Anke Lüdeling/Amir Zeldes

Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics

Abstract

Digital corpora are used as a data source in corpus linguistics, literary computing and computational linguistics. Although differences in these disciplines dictate different kinds of work with corpora, many of their respective methods either are applied or could be applicable in the other disciplines. With the recent emergence of richly annotated multi-level and multi-purpose corpora in mind, we review differences and similarities in research questions, corpus resources and their qualitative and quantitative exploitation in the three disciplines, along with suggestions for further development and mutual enrichment.

1. Introduction

It all started with Roberto Busa's famous *Thomas Aquinas corpus*[1] from the late 1940s, which is claimed by at least corpus linguistics and literary computing[2] as a starting point for their disciplines[3], but is equally at the basis of the corpora in use in contemporary computational linguistics.[4]

---

1 Busa (1974, 1980).

2 A few words on terminology: we use the term ›literary computing‹ to denote computational approaches to the study of literature (similar to the German *Computerphilologie*, see Jannidis [2007]). The terms ›humanities computing‹ or ›digital humanities‹, while sometimes used in a similar way (see for example Zampolli [2001]), very often encompass all aspects of the use of computers in the humanities (see for example articles in Schreibman et al. 2004). The scope of the term ›computational linguistics‹ will be limited in this discussion to refer to natural language processing, which is its most relevant subfield in the context of corpora.

3 See for example McEnery/Wilson (2001) and Hockey (2004).

4 Most histories of computational linguistics claim as a starting point the interest in machine translation and the development of automata theory in the 1950s and 1960s (see for instance Menzel [2004], Jurafsky/Martin [2000], Dipper, forthcoming). Roberto Busa's work is nevertheless often acknowledged (Bátori [1989], Jones/Sondrup [1989]).

Between 1949 and today, all three of these disciplines have formed and developed. All of them use corpora or electronic versions of texts both qualitatively and quantitatively, but very often they seem to be unaware of each other's work.[5] At a time when multi-level corpus architectures, XML-based standards and standoff annotation allow unprecedented expressivity and coexistence of multipurpose and even conflicting annotations, we feel it is appropriate to review the role of corpora in different fields, the ways in which they can learn from each other and exploit the same or similar resources, as well as harness the latest advances for their own respective uses.

In this article we therefore want to explore the similarities and differences between the three disciplines' approaches to corpora and argue that new corpus architectures and distributed computing might help the disciplines to come together again. Besides large corpora, we will pay special attention to small corpora that are difficult to acquire and might need (partly) manual annotation. We will however concentrate on text corpora and will not deal with spoken corpora and multi-modal corpora.

There are, of course, many articles which deal with the history and comparison of two of the three disciplines[6], and the variety of techniques and methodologies which they could (but often do not) learn from each other. However by focusing on corpora, the common resource at their hearts, we target the element that can be most readily used to enrich one discipline through the other. The differences and similarities in corpus use that we will be dealing with can manifest themselves in three areas: (a) research questions, (b) resources, and (c) exploitation. We begin in the next section with the topic of research questions and goals, since these determine the choice of both resources and methods of exploitation. The discussion of resource types and characteristics in the following section deals separately with corpus design (that is the choice of ma-

---

[5] This has to be qualified somewhat. The fact that neither the *Oxford Handbook of Computational Linguistics* Mitkov (2003), the forthcoming Handbook on Corpus Linguistics Lüdeling/Kytö, (forthcoming) nor the recent issues of the major journals in these fields contain any articles on literary computing shows that computational and corpus linguists tend to overlook work in this area. On the other hand, the fact that a number of corpus linguistics and computational linguistics articles have appeared in literary computing journals and handbooks (see for example Schreibman et al. (2004) and recent issues of the *Jahrbuch für Computerphilologie* or *Digital Humanities Processing*) seems to show that literary computing is more open towards corpus linguistics and computational linguistics approaches.

[6] See for example Zampolli (2001), Hockey (2003), Hajič (2004), Hockey (2004), Dipper (forthcoming).

terials entering a corpus) and corpus annotation and architecture (that is how this raw data is encoded and enriched). The final section on exploitation methods will focus on the dichotomy of qualitative and quantitative approaches to corpus use, and how they can be combined to maximize the benefits of working with corpora in all fields.

2. Research questions and goals

In this section we will sketch some example research questions in corpus linguistics, literary computing, and computational linguistics and discuss the status of corpus data as an empirical basis in each of these fields. By its nature as a linguistic methodology, *corpus linguistics* is concerned with the study of language systems, and not with individual texts, which form instances of the output of those systems. Corpus linguistic research questions therefore tend to concentrate on properties of the language system, with the goal being to substantiate or disprove theories about these properties by using text (be it written or spoken) as evidence. There are areas in linguistics which have traditionally – before electronic corpora were possible – relied on textual data more or less exclusively, such as lexicography, historical linguistics, and even traditional grammar writing, while others, such as sociolinguistics and language acquisition, have relied on textual evidence next to questionnaire data, elicited data, and psycholinguistic findings. Corpus data can thus either be used as the only kind of data, or as one kind of data among others.[7]

Corpus findings can be interesting in themselves but often they are integrated into a larger theory. A good example of a research question that is forced to rely on corpus data but is grounded in a linguistic framework

[7] Tognini Bonelli (2001) proposes a distinction between corpus-driven and corpus-based approaches (compare Xiao, forthcoming). Corpus-based approaches essentially take corpora as corroborative evidence for existing theories reached by other means (for example introspection, but also other empirical means such as psycholinguistic experiments), or else as a source of counterexamples for such theories. Corpus-driven studies, by contrast, attempt to approach the data with as few preconceptions as possible, ideally deriving categorizations directly from the data. We are, however, skeptical about corpus-driven research and argue that it is not possible to do any kind of research without some previous classification; even the splitting of a text into minimal units – tokenization – requires linguistic decisions (see Lüdeling [2007]; on tokenization see Schmid, [forthcoming]).

is the study of Kytö and Romaine[8], which examines the distribution and diachronic development of inflectional, periphrastic and double adjectival degree marking in English (for example forms like *easier*, *more easy* and *more easier* respectively) in two diachronically disparate corpora. The research question is oriented towards an existing theoretical framework in the sense that it adheres to the variationist theory of language change. The findings are therefore not only interesting in themselves but can be used as a building block in a larger theory. An example of a study that uses corpus data together with other data is the study of morphological productivity by Baayen and many others[9] which builds on work in theoretical morphology and uses corpus figures to model productivity and make predictions about the behavior of a given morphological process. The corpus findings are then integrated with psycholinguistic evidence for a cognitive theory of productivity in the mental lexicon.

Corpus data is often used to model complex quantitative dependencies that cannot be found in any other way. For example, using such methods as multivariate analysis it is possible to consider and compare the significance of multiple factors represented by overlapping or complementary annotation schemes. In one study[10], Gries analyzes different postulated factors that may be responsible for positioning English phrasal verbs before or around their objects (for example *pick up a book* versus *pick a book up*), ranking over a dozen significant factors. By ana-lyzing annotated corpus data, a prediction accuracy on phrasal verb construction choice of upwards of 84% is reached in this study. Corpus linguistics is thus not limited to verifying or falsifying theories, but can investigate the interaction of factors in the data and give predictions with quantifiable accuracy.

