Exploring Lexical Patterns in Text: Lexical Cohesion Analysis with WordNet

Elke Teich[*], *Peter Fankhauser*[**]

Technische Universität Darmstadt[*], FhG-IPSI, Darmstadt[**]

We present a system for the linguistic exploration and analysis of lexical cohesion in English texts. Using an electronic thesaurus-like resource, Princeton WordNet, and the Brown Corpus of English, we have implemented a process of annotating text with lexical chains and a graphical user interface for inspection of the annotated text. We describe the system and report on some sample linguistic analyses carried out using the combined thesaurus-corpus resource.

1 Introduction

In recent years, there has been an increasing activity in building up corpora annotated at multiple linguistic levels (syllable, word, clause, text) and strata (phonology, grammar, semantics). With the growing interest in such *multi-layer corpora* comes the need for tools that support corpus annotation and exploration of the resulting annotations as well as facilitate further computational processing.

For the lower levels of the linguistic system (grammatical units, such as words, phrases, clauses), there are plenty of tools that provide the necessary functionalities. For instance, at the stratum of grammar, part-of-speech tagging and shallow phrase structure parsing can be carried out automatically at reasonable accuracy, with hardly any human intervention. Also, there are some rather mature tools for corpus inspection, such as special-purpose query (e.g., CQP (Christ, 1994a)), TIGERSearch (Lezius and König, 2000; König and Lezius, 2003), concordancers (e.g., XKwic (Christ, 1994b)) and browsers for tree structures (e.g., Annotate (Plaehn and Brants, 2000)).

Interdisciplinary Studies on Information Structure 02 (2005): 129–145
Dipper, S., M. Götze and M. Stede (eds.):
Heterogeneity in Focus: Creating and Using Linguistic Databases

However, when it comes to the unit of *text* and the analysis of *meaning*, the situation is difficult in two respects. First, fully automatic annotation is often not possible; second, tools supporting annotation and exploration exist only for selected aspects of textual analysis, e.g., for rhetorical structure (O'Donnell, 1997). Rhetorical structure is clearly an important aspect of a text's organization and vital for a full-blown interpretation of a text. But there are many other meaning-creating features in a text, which are interesting from the viewpoints of both linguistic theory and computational-linguistic processing. One such feature is *cohesion*.

1.1 Corpora Annotated for Cohesion: Motivation, Goals, Tools

Cohesion is defined as the set of linguistic means we have available for creating *texture* (Halliday and Hasan, 1976, 2), i.e., the property of a text of being an interpretable whole (rather than unconnected sentences). Cohesion occurs "where the interpretation of some element in the text is dependent on that of another. The one presupposes the other, in the sense that it cannot be effectively decoded except by recourse to it." (Halliday and Hasan, 1976, 4).

The most often cited type of cohesion is *reference*.[1] Consider example (1) (from Halliday and Hasan, 1976, 2).

(1) Wash and core six cooking *apples*. Put *them* into a fireproof dish.

In example (1), it is the cohesive tie of coreference between *them* and *apples* that gives cohesion to the two sentences, so that we interpret them as a text. The detection of such referential ties is clearly essential for the semantic interpretation of a text. Corpora annotated for reference relations are thus of interest for both linguistics, e.g., for testing theories of information structure (loci

[1] Also known as *coreference* or *anaphora* and often taken to include substitution and ellipsis, i.e., *one*-anaphora and zero-anaphora.

of high/low informational load, informational statuses (Given/New)), and computational processing, e.g., for applications such as information extraction or information retrieval.

Another type of cohesion, coacting with reference to create texture, is *lexical cohesion* (cf. Halliday and Hasan, 1976). Lexical cohesion is the central device for making texts hang together experientially, defining the aboutness of a text (cf. Halliday and Hasan, 1976, chapter 6). Typically, lexical cohesion makes the most substantive contribution to texture: According to Hasan (1984) and Hoey (1991), around fourty to fifty percent of a text's cohesive ties are lexical.

In its simplest incarnation, lexical cohesion operates with *repetition*, either simple string repetition or repetition by means of inflectional and derivational variants of the word contracting a cohesive tie. The more complex types of lexical cohesion work on the basis of the semantic relationships between words in terms of *sense relations*, such as synonymy, hyponymy, antonymy and meronymy (cf. Halliday and Hasan, 1976, 278–282). See examples of a meronymic relation (highlighted in italics) and an antonymic relation (highlighted in bold face) in (2) below; the latter at the same time is a case of repetition.[2]

(2) Tone languages use for **linguistic** contrasts *speech* parameters which also function heavily in **non-linguistic** use. [...] The problem is to disentangle the **linguistic** parameters of *pitch* from the co-occurring **non-linguistic** features.

