diagonal, and darker grey shadows out of the diagonal.

Taking into account all indicators, annotators of English texts show a slightly higher IAA than those of the German ones. This may be due to the higher number of repetitions in English than German, which can be more easily identified than e.g. relations of synonymy. Moreover, lower inter-annotator agreement in German may go along with a higher degree of lexical specification.

6 Conclusion and Discussion

In the present paper, we have provided an insight into the annotation procedures underlying our analysis of lexical cohesion in English and German spoken and written texts. Our initial approach included the combination of automatic and manual procedures. The automatic annotation procedure employs basic heuristics linking nouns and noun phrases greedily if a certain type of a link can be found in WordNet. In some respects, this is a high-recall strategy desirable in scenario combining an automatic pre-annotation and manual post-correction, which was originally our intention.

We have performed a thorough evaluation of the automatic annotation calculating IAA between the automatic system and the human annotators. The main challenges for the evaluation are unitisation issues hindering comparability (see Wacholder et al. (2014) for a similar scenario) and the complexity of assessing multiple annotation choices (candidate identification, chain membership, link assignment, and sense relation assignment) which are comprised in the task of building lexical chains. Our evaluation goes beyond previous intrinsic evaluations of this task.

Although the number of markables and chains seems to be similar in both datasets, the representation of lexical cohesion by means of lexical chains and the internal structure of the chains differs in automatic and human annotations, as shown in the evaluation of chains and relations. The performance of the automatic system presented in Section 4.2 is much lower than the human reference quantified in terms of precision, recall and IAA. These differences have an important effect on higher level dimensions such as topic development and overall semantic variation. Their correction turned out to be even more time consuming than a purely manual procedure. This was confirmed by the feedback provided by human annotators, who considered it much easier to build lexical chains and annotate relations from scratch than post-editing the system's output.

The evaluation of our manual procedures show that overall, we achieve a good IAA in the annotation of both German and English texts. The agreement scores however show that annotating lexical cohesion chains is a difficult task even for humans. Annotators showed a higher degree of agreement in English than in German across all levels of comparison. The challenge not only arises from the high conceptual level of the linguistic analysis but also from the complexity of the annotation which is made up of different subtasks.

Acknowledgments

This paper is based on work carried out in the frame of the GECCo[5] project funded by the German Research Foundation (DFG) under GZ STE 840/6-2 and KU 3129/1-2 *Kohäsion im Deutschen und Englischen – ein empirischer Ansatz zum kontrastiven Vergleich*.

References

Ron Artstein and Massimo Poesio. 2008. Survey Article Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 4(34):555–596.

Sabine Bartsch, Stefania Degaetano, Tomek Grubba, Nina Petrychka, David Sullivan, Christoph Tragl, and Claudio Weck. 2009. ObamaSpeeches.com Building and Processing a Corpus of Political Speeches. In Elke Teich, Andreas Witt, and Peter Fankhauser, editors, *Poster at Proceedings of GSCL Workshop: Linguistic Processing Pipelines.*, pages 41–42, Potsdam, sep.

Regina Barzilay and Michael Elhadad. 1999. Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121.

Ildikó Berzlanovich. 2008. *Lexical cohesion and the organization of discourse. First year PhD report*. University of Groningen, Groningen.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

William Doran, Nicola Stokes, Joe Carthy, and John Dunnion. 2004. Comparing lexical chain-based summarisation approaches using an extrinsic evaluation. In P Soijka, K Pala, P Smrz, C Fellbaum, and P Vossen, editors, *Proceedings of the 2nd Global WordNet Conference, 20 - 23 January*, page 112, Brno, jan. Masaryk University.

Peter Fankhauser and Elke Teich. 2004. Multiple perspectives on text using multiple resources: Experiences with XML processing. In *Proceedings of LREC Workshop on XML-based richly annotated corpora, 4th Conference on Language Resources and Evaluation (LREC)*, pages 15–20, Lisbon, Portugal, may.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*, volume 71. MIT Press, Cambridge, MA.

[5] http://www.gecco.uni-saarland.de

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London, New York.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, jul.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

Ruqaiya Hasan. 1985a. The structure of a text. In M.A.K. Hallday and R. Hasan, editors, *Text and context: aspects of language in a social-semiotic perspective*, pages 52–96. Oxford University Press, Oxford.

Ruqaiya Hasan. 1985b. The texture of a text. In M.A.K. Hallday and R. Hasan, editors, *Text and context: aspects of language in a social-semiotic perspective*, pages 70–96. Oxford University Press, Oxford.