In recent years the role of corpus evidence has been discussed anew in theoretical linguistics and corpus linguistics.[11] Generative linguistics, which for a long time only accepted grammaticality judgments (and sometimes psycholinguistic findings) as evidence, has started to use corpus data in these paradigms for some questions as well.[12]

---

8 Kytö/Romaine (1997).

9 For an overview see Baayen (forthcoming).

10 Gries (2001).

11 See for example the articles in Bod et al. (2003) or in Reis/Kepser (2005).

12 Examples are Featherston (2005) who uses frequency data to discuss graded grammaticality or Meurers/Müller, forthcoming, who use corpus data qualitatively to study unclear syntactic phenomena.

Unlike corpus linguistics, *literary computing* concentrates on particular texts, and is therefore centered on properties and interpretations of the data in itself, and not on making predictions outside of the corpus for new input. However, this does not mean that the researcher is limited to the contents of a particular edition or manuscript that he or she has available: a text in this context is understood to mean the content of a particular work as it was formulated or used in a particular context in time and space. For historical works, this text often does not exist in its entirety in any one manuscript, but must be abstracted from multiple witnesses. It is thus possible to speak of different diatopically or diachronically disparate variants of a text. Many historical projects in literary computing therefore begin by applying the methods of stemmatics, pioneered by Karl Lachmann, in order to reconstruct the contents of each version, going back to the earliest possible text by a process of collation and comparison. The computer makes such reconstructions (or stemmata) easier than ever before, with programs such as *Collate*[13] and *TUSTEP* [1] (*Tübinger System von Textverarbeitungs-Programmen*) easing the work of comparing (digitized) manuscripts and producing a critical apparatus of disagreements between witnesses automatically. Bakker, for example, uses *Collate* on multiple digitized manuscripts of the Old Church Slavonic New Testament in an attempt to reconstruct the original Slavic translation of the New Testament prepared by Sts. Cyril and Methodius in the 9th century.[14] For a more recent example using phylogenetic methods, see the papers in Macé et al.[15]

The first research question in literary computing may thus often be »what is the text?«, and this question can also be relevant to corpus linguistics, albeit usually not as a research question in itself. This question is also ultimately related to one of the most crucial practical questions in literary studies in general and in philology in particular: »how should the text be presented to the reader?« Here, the advent of literary computing has had perhaps the most revolutionary effect on the goals of the traditional philologist: it has made the visualization of ambiguous text possible. A good example of an interactive environment for the study of textual variation can be found in the *Canterbury Tales Project* [2], where the correct presentation of its results to the reader has been an explicitly stated main goal of the work.[16] There is thus a substantial range of digital

[13] Robinson (1994).
[14] Bakker (1996).
[15] Macé et al. (2006).
[16] Robinson (2003).

editions, from ones that simply mean to present a text statically, often oriented towards qualitative study of the text similar to that possible using a printed copy, to ones offering dynamic interactivity, with no one view of the text being predetermined by the editor, since a particular view may place limits on the research questions that can be addressed.

A good critical digital edition is therefore not only concerned with producing an authoritative text, which is not always possible, but also with its usability as the basis for a variety of research questions, possibly involving only small parts of texts. For example, the *Canterbury Tales Project's* CD-ROM edition was used by Solopova[17] and Kennedy[18] to discuss the controversial *Wife of Bath's Prologue*, a contested section missing from most manuscripts of the text but considered by some to originate in an unfinished draft by Chaucer. Importantly in such widely used editions, and similarly to corpus linguistics, the use of an agreed upon data base as a source of evidence facilitates the comparison of different studies, and ensures maximal reusability and reproducibility, sparing researchers all too common doubled work efforts.

In some cases, especially if too few or even only one manuscript of a text is available, no attempt is made to reconstruct stemmata. For more modern works, where an authoritative print edition forms the basis of the studied text, there is also no need to do so. But whether or not a text is abstracted from the data or is simply available for study, the digital edition can be only the first step in deeper investigations into a text and its context. Greater value can be attained in texts that are also annotated for metadata (more on which in the next section) – even physical properties of codices, or the logical structural divisions of line and page breaks, or metrical and typographical information can be retained in corpora, possibly alongside digitized, aligned facsimiles. Such high-quality corpora provide unprecedented access to original documents for researchers worldwide using nothing more than a web browser, and consequently allow the diverse traditional research questions to benefit from the digital corpus. But the ultimate goal is to create resources that not only allow any traditional research to be carried out on-line, but also to open up new directions and research questions, especially in quantitative studies. Semino and Short, for example, use a corpus of English prose, newspapers and biographical material to study the different ways in which speech and thoughts are presented in writing, analyzing data both quali-

[17] Solopova (1997).
[18] Kennedy (1997).

tatively and quantitatively.[19] Another example is the *Charikleia* project, which proposes studying the historical emergence of the German novel and the narratological development of this genre using a corpus of German literature from 1500 to 1900 (for more on quantitative methods see the section on quantitative corpus exploitation below).[20]

Corpus linguistics and literary computing use computational methods to study underlying data. *Computational linguistics* by contrast harnesses linguistic resources as a means to an end, usually in order to create systems that can cope with unseen, but similar linguistic input, and process it. This is not to say that computational linguistics is detached from linguistic theory – on the contrary, many computational linguists attempt to formalize and implement linguistic theories such as LFG (Lexical Functional Grammar) or HPSG (Head-Phrase Structure Grammar) in parsers, and have to address such issues as the computational complexity of these models.[21] However beyond theoretical questions such as what computers can or cannot do linguistically (for example whether or not a computer can really »speak«, or pass the Turing Test, fooling a human into thinking they are engaged in a conversation with another human), computational linguistics can be thought of as more goal-oriented than research question-oriented in the sense of the other two domains discussed here.[22]

A typical goal of computational linguistics can be found in what is perhaps its defining task, and certainly the original motivation driving the development of the field in its early days: machine translation. Its task is simply put to take input in a source language and output its translation in a destination language.[23] However since at least in corpus-based systems the probabilities of different possible translations are calculated based on examples from a parallel bilingual training corpus, the translation is only as good as the corpus it is based on, or more exactly, its quality depends on how closely the corpus resembles future input (for example in domain, register et cetera). Computational linguistics tasks are typically evaluated in terms of precision and recall (that is how much of the output is incorrect, and how much of the desired output was achieved), meaning that they rely on some sort of ›gold standard‹, often a manually annotated

---

[19] Semino/Short (2004).

[20] Jannidis et al. (2006).

[21] Compare Dipper (forthcoming).

[22] See the articles in Mitkov (2003); for the role of corpora in computational linguistics see Dipper, forthcoming.

[23] For an overview of different approaches and some background, see Nirenburg et al. (2003) and Somers, forthcoming.

output or a set of guidelines for humans to produce the desired output,[24] which may also influence a task's formulation.

Other computational linguistics tasks are intimately involved in the preparation of linguistic corpora and databases for the digital humanities in general, such as lemmatization and orthographic normalization or fuzzy search[25], part-of-speech tagging, syntactic parsing, or even high-level tasks such as anaphor and co-reference resolution or named entity recognition.[26] Many approaches to these tasks rely on sample corpora for training statistical models, meaning for example that a normalized edition of a small text or part of a text may be needed in order to create one of a larger text automatically or semi-automatically.