In a text, potentially any occurrence of repetition or relatedness by sense can form a cohesive tie; but not every instance of semantic relatedness between two words in a text does necessarily create a cohesive effect. For example, if a word *linguists* occurring in sentence 1 of a text containing eighty sentences is

[2] The example is taken from text j34 of the Brown corpus.

repeated in sentence 76, a cohesive effect is rather unlikely. Also, there seem to be stronger cohesive effects involving the register-specific vocabulary rather than the “general” vobulary (cf. Section 3).

Detailed manual analyses of small samples of text (e.g., Hoey, 1991) can bring out some tendencies of how lexical cohesion is achieved; but in order to arrive at any generalizations, large amounts of texts annotated for lexical ties are needed. Manual analysis is very labor-intensive, however, and the level of inter-annotator agreement is typically not satisfactory. Thus, an automatic procedure is called for. Fortunately, lexical cohesion analysis is a suitable candidate for automization: Texts systematically make use of the semantic relations between words and detecting lexical cohesive ties simply means checking the relatedness of words in a text against a thesaurus or thesaurus-like resource. A few additional constraints must be added to arrive at plausible lexical chains, such as, e.g., the afore mentioned distance between words in a text or the specificity of the vocabulary (see also Section 2).

Automatic lexical cohesion analysis has been applied in computational linguistics for automatic text summarization (e.g., Barzilay and Elhadad, 1997). Our own motivation for building a system that automatically annotates text in terms of lexical cohesion has been to be able to explore the workings of lexical cohesion in more detail, asking questions such as (cf. Fankhauser and Teich, 2004): In a given text, what are the dominant lexical chains (indicating what the text is mainly about)? Are there differences in the strength of lexical cohesion according to the register and/or genre of a text? In a given register/genre, are there any patterns of lexical cohesion (e.g., hyponymy-hypernymy, holonymy-meronymy) that occur significantly more often than others? Can the internal make-up of lexical chains tell us anything about the genre of a text (e.g., narrative vs. factual)?

1.2 Summary; Overview of Paper

With the growing interest in richly annotated corpora, there is an increasing need for tools supporting annotation as well as exploration of corpus resources, both for linguistic and for computational purposes. The corpus processing of grammatical units is pretty well understood, but there are many unresolved issues when it comes to processing corpora at the level of text. The system we present in this paper addresses one such issue, namely the annotation and exploration of lexical cohesion.

Section 2 introduces our approach to annotation of lexical cohesion and describes the functionalities of the system. Section 3 provides some examples of linguistic analysis that we have carried out using the data generated by our system. Finally, we conclude with a summary and outlook on future research (Section 4).

2 Automatic Analysis of Lexical Cohesion

The basic means for lexical cohesion analysis are so called lexical chains, which consist of words that are related by a lexically cohesive tie. Using the SEMCOR version of the Brown Corpus, which is sense tagged with so called synsets from the Princeton WordNet (version 1.6), these ties can be determined by navigating along the relationships (synonymy, hypernymy, hyponymy, antonymy, and various kinds of meronymy) in WordNet. In addition to the direct relationships we also take into account indirect relationships, including transitive hypernymy, hyponymy, and meronymy, co-hypernymy, and co-meronymy, and ties observable directly from the text, including repetition of lemmas and of proper nouns. A more detailed description of the resources and the processing steps is given in Fankhauser and Teich (2004).

Not all the ties automatically determined in this way are necessarily cohe-

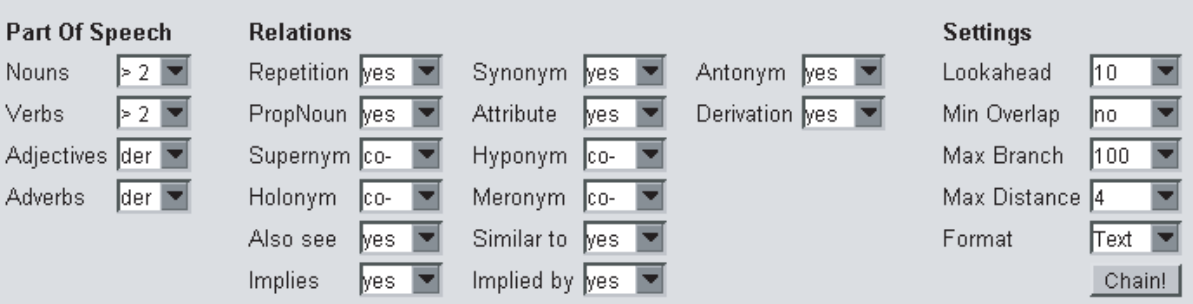


Figure 1: Options for cohesion analysis

sive. A number of factors can help in ruling out non-cohesive ties:

- Specificity and part-of-speech: A specific noun like *tone_system* is more likely to contract a lexically cohesive tie than a general verb like *be*.
- Kind of the semantic relationship: Repetition and synonymy form stronger ties than hypernymy or meronymy.
- Strength of the relationship: The direct hypernym *phonologic_system* forms a stronger cohesive tie with *tone_system* than the remote hypernym *system*.
- Distance in text: Words with many intervening words, sentences, or paragraphs are less likely to contract a cohesive tie than close words.

Our system allows fine-tuning these factors as shown in Figure 1.

The depicted settings (Part Of Speech) take only into account ties between specific nouns and verbs, which are at least at depth 3 in the WordNet hypernymy hierarchy, and include adjectives and adverbs only if they are directly related to an included noun or verb. Moreover, ties may not span more than 10 sentences (Lookahead), and transitive relationships may comprise at most 4 steps (Max Distance) with a branching factor of at most 100 alternative paths

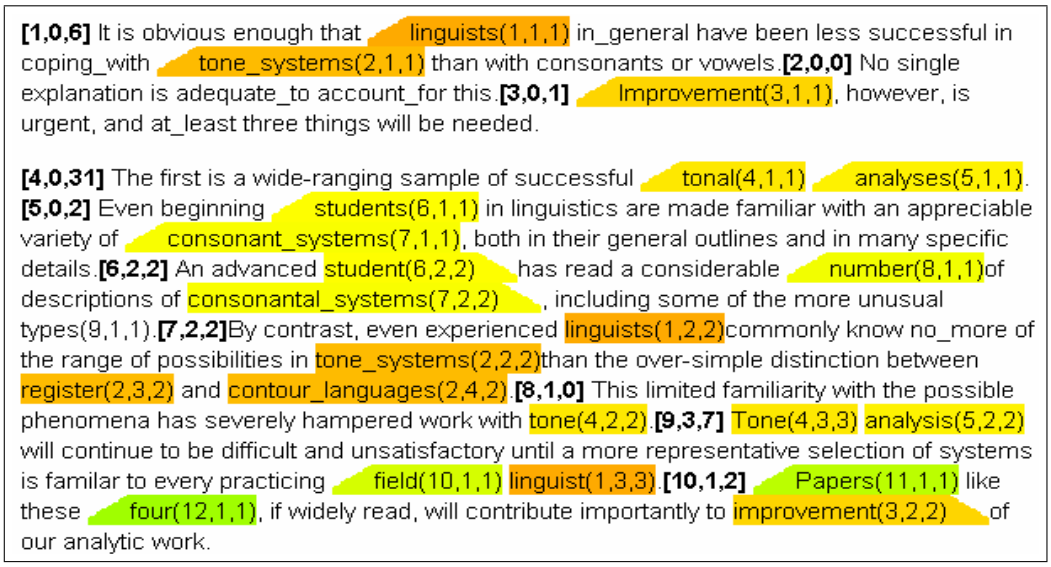


Figure 2: Text view on annotated text

(Max Branch). The kinds of relationships are not further constrained in the example setting.

Lexical chains can then be inspected from three perspectives. In the *text view* (Figure 2), each lexical chain is highlighted with an individual color, in such a way that chains starting in succession are close in color. In addition, for each sentence its number, the number of preceding sentences and the number of following sentences with a word in the same chain are given. This view can give a quick grasp on the overall topic flow in the text to the extent that it is represented by lexical cohesion.