Verena Henrich and Erhard Hinrichs. 2010. Gernedit – the germanet editing tool. In *The Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, Valletta, Malta.

Bill Hollingsworth and Simone Teufel. 2005. Human annotation of lexical chains: coverage and agreement measures. In *he ACM International Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications (ELECTRA 2005) held at SIGIR 2005*, volume 39, New York, NY, USA. CM SIGIR Forum Homepage archive.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational linguistics*, 32:485–525.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and 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, jun.

Kerstin Kunz. 2014. Annotation guidelines for lexical cohesion. , Universität des Saarlandes.

Ekaterina Lapshinova-Koltunski and Kerstin Kunz. 2014. Annotating cohesion for multillingual analysis. In *Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Reykjavik, Iceland, May. LREC.

Ekaterina Lapshinova-Koltunski, Kerstin Kunz, and Marilisa Amoia. 2012. Compiling a multilingual spoken corpus. In Tommaso Raso Heliana Mello, Massimo Pettorino, editor, *Proceedings of the VIIth GSCP International Conference: Speech and corpora*, pages 79–84, Firenze. Firenze University Press.

John Lyons. 1977. *Semantics*, volume 1–2. Cambridge University Press, Cambridge, UK.

James R. Martin. 1992. *English Text. System and Structure*. John Benjamins, Amsterdam, Netherlands.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17:21–48.

Christoph Müller and Michael Strube. 2006. Multilevel annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Daniel Naber. 2005. Openthesaurus: ein offenes deutsches wortnetz. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beitrge zur GLDV-Tagung 2005*, pages 422–433, Bonn. Peter-Lang-Verlag, Frankfurt.

Sanna-Kaisa Tanskannen. 2006. *Collaborating towards Coherence*. John Benjamins, Amsterdam, Netherlands.

Elke Teich and Peter Fankhauser. 2004. WordNet for lexical cohesion analysis. In P Soijka, K Pala, P Smrz, C Fellbaum, and P Vossen, editors, *Proceedings of the 2nd Global WordNet Conference, 20 - 23 January*, pages 326–331. Masaryk University, Brno, Czech republic.

Elke Teich and Peter Fankhauser. 2005. Exploring lexical patterns in text: lexical cohesion analysis with WordNet. In *Heterogeneity in focus: Creating and using linguistic databases. Interdisciplinary studies on information structure*, volume 2, pages 129–145. Universität Potsdam.

Nina Wacholder, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. 2014. Annotating Multiparty Discourse: Challenges for Agreement Metrics. In Lori Levin and Manfred Stede, editors, *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 120–128, Dublin, aug.

Li Wang, Diana Mccarthy, and Timothy Baldwin. 2011. Predicting Thread Linking Structure by Lexical Chaining. In Diego Mollá and David Martinez, editors, *Proceedings of the Australasian Language Technology Association Workshop 2011*, volume 9, pages 76–85, Canberra, dec. Australasian Language Technology Association.

Chaffin R. Winston, M.E. and D. Herrmann. 1987. A taxonomy of part-whole relations. In *Cognitive Science*, number 11, pages 417–444.

A Figures

```
candidates = []
for token in tokens_of_text:
  if token == noun:
    append.candidates(token)
ties = []
for candidate in candidates:
  all_pairs = get_all_pairs(candidate,
    ↪  candidates)
  all_pairs = filter_pairs(all_pairs)
  append.ties(all_pairs)
chains = []
for chain in chains:
  for tie in ties:
    if tie[0] in chain or tie[1] in
      ↪  chain:
      append.chain(tie)
    else:
      new_chain = [tie]
      chains.append(new_chain)
for chain in chains:
  chain = link_ties(chain)
chains = filter(chains)
```

Figure 1: Pseudo-code for lexical chain building algorithm

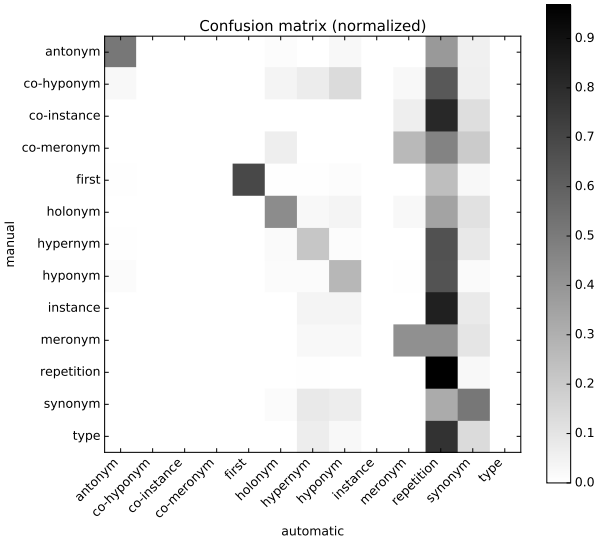

Figure 2: Confusion matrix for annotation of semantic relations manual vs. automatic in English.