Often, however, computational linguistics is less preoccupied with annotating data that researchers, or humans in general, will be interested in explicitly searching for, but rather in resolving the fuzziness that exists in users and their needs themselves. For example information retrieval, a domain in computational linguistics dealing with searching for and retrieving all and only the data that a user is interested in from a collection of documents, is concerned with bridging the gap between an explicit but inaccurate query, and the possibly more accurate but inexplicit intent of the user. In order to fulfill this goal, user input can be expanded by using lexical semantic resources such as formal ontologies, which provide alternative ways in which the user's intent might match actual text (for example to search for *poodle* too when *dog* is input). At the same time, the document set being searched, which is in many ways similar to a corpus, despite its non-linguistic design and motivation, is enriched with relevant semantic tagging, such as tags denoting whether entities are human, animate, edible, sub-parts of other entities and so on. It goes without saying that the development of many such resources and their testing also involve large corpora, which, as we shall see, have different properties than linguistic and literary ones.

In a sense, the goals of computational linguistics are thus partly determined by the needs of other disciplines, which require taggers and parsers, and partly by commercial interests, which are more involved in the development of search engines and machine translation systems. At

---

[24] An example of a machine translation evaluation measure is the BLEU score (Papineni et al. [2002]). For criticism see Callison-Burch et al. (2006).

[25] The latter two are especially relevant for achieving searchability of non-standard and historical texts, see Pilz et al., forthcoming.

[26] See for example Manning/Schütze (1999), Jurafsky/Martin (2000), and the articles in Mitkov (2003).

the same time, computational linguistic methods are constantly being fed by the resources and theories that other disciplines produce. In the next section we discuss some of these resources in greater depth.

3. Resource types

As text-based disciplines, corpus linguistics, literary computing, and those areas of computational linguistics which are concerned with the processing of natural language texts, all make use of digital corpora. However, there are several key differences in the resources each of these disciplines uses, both from a practical and a theoretical point of view. The differences pertain to corpus design (that is the question »what goes into the corpus?«), which is discussed in the next subsection, and corpus annotation and architecture, which are addressed in the following one.

3.1 Corpus design

The design of a corpus is first and foremost dependent on the research questions it is meant to answer. Corpora range from very specific to opportunistic[27]. Which researchers use what kind of corpora is a matter of degree rather than one of principle: while all disciplines use specific corpora (with corpora that contain one text of one author, which are more common in literary computing, being the extreme), probably no philologists, only very few corpus linguists but many computational linguists use large opportunistic corpora.[28]

---

[27] That is everything one can get, for example corpora harvested from the Web, see the papers in Baroni/Bernardini (2006b) or Bergh/Zanchetta (forthcoming).

[28] Admittedly, the considerations behind selecting a text for any study may be partly dictated by availability – a text already available digitally e.g. from *Project Gutenberg* [3] or from the *German digital library zeno.org* [4] is often more attractive than one requiring digitization. Still, the choice of text is liable to be much more particular in literary computing. Many corpus linguists, on the other hand, would perhaps not even call opportunistic collections ›corpora‹ since often the fact that the collection strategy is dependent on given research questions and goals is part of the definition of ›corpus‹. Compare for example the definition given by the Expert Advisory Group on Language Engineering Standards: »A *corpus* is a collection of pieces of language that are selected and

Since corpus linguistics, as already mentioned, is concerned with the study of abstracted language systems, and not one particular text or another, one of its primary concerns is obtaining resources which are ›representative‹ of the language in question. Representativeness is ultimately impossible to achieve for an infinite body of language (such as any living variety of a language), where the distributions of texts according to certain parameters cannot be determined and consequently cannot be mapped onto the corpus design. The term ›representative‹ must therefore be used with caution. Although representativeness is often equated with an attempt to get corpora which are as large as possible, in order to cover more of the language, corpora attempting to be representative should more importantly focus on giving a balanced sample of the possible variability in the language population being investigated.[29] In the case of very large, and especially (national) reference corpora, this often means incorporating both written and spoken (usually transcribed) data, as well as a classification of texts according to such factors as genre, register, dialect et cetera. The corpus can then be designed to include controlled amounts of material from each category in order to be ›representative‹. The BNC [6], for example, contains 90% written and 10% transcribed spoken British English. Written texts are classified according to the time they were composed, subject matter and publication medium (novels, journals et cetera), while spoken data covers material from speakers of diverse age, locations, social class and sex, as well as material from formal speech such as radio broadcasts. However, beyond the ratio of the different text types, the identity of the texts is supposed to play no role in the usage of the corpus, which is seen as a sample of the infinite potential texts that could or do occur in modern British English.

Linguists often study corpora of ›non-standard‹ varieties such as *dialect corpora*[30], corpora of specific social groups[31] or certain registers[32]. Many text types are not available in these varieties, and standardization is often problematic, since some of these varieties are not usually written – such corpora are therefore often opportunistic and small. Linguists use such corpora to study lexical, morphological, syntactic and many other prop-

ordered according to explicit linguistic criteria in order to be used as a sample of the language« [5].

[29] As pointed out for example by Biber (1993: 243).

[30] Hollmann/Siewierska (2003), Anderwald/Szmrecsanyi (forthcoming).

[31] For example London teenagers in the COLT corpus, Haslerud/Stenström (1995).

[32] On the multidimensional model for the description of registers see for example Biber/Conrad/Reppen (1998).

erties of the given variety – again, the specific text is not important as long as the corpus represents the variety.

Literary and philological corpora, by contrast, are in general unique – that is, they are not interchangeable with other, comparable corpora in the same language. In most cases, a philological corpus is closed, meaning no new texts are expected to be added to the corpus, though corpora of living authors can of course grow, and occasional discoveries in corpora of historical authors are also possible. *Historical corpora* are not only closed, but their contents are often dictated by external factors, which therefore determine the corpus design. This is perhaps less problematic for literary computing than for historical linguistics, since as the selection of available texts becomes smaller the further back one looks, corpora become less and less linguistically representative, and often the available data is less than ideal. In practice, the same historical corpus can be (but often is not) used by linguists and literary scholars. Donhauser[33] describes a corpus of a 9th century interlinear Latin and Old High German translation of Tatian that is used for the study of information structure in Old High German. While the translated German text in this corpus often adheres to the word order from its interlinear Latin original, discrepancies between the texts can be used to draw conclusions on the development of Germanic word order.[34] Note that while this is not very different from a description of a particular text's language in literary computing, the aim is to make statements about Old High German in general (ideally supported by further comparative, typological or other types of evidence from outside the corpus), and not about the Tatian text itself.

All this does not mean that historical corpora must be small – good counterexamples can be found in corpora in the Classics, which although essentially closed, are in fact massive. For example, the *University of California Irvine* hosts the *Thesaurus Linguae Graecae* [7], offering 99 million words of texts including Greek authors ranging between Homer and the fall of the Byzantine Empire. *Tufts University's Perseus Project* hosts a freely available Classics collection [8] containing over 7.8 million words of Greek and over 5.2 million words of Latin. It also includes close to 39 million words of English translations and reference works for scholarly use, hyperlinked from and to the source texts, as well as advanced graphical resources such as maps and photographed archeology collections. However being a philological resource does not contradict offering a wide variety of tools that are of interest to linguists: *Perseus* also contains

[33] Donhauser (2007).
[34] See Petrova (2006); Hinterhölzl et al. (2005).

hyperlinked morphological analysis tools allowing users to analyze and lemmatize inflected word forms, as well as to access corresponding dictionary entries in multiple digitized resources. Quantitative corpus linguistic studies are also supported with automatic usage statistics for words in different authors or text types. However importantly for linguistic use, these corpora make no attempt at being balanced; rather, they try to be exhaustive. We can find out how often Aeschylus uses a certain word, or how much more frequent a word is in Homer than in Hesiod, but estimating how frequent a word was in Classical Greek in general, or in the dialect or idiolect of even one author, is methodologically compromised by imbalances in corpus design that depend on the coincidences that preserved one text but lost another, or in the case of collections that have only selectively digitized some of the available texts, the content-based preferences of the editors.[35] In such cases, researchers may need to hand-craft appropriate subcorpora from the material available.