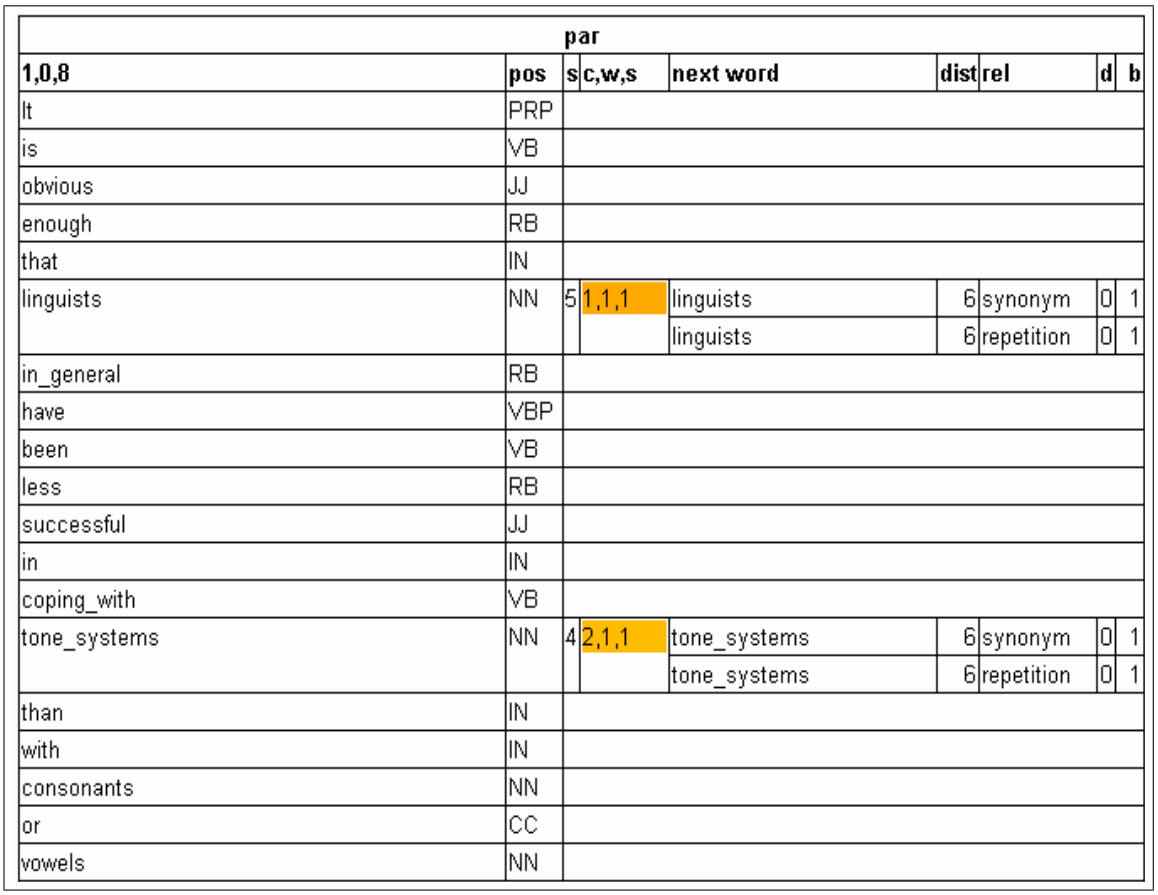
The *chain view* (Figure 3) presents chains as a table with one row for each sentence, and a column for each chain ordered by the number of words contained in it. In addition, each chain gives its most frequent word (*domwf*), and the absolute and relative number of kinds of relationships forming a tie (*repsyn* for repetition with synonymy, *rep* for repetition without synonymy, etc.). This view also reflects the topical organization fairly well by grouping the dominant chains closely.

#s, #w	22, 30	23, 28	10, 14	11, 11	5, 7	6, 7	6, 6	4, 6	5, 5	3, 5
domwf	tone	morphophonemics	phonemic_system	orthography	intonation	rule	field	theory	linguist	tone_system
repsyn	16 (55,2%)	6 (27,3%)	1 (8,3%)	5 (55,6%)	6 (100%)	6 (100%	5 (100%	5 (100%)	4 (100%)	1 (25%)
rep	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
syn	0 (0%)	1 (4,5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
super	3 (10,3%)	0 (0%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (25%)
hypo	2 (6,9%)	0 (0%)	1 (8,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
cohypo	0 (0%)	3 (13,6%)	9 (75%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (25%)
part	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (25%)
other	8 (27,6%)	12 (54,5%)	0 (0%)	4 (44,4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
par	4	19	33	53	30	42	10	16	1	2
1									linguists	tone_systems
2										
3										
par	4	19	33	53	30	42	10	16	1	2
4	tonal									
5										
6										
7									linguists	tone_systems register contour_languages
8	tone									
9	Tone						field		linguist	

Figure 3: Chain view on annotated text

Finally, the *tie view* (Figure 4) displays for each word all its (direct) cohesive ties together with their properties (kind, distance, etc.). This view is mainly useful for checking the automatically determined ties in detail.