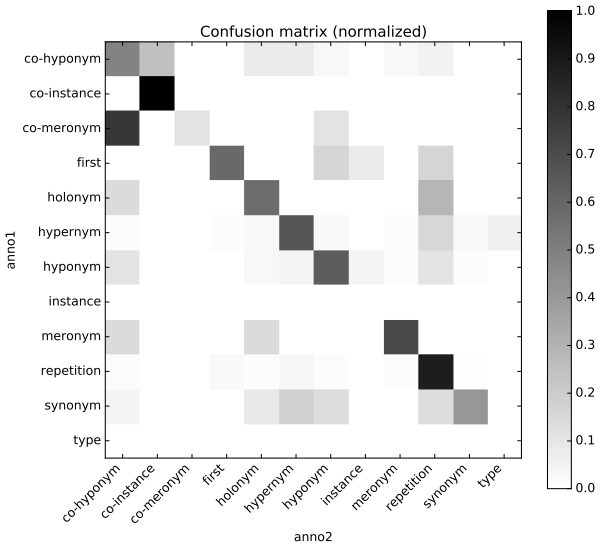

Figure 3: Confusion matrix for annotation of semantic relations Annotator 1 vs. Annotator 2 in English.

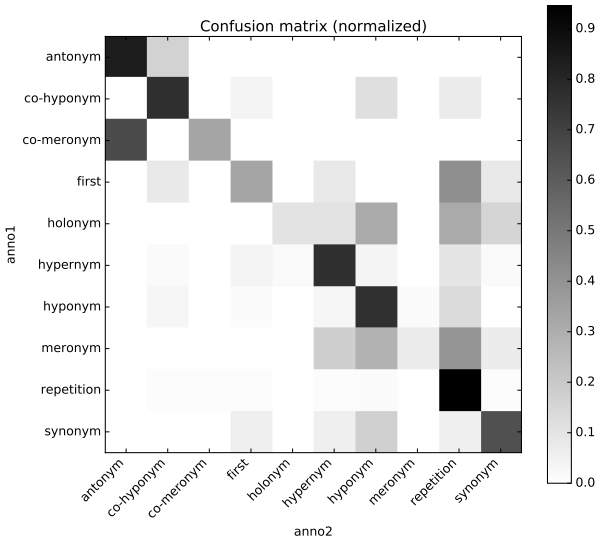

Figure 4: Confusion matrix for annotation of semantic relations Annotator 1 vs. Annotator 2 in German.

```
I think that is built into <lexicalcohesion id="markable_313" lexical_type="hyponym"
  ↪ lexical_chain="set_49"> broadcast legislation </lexicalcohesion> but it there
  ↪ it is not there for the <lexicalcohesion id="markable_377"
  ↪ lexical_type="co-hyponym" lexical_chain="set_49"> cinema legislation
  ↪ </lexicalcohesion>, for <lexicalcohesion id="markable_378"
  ↪ lexical_type="co-hyponym" lexical_chain="set_49"> film  legislation
  ↪ </lexicalcohesion>. There is no <lexicalcohesion id="markable_316"
  ↪ lexical_type="repetition"  lexical_chain="set_49"> film legislation
  ↪ </lexicalcohesion>. I know that the <lexicalcohesion id="markable_312"
  ↪ lexical_type="repetition"  lexical_chain="set_113"> UK </lexicalcohesion> 's
  ↪ quite advanced, isn't it, in terms of audiodescription compared with the rest
  ↪ of  <lexicalcohesion id="markable_318" lexical_type="holonym"
  ↪ lexical_chain="set_113"> Europe </lexicalcohesion> , for example . What do you
  ↪ think it is that makes us, or makes the <lexicalcohesion id="markable_319"
  ↪ lexical_type="meronym"  lexical_chain="set_113"> UK </lexicalcohesion> UK, so
  ↪ good at doing this?
```

Figure 5: Annotated lexical chains in the corpus