Computational linguistics, by contrast, tends to prefer maximally extensive corpora. Many applications rely on statistics, so it is often necessary to have large corpora to achieve both statistical validity in theory and adequate performance in practice. This leads to the fact that many corpora in computational linguistics are mostly of contemporary text, which is already (and cheaply) available electronically. The texts can be literary, but are often more restricted to newspaper language. Commonly applications are developed using only a large part of such a corpus for training, leaving a smaller part as data unseen by the system for testing performance. Another type of corpus which figures prominently in computational linguistics is the *parallel corpus*, which is used especially in statistical machine translation systems, but also in computational lexicography for parallel multilingual terminology extraction (that is creating dictionaries for specialized technical domains) and the preparation of multilingual documentation. One of the largest and most frequently used corpora in this area is the *EuroParl corpus*[36], which contains proceedings of the European Parliament in 11 European languages, with between 26 and 44 million words of sentence-aligned text for each language.

---

[35] This is especially pertinent for historical corpora of more recent periods, which are forced to selectively digitize samples from each period on account of the vast amounts of material, for example the corpus proposed in Jannidis et al. (2006) mentioned above.

[36] Koehn (2005).

Parallel corpora are also of interest for comparative linguistics and the study of translated language, as the following examples show. The Regensburg *Parallel Corpus*[37], for example, currently offers 31 parallel pos-tagged texts, available in any number of 10 languages (Slavic languages, English, and German), and totaling some 9.4 million tokens. Users can query subcorpora to find and quantify occurrences of part-of-speech tags, lemmas or word forms in one language, depending on whether or not another part-of-speech, lemma or word form is found in the available parallel texts (for example to find the frequency of German *Haus* translating English *house* versus English *home*). Using regular expressions to define variable length token chains, it is even possible to investigate the frequency of certain syntactic phenomena (for example which elements in the article-less Slavic languages co-vary with the use of English definite versus indefinite noun phrases). Zeldes[38] demonstrates how a parallel historical Bible corpus can be used to study syntactic and lexical change, by automatically extracting correspondences between lemmas, morphological suffixes and recurrent token sequences in texts from different stages of the Polish language. Another parallel corpus used for a study of translated language in itself (sometimes called »translationese«, a term due to Martin Gellerstam) can be found in the work of Baroni and Bernardini.[39] The study used support vector machines (SVMs), a machine learning technique, on a corpus of some 2 million words of original Italian journal text, and over 877,000 words of articles translated into Italian in the same journal, from several source languages. The authors report that the machine learning algorithm trained on the corpus was able to distinguish translated language from original Italian, on average with high accuracy (86.7%), outperforming the judgment of even human translators, based solely on the corpus example data.[40]

---

[37] Von Waldenfels (2006).

[38] Zeldes (2007).

[39] Baroni/Bernardini (2006a).

[40] For more on parallel corpora in contrastive and translation studies see Johansson (2007).

3.2 Annotation and corpus architecture

While corpora in the different disciplines may vary considerably with respect to contents, all three share the need for and use of metadata.[41] Metadata can be classified in many ways; one traditional classification distinguishes between header information (information about the whole text), structural information (information placed between tokens to mark the graphical or logical structure of the text) and positional information (information about the smallest units, the tokens). The levels and types of metadata differ markedly between the disciplines. Header information gives users information about the corpus and the texts in it on a macroscopic level, providing such details as the time, place and language of composition, as well as the authorship, or for historical texts often the scribe or copyist who prepared a manuscript. Other kinds of metadata describe the corpus coding itself, for example the annotation layers available in the corpus or the symbols used in the text, or the corpus structure, such as divisions into chapters, paragraphs et cetera. There are many standards available for encoding corpora and their metadata, with no consensus having emerged yet. Structural divisions of texts are often captured in TEI XML, or its simplified version TEI Lite, which are hierarchical XML specifications created by the Text Encoding Initiative [9]. This format is especially common in literary and historical corpora, since it offers many options for the description of logical and also graphical elements that may become fairly complex in attempts to faithfully describe manuscript material. TEI also offers its own extensive format for header data to describe corpora, and annotations to describe the text, with specialized fields used, for example, to mark up rhyming or meter in verse texts, information on stage directions and the cast in performance texts, and much more. Other formats concentrate on metadata used to identify linguistic characteristics of a text, and are most useful for linguists wanting to establish what kind of language a corpus is a sample of.

High-quality closed corpora with multiple layers of rich annotation are probably more typical for philological resources, but there are also examples of richly annotated linguistic corpora, such as the above-

[41] Notwithstanding approaches that avoid metadata since it is always an interpretation of the text, and therefore perhaps controversial. Except in a few very special cases, such as perhaps the segmentation algorithm in Golcher (2006), we believe that metadata is always useful (compare the data-driven approach mentioned in footnote 4 and the criticism mentioned there).

mentioned Tatian corpus.[42] Beyond ordinary grammatical annotations, the corpus contains detailed annotations regarding information structure in the text, including topical and focal elements, givenness, definiteness and more, with the goal of linguistically studying discrepancies in word order between the Latin and Old High German texts. This type of research requires specialized annotation schemes that would not be available in a general purpose literary corpus of the same texts, and at the same time tools for quantitative analysis on, for example, how often we find verb first, second, or last position in the corpus, depending on syntactic or information structural considerations. Other examples of corpora with rich annotation are the learner corpus *Falko*[43] which has a multi-layer error annotation, or the *Potsdam Commentary Corpus* (PCC)[44], which is annotated with rhetorical structure[45], information structure and coreference, alongside syntactic annotation.

Some richly annotated schemes also allow competing annotations for the same metadata field. The freely available version of the *Europarl corpus* [10] is an interesting resource in this context, since it contains multiple competing part-of-speech annotations for some languages, in the form of tags assigned by several taggers to the same text (up to six tags for each token in the English version). This can be useful, since different tagging schemes may be more suitable for different applications. This contrasts however with a typical linguistic point of view, in which a certain tagging scheme is selected for more or less well thought out theoretical reasons, and treated (often too lightheartedly) as a ground truth for further study (for example allowing statements on the absolute or relative frequency of certain grammatical categories, et cetera).

No matter what the annotation categories themselves or their values, all three disciplines face the same issues of storing and querying corpus data with diverse multi-layer annotations. In recent years, therefore, much effort has been spent on developing multi-layer architectures that separate data and annotation, instead of committing to one inline annotation layer, which is difficult to expand and modify. Standoff architectures[46], which are designed to allow separate annotation files to refer to corpus data, are opening up new ways in which different annotations

[42] Donhauser (2007).
[43] Lüdeling et al. (2008).
[44] Stede (2004).
[45] Based on RST, Mann/Thompson (1987).
[46] Thompson/McKelvie (1997); compare Dipper (2005).

with conflicting hierarchies, different categories and ambiguous values can serve multiple disciplines more adequately at the same time.