In addition, all views provide hyperlinks to the WordNet classification for each word in a chain to explore its semantic neighborhood. Moreover, some statistics, such as the number of sentences linking to and linked from a sentence, and the relative percentage of ties contributing to a chain are presented. These and some other statistics can then also be exported to a standard statistics package, such as MS Excel or SPSS.



par								
1,0,8	pos	s	c,w,s	next word	dist	rel	d	b
It	PRP							
is	VB							
obvious	JJ							
enough	RB							
that	IN							
linguists	NN	5	1,1,1	linguists	6	synonym	0	1
				linguists	6	repetition	0	1
in_general	RB							
have	VBP							
been	VB							
less	RB							
successful	JJ							
in	IN							
coping_with	VB							
tone_systems	NN	4	2,1,1	tone_systems	6	synonym	0	1
				tone_systems	6	repetition	0	1
than	IN							
with	IN							
consonants	NN							
or	CC							
vowels	NN							

Figure 4: Tie view on annotated text

3 Exploring lexical cohesion

On the basis of the annotated data, we have generated some statistics concerning the average **chain lengths** (in no. of sentences/words participating in a chain), according to register, of both all the chains and the dominant (i.e., the longest) chains and the distribution of **types of lexical cohesion** (repetition, synonymy, hyponymy, etc.) according to register.

As will be seen, the dominant chains in a text give a good indication of a text's topic; also, the distribution of types of lexical cohesion turns out to be a possible measure for discriminating between registers.

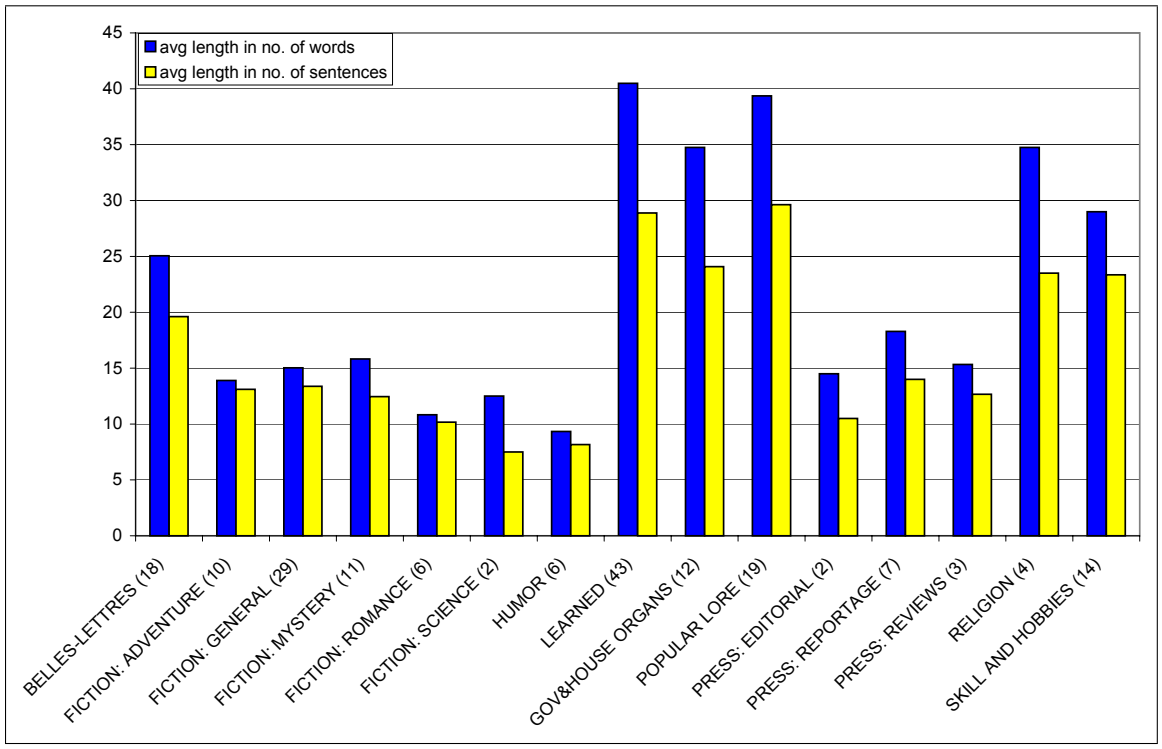


Figure 5: Average length of dominant chains by register

3.1 Chain Length

Comparing registers, the average length of lexical chains does not show substantial differences at a first glance. Most registers average between 3 and 4.5 in terms of the number of words participating in a chain and between 3 and 4 in terms of the number of sentences a chain stretches over. This means that the texts in the corpus are similarly cohesive.