4. Exploitation

As already mentioned, all three disciplines exploit corpora both qualitatively and quantitatively. The differences one finds are of degree and not of principle. Computational linguistics in recent years has almost exclusively used statistical methods (as can be seen for example in papers featured at the conferences of the Association for Computational Linguistics), whereas many scholars in literary computing and corpus linguists concentrate more on qualitative methods.

One interesting difference between corpus linguistics and literary computing stems from the fact that scholars in literary computing see themselves mainly as humanities scholars whereas at least some corpus linguists see themselves as natural scientists and conduct their research to meet certain standards of experimental design, reproducibility of results et cetera.[47] Since the same techniques can be, and increasingly often are used in multiple domains, the next sections are arranged according to methodologies and not discipline by discipline.

4.1 Qualitative methods

The qualitative use of corpora in general has concentrated on the key word in context (KWIC) concordance,[48] as can be gleaned from the wide variety of concordancing tools available. KWIC concordances are essentially a list of corpus data segments matching a search criterion, surrounded by its context (that is the words before and after it). The concordance allows researchers to get an overview of the different contexts in which a target item (be it a word, a lemma, a complex annotation or syntactic construction, or any combination of these) may appear. The ultimate goals of such a search can be very varied: a linguist may be interested in finding a counter example to a theory predicting that a certain

[47] Hajič (2004), Baroni/Evert (forthcoming), Biber/Jones (forthcoming).
[48] For an overview of the development of concordancing see Jones/Sondrup (1989).

construction will not appear, while a literary scholar may try to find all mentions of certain characters or places in a novel. Computational linguists may be more interested in using such tools to find examples of constructions their systems have trouble handling, or indeed to be able to foresee if the presuppositions their systems depend on are supported by the corpus data. An example might be searching for pronouns in various constellations to determine if and how often an anaphor resolution heuristic would be correct, before one sets out to implement it.

While most search engines rely on users being able to formulate more or less complex queries in a query language, providing an appropriate query builder makes exploitation much easier for the uninitiated (though this is not meant to replace an expert interface allowing the full functionality and power of the underlying search engine). A particularly noteworthy idea on the border between corpus and computational linguistics is the *Linguist's Search Engine* [11], which allows users to input an example sentence to be parsed by an on-line parser, and have the search engine retrieve syntactically similar sentences from a corpus. This type of query could doubtless be useful for literary scholars interested in the language or style of certain authors or works, who may not be familiar with the syntactic formalism used to annotate the corpus, and might therefore find phrasing the necessary queries directly difficult or cumbersome.

Once the desired query is formulated and a concordance has been retrieved, an immediate second step is usually a classification of the results into meaningful categories. These can be a simple binary decision (is this the construction being searched for or not, for example the linguist's necessary counter example), or a more complex classification (such as semantically classifying matched adjectives into color terms, other physical properties, value judgments and so on). In this context it is often interesting to classify corpus results by their contexts. The literary scholar may want to know which characters appear when a certain term is mentioned, who mentioned it, or in what setting it was mentioned. If the element determining the classification can be defined in machine decidable terms, concordances can simply be sorted to produce the classification (for example all results for a certain adjective followed by any noun can be sorted by that noun alphabetically).

Naturally, the literary scholar is often concerned with more context than can be conveniently displayed in a KWIC concordance, which is why most literarily oriented concordance interfaces offer hyperlinking functionality between concordances and expanded context views of the corpus. The advantage of using both views in conjunction is that potentially interesting results can be reviewed easily in the plain-text concor-

dance, possibly with helpful highlighting functions and annotations, whereas a detailed view navigated to from this list can contain both more text, and representations that are more taxing to interpret, such as aligned facsimiles. A good example of this mode of operation can be found in the *Canterbury Tales Project*, which also offers special marking for variants in the collation, so that different versions of a search result can be navigated to on the fly. Although these functions have been developed largely with literary computing in mind, they are entirely applicable to corpus linguistics as well. Many linguistic domains require relatively large contexts, and many corpora correspondingly offer not only adjustable context width for concordances, but also dedicated text-length context views, which are especially appropriate for studying text-wide dependencies. The rhetorical structure annotated in the above mentioned Potsdam Commentary Corpus, for example, cannot be adequately interpreted without very large context, and often requires reading an entire text. Corpora comprised of short news stories or essays can also be studied at text level, using searches to retrieve text containing interesting phenomena. This allows researchers, for instance, to study constructions typical of the beginning or end of a text, and their dependencies on various features being found in or absent from the entire text. This means that the same corpus can be exploited by researchers in different fields, or even used to examine interdependencies between different layers (for example the effect of information structure on syntax). More and more types of annotation, often created by work-intensive manual methods, are being proliferated, for example verbal argument annotations in *PropBank*[49] and discourse annotations for connectives like *because* or *although* in the *Penn Discourse Treebank*[50]. New research methods taking advantage of such annotations simultaneously may reveal as yet unknown interactions between different linguistic levels.

The integration of scholarly works into corpora is another trend which has grown in literary computing, but has not to date been carried over to linguistic corpora. Since literary corpora often render existing editions, which may contain footnotes commenting on various aspects of the text and citing previous research, such additional data has been digitized alongside the text in some resources. *Perseus*, for example, offers linked commentary works and translations of many original texts, which often amount to much more material than the actual corpus data, and can be of immense use to users wanting to exploit the text for their own

[49] See Kingsbury/Palmer (2003).
[50] Miltsakaki et al. (2004).

research. While linguistic corpora sometimes offer connectivity with lexical resources,[51] and parallel corpora naturally contain aligned translations, in the future corpora could offer access to digitized linguistic scholarly works and commentary, either through license-based internet vendors like JSTOR [12], or through archives of freely available materials, conference proceedings et cetera. The corpus could thus become a true linguist's workbench, where he or she can not only find attestations of phenomena, but also learn what has been written about them by other researchers.

An exciting prospect in this context is the possibility of integrating Web 2.0 functionality into such commentary and linking, allowing users to tag their own analyses[52] and link search results to relevant available works, or voting (either directly or using link usage statistics) for the most relevant commentaries. Far from being distant possibilities, this type of services is already being offered for a linguistic application in *Perseus*' latest version (4.0), which allows users to vote for and rely on choices of alternative morphological analyses in word forms that are ambiguous. This creates user-based positional information (or rather a weighting of available conflicting annotations) telling us that the same form may be an accusative in a certain chapter of the *Iliad*, but nominative in another text by Herodotus, according to most users. The potential for harnessing users to develop a resource further simply by letting them exploit it and browse through it is thus limitless.

4.2 Quantitative methods

Beyond the advantages of advanced search capabilities facilitating once very time consuming qualitative research, the added value of digital corpora really lies in the possibility of quantitative analyses. Although probably used by only a minority of literary scholars accessing corpora, and certainly not by all corpus data-based linguists, basic frequency counts of word forms, lemmas et cetera have been offered by corpus interfaces and used successfully for a long time. However, manipulating quantitative data to form meaningful statements has often required the develop-

[51] This includes dictionaries, morphological analyzers, or even lexico-semantic resources such as Princeton's WordNet [13], which already more than 10 years ago (version 1.4) included a WordNet tagged version of the Brown corpus.

[52] Compare Smith et al. (2007).

ment of specialized systems, which have had less penetration as exploratory tools for larger communities. For example, Rayson et al. describe a study on key lexical items best distinguishing speakers according to gender, age and social class in the spoken part of the *British National Corpus*.[53] The study used software developed at UCREL at *Lancaster University* to rank items using chi-squared distribution values determining the significance of deviations in the frequency of items in one category versus another (for example male versus female speakers).

More recently, corpus interfaces have begun integrating advanced tools for quantitative analysis. While technically not difficult to implement, these tools immediately deliver utility of a higher order of magnitude to users who are not in a position to write scripts to manipulate raw corpora themselves. Advances have been made especially in the field of collocation extraction, that is the automatic identification of (ideally meaningful) combinations of words whose cooccurrence is statistically significant. Measures of collocability such as Log Likelihood (LL)[54], mutual information (MI)[55] and others,[56] which were originally developed in computational linguistics for tasks like signal processing, technical terminology extraction, automatic lexicon acquisition and machine translation, are now being offered within corpus interfaces. For example, the corpus of the digital German dictionary *DWDS* (*Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts* [14]), allows users to switch from concordance to collocate view for a node word, to see which other words are most significantly associated with it. Users can choose between three association measures (LL, MI and the t-test), and sort results either by one of the measures, the frequency of the collocation or the frequency of the non-node word. The interface developed by Mark Davies at *Brigham Young University* for a collection of large corpora [15] goes further in offering not only integrated collocations for individual items, but also automatic comparison of either common or mutually exclusive collocates of multiple query items. This allows users to study subtle differences in near synonyms and to find areas of semantic overlap.[57]

Both of the above interfaces also allow users to graphically compare the distribution of items across genres by computing frequency counts

---

53 Rayson et al. (1997).

54 Dunning (1993).

55 See Daille (1995) for implementations and discussion.

56 For an extensive overview of collocation measures see Evert (2005).

57 See Manning/Schütze 1999: 166–168 for an example comparing which English nouns combine with the adjective *strong* more often, and which with *powerful*.

for different subcorpora, and in the case of some of the corpora at *Brigham Young University*, also based on divisions into time periods (by decade, or century in historical data). Literary corpora have for a long time been organized around entire, often relatively small texts, and have naturally allowed quantitative results to be compared across such subcorpora. In linguistics, where large newspaper corpora are common, functionality comparing for example the frequency of items in each individual article is rarely offered. In the future both domains could benefit from more flexible abilities to define ad hoc subcorpora on the fly, based on metadata or query results (that is searching within a list of matches or saving one as a subcorpus).

Another useful quantitative tool coming from lexicographically oriented computational and corpus linguistics is the *Sketch Engine*[58], which offers a corpus-based one page summary for each word, including its most common collocates in various constructions (for example most common nouns in subject and object positions for verbs, associated prepositions, adjuncts et cetera).[59] Such functionality, especially used comparatively in conjunction with subcorpora and larger monitor corpora can reveal where texts differ semantically from each other, and from a more »average« usage as represented in the larger corpus. Integrating these tools, based on data which is essentially already there, should be a top priority for both linguistic and literary corpora, and may have considerable value in computational linguistics too as a diagnostic tool for evaluating differences in domain-specific texts.

A special area of quantitative research equally related to literary research, linguistics, and computational linguistics is statistical stylometry. One of its typical tasks is using sample texts from different authors to establish corpus-based parameters (or »discriminators«) characterizing their work, in order to identify the author of an unattributed work out of given options. Some researchers use the relative frequencies of function words, which are thought to be topic independent but characteristic of particular writers: Merriam and Matthews used a multi-layer perceptron, a neural network-based machine learning technique, to determine authorship of plays and parts of plays that may have been written by either Shakespeare or Marlowe, based on the relative frequency of ten common words such as *the*, *not* and *that*.[60] Burrows considers a much larger range

[58] Kilgarriff et al. (2004) [16].

[59] For an example comparing the behavior of the lemmas *man* and *woman* in the BNC using the Sketch Engine see Pearce (2007).

[60] Merriam/Matthews (1993).

of authors at once, using a distance measure to evaluate similarity in a collection of texts from 25 poets of the English Restoration period.[61] Other studies also use the corpus to automatically decide which words or constructions would make the best discriminators, and it is even possible to use the frequencies of all possible substrings of a text to compute a measure of its repetitiveness which is different and characteristic for different authors.[62] Stylometry can also be used to analyze the speech of individual characters in novels: DeForest and Johnson classify Jane Austen characters according to the proportion of Latinate versus Germanic words they use in their dialog and letters.[63] For more information see Oakes'[64] overview of corpus-based stylometry

5. Conclusion and outlook

In this paper we have discussed the role of corpora in linguistics, literary computing and computational linguistics. As we have shown, research questions, specific resources and the methods of their exploitation may differ considerably between these disciplines, yet they must all deal with similar and overlapping issues in corpus design and annotation, and may benefit from adapting each other's methods. A relatively new and exciting direction for the future of work within these areas, and for interdisciplinary work as well, is multi-layer annotations and architecture on the one hand, and methods of taking advantage of data from such multi-layer corpora on the other. Where in the past computational linguists may have used a linguistic corpus to create taggers and parsers, and linguists in turn used these tools on corpora digitized by the digital humanities, we are entering a stage where work using the same resource is becoming possible on both an interdisciplinary level and an interpersonal level, between researchers working separately.

The technologies available today enable multiple users to engage in independent research on and annotation of the same data. As we have seen, first applications offering so-called Web 2.0 functionality for corpora are emerging, which will allow scholars in different fields to communicate through the use of shared resources, and keep them more in-

[61] Burrows (2002).
[62] See Golcher (2007).
[63] DeForest/Johnson (2001).
[64] Oakes (forthcoming).

formed and more up to date about work relevant to the resources they are using. Offering multiple views of the same data and allowing users to develop resources further is especially relevant for data that is difficult, time-consuming, or expensive to acquire, such as historical data.[65] A wide usage of such texts by as many people as possible is therefore highly desirable. One project that explores how far this idea may go is the *TextGrid* project [17], which uses grid computing to combine resources such as corpora or lexicons and techniques such as lemmatization from many different sources. These resources, combined with the right computational and statistical tools, could give scholars not only a convenient way to continue traditional modes of work, but also to develop new and especially quantitative approaches that may not have been practicable only a few years ago.

Researchers in all text-based disciplines are finding themselves witnessing a massive process of digitization of written human knowledge.[66] Now more than ever it is up to research communities to take advantage of the resources which are becoming available, and shape them to their research needs, giving us not only three, but in fact an unlimited number of views on corpora.

Bibliography

Anderwald, Lieselotte/Benedikt Szmrecsanyi
Forthcoming Corpus Linguistics and Dialectology. In: Lüdeling/Kytö.

Baayen, Harald
Forthcoming Corpus Linguistics in Morphology: Morphological Productivity. In: Lüdeling/Kytö.

Bakker, Hette Popke Sjoerd
1996 Towards a Critical Edition of the Old Slavic New Testament. PhD Thesis, University of Amsterdam.

Baroni, Marco/Silvia Bernardini
2006a A New Approach to the Study of Translationese: Machine-learning the Difference Between Original and Translated Text. In: Literary and Linguistic Computing 21(3), p. 259–274.

2006b (eds.): WaCky! Working Papers on the Web as Corpus. Bologna: Gedit.

Baroni, Marco/Stefan Evert

[65] For a proposal on how this might look see Lüdeling et al. (2005).

[66] Compare Google's scanning thousands, or in the near future even millions of books, [18]. For an academic perspective on these developments, see Crane (2006).