However, when we compare the average length of the dominant chains across registers (i.e., the longest chains), two groups of registers stand out (cf. Figure 5): texts from the registers of LEARNED, GOVERNMENT & HOUSE ORGANS and RELIGION have relatively long dominant lexical chains and texts from PRESS and FICTION have relatively short dominant lexical chains. For example, the average length of the dominant chains in LEARNED is 40, in FIC-

TION:GENERAL it is only 15.[3]

When we look at the concrete words that make up the dominant chains, we can observe that they are good indicators of the topic of a text.[4] Short chains (with few participating words) have a different function in that they "glue" a text together locally. For example in text j34 from LEARNED (see also Figure 3), the dominant chains are built around *tone* and *phonology/morphophonemics* — this places the text in the area of linguistics, in particular phonology, and it gives us the topic of the text, which is tone. The shorter chains in this text are built around, for example, groups of words such as *explanation, theory, hypothesis, assumption* or *analysis, investigation*. One hypothesis that could be derived from such observations for this particular register is that the dominant chains are built around the register-specific vocabulary and shorter chains around the "general" vocabulary (cf. also Hoey, 1991). This hypothesis would need to be tested on more data than we have available here, however, and require a proper definition of what register-specific vocabulary means.

3.2 Types of Lexical Cohesion

Among the different types of cohesion (repetition, synonymy, hyponymy/ hypernymy, meronymy/holonymy), the most frequent means employed throughout the corpus is repetition co-occurring with synonymy with over 50% (see Figure 6, rightmost bar).

However, contrasting the different registers, there are differences in the distribution of repetition, hypernymy+(co)hyponymy and meronymy. Texts from LEARNED, RELIGION, and PRESS exhibit a higher frequency of hypernymy plus

[3] For all the data discussed here, tests for significance would have to be carried out, of course. For the time being, we conceive of the analyses reported on as purely exploratory.

[4] This observation conforms to the findings of e.g., Barzilay and Elhadad (1997), who use the dominant chains as a basis for summarization. Also, the words found in dominant chains usually have high inverse document frequency, a measure used in information retrieval.

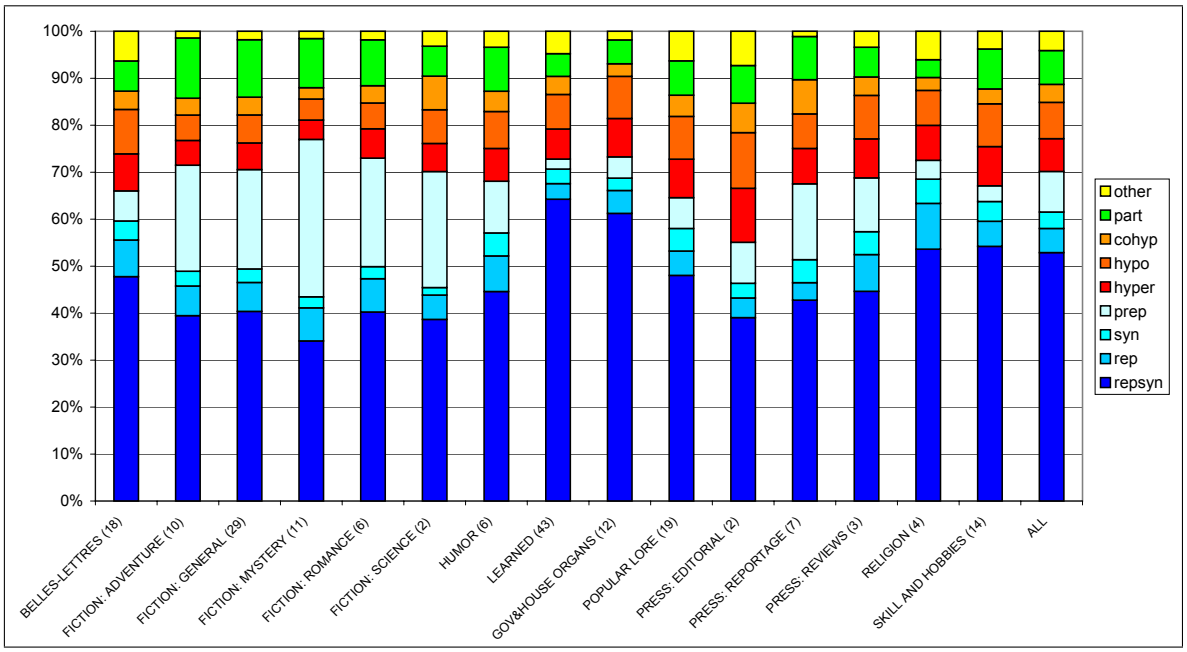


Figure 6: Types of lexical cohesion by register

(co)hyponymy than texts from FICTION. Interestingly, LEARNED and RELIGION also have the longest lexical chains relative to other registers (cf. Section 3.1). This does not come as a total surprise, however: We would expect texts from a factual genre, such as academic articles as they are included in the LEARNED register, to exhibit a strong topic continuity, whereas texts from the narrative genre, as the ones contained in the FICTION registers, can be expected to include topic shifts.

Coming back to repetition, in the LEARNED register, there is a high frequency of repetition co-occurring with synonymy, whereas in the FICTION registers repetition occurs significantly less frequently, and there is a larger amount of repetition without synonymy. This can be cautiously interpreted as follows: Texts from LEARNED try to be as unambiguous as possible, using vocabulary consistently in terms of word senses, whereas FICTION texts may actually play with ambiguity and try to be more varied in terms of vocabulary.

Finally, in the FICTION registers we encounter a substantial amount of *proper noun repetition*, which is very rare in the LEARNED register. FICTION registers also exhibit a higher frequency of meronymy. Again, this is not surprising, since fiction texts often deal with individual people who are referred to by name, and physical things, for which meronymy is more comprehensively covered in WordNet than for abstract concepts.

3.3 Summary

In summary, the findings based on the statistics presented in this section, are the following:

- **Cohesion across registers**
  - All registers included in the corpus show roughly the same degree of cohesion (where individual texts may still vary considerably in cohesive strength).
  - In different registers, cohesion is achieved by different means.
- **Cohesive patterns across registers**
  - Repetition is the most frequently used means of cohesion across registers.
  - Apart from repetition, individual registers may have a preference for a particular type of cohesion.
- **Cohesion in individual texts**
  - The dominant lexical chains (stretching over many sentences with many words participating) indicate the topic of a text.
  - In factual texts, the dominant chains tend to be made up of register-specific vocabulary.

4 Summary and conclusions

As the interest in richly annotated corpora is growing, so is the need for tools supporting annotation and exploration of multi-layer corpora. In particular, recently there is an increasing interest in the analysis of *texts*, be it for building linguistic descriptions, for testing linguistic theories or for computational applications, such as automatic summarization, text classification, information extraction or ontology building. The common interest is the interpretation of text in terms of the meaning(s) it encodes, be that rhetorical structure, information distribution or informational content.

While there is no comprehensive corpus tool available that can cater for all the linguistic needs involved in annotating text and exploring richly annotated corpus resources,[5] it has become common practice to use/build special-purpose tools that are geared to a particular annotation and/or corpus analysis task. The system we have presented in this paper is one such tool. The specific purpose it is dedicated to is to support the analysis of texts in terms of lexical cohesion. The system automatically annotates text (here: SEMCOR/Brown Corpus) in terms of lexical-cohesive ties on the basis of WordNet. The resulting annotated text can be viewed from three different perspectives, each supporting exploration of lexical-cohesive patterns from a different angle (cf. Section 2). The results of annotation can be statistically processed, simply using a standard statistics program, such as the one included in MS Excel. We have exemplified the use of some such statistics in linguistic analysis (Section 3).

With different tools taking care of different types of corpus-related tasks, special attention has to be paid to their interoperability, notably the interchange of the created corpus data. Here, the common practice now is to represent corpus resources using a standard format and data model, typically XML (see Dipper

[5] One project in this direction was the MATE project (McKelvie et al., 2001). Unfortunately, the project did not result in a scalable implementation (cf. Teich et al., 2001).

et al. (2004b) for an overview of corpus tools relying on XML). The system we have presented follows this policy, solely relying on XML and XSLT/XPath. Thus, the present research is in line with other corpus-based projects currently running or in planning, such as MULI (Baumann et al., 2004b,a), the Potsdam–Berlin SFB No. 632[6], the *Forschergruppe* at Bielefeld[7] or the project *Deutsch Diachron Digital* (Dipper et al., 2004a), only to mention a few.