Forthcoming Statistical Methods for Corpus Exploitation. In: Lüdeling/Kytö.

Bátori, István
1989 Grundprobleme der Anwendungen in der Computerlinguistik. In: István Bátori/Winfried Lenders/Wolfgang Putschke (eds.) Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications. Berlin: Walter de Gruyter, p. 481–489.

Bergh, Gunnar/Eros Zanchetta
Forthcoming Web linguistics. In: Lüdeling/Kytö.

Biber, Douglas
1993 Representativeness in Corpus Design. In: Literary and Linguistic Computing 8(4), p. 243–257.

Biber, Douglas/Susan Conrad/Randi Reppen
1998 Corpus Linguistics. Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Biber, Douglas/James K. Jones
Forthcoming Quantitative Methods in Corpus Linguistics. In: Lüdeling/Kytö.

Bod, Rens/Jennifer Hay/Stefanie Jannedy
2003 (eds.) Probabilistic Linguistics. Cambridge, MA: MIT Press.

Burrows, John
2002 ›Delta‹: A Measure of Stylistic difference and a Guide to Likely Authorship. In: Literary and Linguistic Computing 17(3), p.267–287.

Busa, Roberto
1974 Index Thomisticus. Stuttgart: Frommann-Holzboog.

1980 The Annals of Humanities Computing: The Index Thomisticus. In: Computers and the Humanities 14, p. 83–90.

Callison-Burch, Chris/Miles Osborne/Philipp Koehn
2006 Re-evaluating the Role of Bleu in Machine Translation Research. In: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL). Trento, Italy, p. 249–256.

Crane, Gregory
2006 What do you do with a Million Books? In: D-Lib Magazine 12(3).

Daille, Béatrice
1995 Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. Unit for Computer Research on the English Language Technical Papers 5, Lancaster University.

DeForest, Mary/Eric Johnson
2001 The Density of Latinate Words in the Speeches of Jane Austen's Characters. In: Literary and Linguistic Computing 16(4), p. 389–401.

Dipper, Stefanie
2005 XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: Proceedings of Berliner XML Tage 2005 (BXML 2005). Berlin, p. 39–50.

Dipper, Stefanie
Forthcoming Theory-driven and Corpus-driven Computational Linguistics and the Use of Corpora. In: Lüdeling/Kytö.

Donhauser, Karin
2007 Zur informationsstrukturellen Annotation sprachhistorischer Texte. In: Zeitschrift für Sprache und Datenverarbeitung 31 (1-2), p. 39-45.

Dunning, Ted
1993 Accurate Methods for the Statistics of Surprise and Coincidence. In: Computational Linguistics 19(1), p. 61–74.

Evert, Stefan
2005 The Statistics of Word Cooccurrences: Word Pairs and Collocations. Doctoral thesis, University of Stuttgart.

Featherston, Sam
2005 The Decathlon Model: Design Features for an Empirical Syntax. In: Marga Reis/Stephan Kepser (eds.): Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives. Berlin: Mouton de Gruyter.

Golcher, Felix
2006 Statistical text segmentation with partial structure analysis. In Proceedings of KONVENS 2006, Konstanz, p. 44–51.

Golcher, Felix
2007 A New Text Statistical Measure and its Application to Stylometry. In: Proceedings of Corpus Linguistics 2007, Birmingham.

Gries, Stefan Th.
2001 Multifactorial Analysis of Syntactic Variation: Particle Movement Revisited«. In: Jour-nal of Quantitative Linguistics 8, p. 33–50.

Hajič, Jan
2004 Linguistics Meets Exact Sciences. In: In: Susan Schreibman/Ray Siemens/John Unsworth (eds.): A Companion to Digital Humanities. Oxford: Blackwell, p. 79–87.

Haslerud, Vibecke/Anna-Brita Stenström
1995 The Bergen Corpus of London Teenager Language (COLT). In: Geoffrey Leech/Greg Myers/Jenny Thomas (eds.): Spoken English on Computer. London: Longman, p. 235–242.

Hinterhölzl, Roland/Svetlana Petrova/Michael Solf
2005 Diskurspragmatische Faktoren für Topikalität und Verbstellung in der althochdeutschen Tatianübersetzung (9. Jh.). In: Shinichiro Ishihara/Michaela Schmitz/Anne Schwartz (eds.): Approaches and Findings in Oral, Written and Gestural Language, Interdisciplinary Studies on Information structure 3. Potsdam: Universitätsverlag Potsdam, p. 143–182.

Hockey, Susan
2003 Digital Resources in the Humanities: Past, Present, and Future. Towards a Universal Digital Library for the Humanities. In: Thomas Burch/Johannes Fournier/Kurt Gärtner/Andrea Rapp (eds.): Standards und Methoden der Volltextdigitalisierung. Stuttgart: Franz Steiner Verlag. (Akademie der Wissenschaften und der Literatur, Mainz)

2004 The history of Humanities Computing. In: Susan Schreibman/Ray Siemens/John Unsworth (eds.): A Companion to Digital Humanities. Oxford: Blackwell, p. 3–19.

Hollmann, Willem/Anna Siewierska

2003 Corpora and (the Need for) Other Methods in a Study of Lancashire Dialect. In: Zeitschrift für Anglistik und Amerikanistik 2006, p. 203 – 216.

Jannidis, Fotis
2007 Computerphilologie. In: Thomas Anz (ed.): Handbuch Literaturwissenschaft. Bd.2: Methoden und Theorien. Stuttgart, Weimar: Metzler, p. 27–40.

Jannidis, Fotis/Gerhard Lauer/Andrea Rapp
2006 Hohe Romane und blaue Bibliotheken. Zum Forschungsprogramm einer computergestützten Buch- und Narratologiegeschichte des Romans in Deutschland (1500–1900). In: Gisi, Lucas Marco/Loop, Jan/Stolz, Michael (eds.): Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien. Available from germanistik.ch at [19].

Johansson, Stig
2007 Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies. Studies in Corpus Linguistics 26. Amsterdam/Philadelphia: John Benjamins.

Jones, Randall L./Steven P.Sondrup
1989 Computer-aided lexicography: Indexes and Concordances. In: István Bátori/Winfried Lenders/Wolfgang Putschke (eds.) Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications. Berlin: Walter de Gruyter, p.490–509.

Jurafsky, Daniel/James H. Martin
2000 Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, NJ: Prentice-Hall.

Kennedy, Beverly
1997 Contradictory Responses to the Wife of Bath as evidenced by Fifteenth-Century Manu-script Variants. In: Norman F. Blake/Peter Robinson (eds.): The Canterbury Tales Pro-ject Occasional Papers Volume II. London: Office for Humanities Communication, p. 23–39.

Kilgarriff, Adam/Pavel Rychlý/Pavel Smrž/David Tugwell
2004 The Sketch Engine. In: Proceedings of the Eleventh EURALEX International Congress. Lorient, France: Université de Bretagne-Sud, p. 105–116.

Kingsbury, Paul/Martha Palmer
2003 PropBank: the Next Level of TreeBank. In: Proceedings of Treebanks and Lexical Theories '03. Växjö Sweden. Available at [20].

Koehn, Philipp
2005 Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proccedings of the Tenth Machine Translation Summit. Phuket, Thailand.

Kytö, Merja/Susan Romaine
1997 Competing Forms of Adjective comparison in Modern English: What could be more quicker and easier and more effective? In: Terttu Nevalainen/Leena Kahlas-Tarkka (eds.): To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen. Helsinki: Société Néophilologique, p. 329–352.