In our future work, we will carry out further linguistic analyses using the data from the Brown Corpus and extend the data set to other corpora and languages (notably German). Possible applications of this research have been mentioned in passing (cf. Section 3). Notably, the data generated by our system can be used in text summarization and text classification.

Bibliography

Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL*, Madrid, Spain, 1997.

Stefan Baumann, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan Kruijff, Ivana Kruijff-Korbayova, Stella Neumann, Erich Steiner, Elke Teich, and Hans Uszkoreit. The MULI project: Annotation and analysis of information structure in German and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisboa, Portugal, 2004a.

Stefan Baumann, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan Kruijff, Ivana Kruijff-Korbayova, Stella Neumann, and Elke Teich. Multidimensional annotation of linguistic corpora for investigating information structure. In *Proceedings of HLT/NAACL Workshop 'Frontiers in corpus annotation'*. Human Language Technology (HLT) Conference/Annual Meeting of the North-American Chapter of the Association for Computational Linguistics (NAACL), 2004b.

[6] http://www.ling.uni-potsdam.de/sfb/

[7] http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/

Oliver Christ. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the 3rd Conference on Computational Lexicography and Text research (COMPLEX 94)*, pages 23–32, Budapest, Hungary, 1994a.

Oliver Christ. The IMS Corpus Workbench User Manual. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, 1994b. (`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench`).

Stefanie Dipper, Lukas Faulstich, Ulf Leser, and Anke Lüdeling. Challenges in modelling a richly annotated diachronic corpus of German. In *Proceedings of the LREC Workshop on XML-based richly annotated corpora*, Lisboa, Portugal, 2004a.

Stefanie Dipper, Michael Götze, and Manfred Stede. Simple annotation tools for complex annotation tasks: an evaluation. In *Proceedings of the LREC Workshop on XML-based richly annotated corpora*, Lisboa, Portugal, 2004b.

Peter Fankhauser and Elke Teich. Multiple perspectives on text using multiple resources: experiences with XML processing. In *Proceedings of the LREC Workshop on XML-based richly annotated corpora*, Lisboa, Portugal, 2004.

MAK Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.

Ruqaiya Hasan. Coherence and cohesive harmony. In J. Flood, editor, *Understanding Reading Comprehension*, pages 181–219. International Reading Association, Delaware, 1984.

Michael Hoey. *Patterns of lexis in text*. Oxford University Press, Oxford, 1991.

Esther König and Wolfgang Lezius. The TIGER language — a description language for syntax graphs, formal definition. Technical report, IMS, Universität Stuttgart, Germany, 2003. (`http://www.tigersearch.de`).

Wolfgang Lezius and Esther König. Towards a search engine for syntactically annotated corpora. In Ernst G. Schukat-Talamazzini and Werner Zühlke, editors, *KONVENS-2000 Sprachkommunikation*, pages 113–116. VDE Verlag, Ilmenau, Germany, 2000.

David McKelvie, Amy Isard, Andreas Mengel, Morten Baun Møller, Michael Grosse, and Marion Klein. The MATE workbench — An annotation tool for XML coded speech corpora. *Speech Communication*, 33(1–2):97–112, 2001.

Michael O'Donnell. RST-Tool: An RST analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duisburg, Germany, 1997.

Oliver Plaehn and Thorsten Brants. Interactive corpus annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, 2000. Athens.

Elke Teich, Silvia Hansen, and Peter Fankhauser. Representing and querying multilayer annotated corpora. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 228–237, University of Pennsylvania, Philadelphia, 2001.

Elke Teich
Technische Universität Darmstadt
Institut für Sprach- und Literaturwissenschaft
Hochschulstr. 1 (S 1 03 / 190)
D-64289 Darmstadt
teich@linglit.tu-darmstadt.de
http://www.ifs.tu-darmstadt.de/linglit_teich/

Peter Fankhauser
Fraunhofer IPSI
Divison I-Info
Dolivostr. 15
D-64293 Darmstadt
fankhaus@ipsi.fraunhofer.de
http://www.ipsi.fraunhofer.de/~fankhaus/