Lüdeling, Anke

2007 Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In: Gisela Zifonun/Werner Kallmeyer (eds.): Jahrbuch des Instituts für deutsche Sprache 2006. Berlin: Walter de Gruyter, p. 28–48.

Lüdeling, Anke/Seanna Doolittle/Hagen Hirschmann/Karin Schmidt/Maik Walter
2008 Das Lernerkorpus ›Falko‹. In: Deutsch als Fremdsprache 2.

Lüdeling, Anke/Merja Kytö
Forthcoming (eds.) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter.

Lüdeling, Anke/Thorwald Poschenrieder/Lukas C. Faulstich
2005 DeutschDiachronDigital – Ein diachrones Korpus des Deutschen. In: Jahrbuch für Computerphilologie 6 (2004), p. 119–136.

Macé, Caroline/Philippe Baret/Andrea Bozzi/Laura Cignoni
2006 (eds.) The Evolution of Texts: Confronting Stemmatological and Genetical Methods. In: Proceedings of the International Workshop held in Louvain-la-Neuve on September 1–2, 2004. (Linguistica Computazionale XXIV–XXV) Pisa-Rome: Istituti Editoriali e Poligrafici Internazionali.

Mann, William/Sandra Thompson
1987 Rhetorical Structure Theory: A Theory of Text Organization. In: ISI Technical Reports RS-87–190. Los Angeles: Information Sciences Institute, p. 1–81.

Manning, Chris/Hinrich Schütze
1999 Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

McEnery, Anthony/Andrew Wilson
2001 Corpus Linguistics. Edinburgh: Edinburgh University Press.

Menzel, Wolfgang
2004 Zur Geschichte der Computerlinguistik. In: Kai-Uwe Carstensen/Christian Ebert/Cornelia Endriss/Susanne Jekat/Ralf Klabunde/Hagen Langer/Michael Schielen (eds.): Computerlinguistik und Sprachtechnologie – Eine Einführung, Heidelberg: Spektrum-Verlag, p. 1–9.

Merriam, Thomas V. N./Robert A. J. Matthews
1993 Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. In: Literary and Linguistic Computing 9(1), p. 1–6.

Meurers, Walt Detmar/Stefan Müller
Forthcoming Corpora and Syntax. In: Lüdeling/Kytö.

Miltsakaki, Eleni/Rashmi Prasad/Aravind Joshi/Bonnie Webber
2004 The Penn Discourse Treebank. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Mitkov, Ruslan
2003 (ed.) The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press.

Nirenburg, Sergei/Harold L. Somers/Yorick A. Wilks
2003 (eds.) Readings in Machine Translation. Cambridge, MA: MIT Press.

Oakes, Michael P.
Forthcoming Corpus Linguistics and Stylometry. In: Lüdeling/Kytö.

Papineni, Kishore/Salim Roukos/Todd Ward/Wei-Jing Zhu

2002 Bleu: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, PA, p. 311–318.

Pearce, Michael
2007 The Collocational and Colligational Behaviour of the Lemmas MAN and WOMAN in the British National Corpus (BNC). Paper presented at the Corpus Linguistics 2007 conference, Birmingham.

Petrova, Svetlana
2006 A Discourse-Based Approach to Verb Placement in Early West-Germanic. In: Shinichiro Ishihara/Michaela Schmitz, Michaela/Anne Schwartz (eds.): Interdisciplinary Studies on Information structure 5. Potsdam: Universitätsverlag Potsdam, p. 153–185.

Pilz, Thomas/Andrea Ernst Gerlach/Sebastian Kempken/Paul Rayson/Dawn Archer
Forthcoming The identification of spelling variants in English and German historical texts: manual or automatic. In: Literary and Linguistic Computing.

Rayson, Paul/Geoffrey Leech/Mary Hodges
1997 Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. In: International Journal of Corpus Linguistics 2(1), p. 133–152.

Reis, Marga/Stephan Kepser
2005 Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives Berlin: Mouton de Gruyter.

Robinson, Peter M.W.
1994 Collate: A Program for Interactive Collation of Large Textual Traditions. In: Susan Hockey and Nancy Ide (eds.), Research in Humanities Computing 3, p. 32–45.

Robinson, Peter M.W.
2003 The History, Discoveries and Aims of the Canterbury Tales Project. In: The Chaucer Review 38:2, p. 126–139.

Schmid, Helmut
Forthcoming Tokenizing and Part-of-Speech Tagging. In: Lüdeling/Kytö.

Schreibman, Susan/Ray Siemens/John Unsworth
2004 (eds.) A Companion to Digital Humanities. Oxford: Blackwell. Available online at [21].

Semino, Elena/Mick Short
2004 Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing. London and New York: Routledge.

Smith, Nicholas/Sebastian Hoffmann/Paul Rayson
2007 Corpus Tools Today and Tomorrow: Incorporating User-Defined Annotations. Paper presented at the Corpus Linguistics 2007 conference, Birmingham.

Solopova, Elizabeth
1997 The Problem of Authorial Variants in »The Wife of Bath's Prologue«. In: Norman F. Blake,/Peter Robinson (eds.): The Canterbury Tales Project Occasional Papers Volume II. London: Office for Humanities Communication, p. 143–164.

Somers, Harold
Forthcoming Corpora and Machine Translation. In: Anke Lüdeling/Merja Kytö.

Stede, Manfred
2004 The Potsdam Commentary Corpus. In: Bonnie Webber/Donna K. Byron (eds.): Asso-ciation for Computational Linguistics (ACL) 2004 Workshop on Discourse Annotation. Barcelona, Spain, p. 96–102.

Thompson, Henry/David McKelvie
1997 Hyperlink semantics for standoff markup of read-only documents. In: Proceedings of SGML Europe'97 [22].

Tognini Bonelli, Elisa
2001 Corpus Linguistics at Work. Amsterdam: John Benjamins.

von Waldenfels, Ruprecht
2006 Compiling a Parallel Corpus Of Slavic Languages: Text Strategies, Tools and the Question of Lemmatization in Alignment. In: B. Brehmer/V. Ždanova/R. Zimny. (eds.): Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. Munich, p. 123–138.

Xiao, Richard
Forthcoming Theory-Driven Corpus Research: Using Corpora to Inform Aspect Theory. In: Lüdeling/Kytö.

Zampolli, Antonio
2001 Language Resources the current situation and opportunities for co-operation between Computational Linguistics and Humanities Computing. In: Domenico Fiormonte/Jonathan Usher (eds.): New Media and the Humanities. Research and Applications. Oxford: Humanities Computing Unit, p. 69–84.

Zeldes, Amir
2007 Machine Tranlsation between Language Stages: Extracting Historical Grammar from a Parallel Diachronic Corpus of Polish. In: Proceedings of Corpus Linguistics 2007, Birmingham. Available online at [23].

Websites

[1] <http://www.zdv.uni-tuebingen.de/tustep/tustep_eng.html> (12.01.2008).
[2] <http://www.canterburytalesproject.org/> (12.01.2008).
[3] <http://www.gutenberg.org> (12.01.2008).
[4] <http://www.zeno.org> (12.01.2008).
[5] <http://www.ilc.cnr.it/EAGLES96/typology/typology.html> (12.01.2008).
[6] <http://www.natcorp.ox.ac.uk/> (12.01.2008).
[7] <http://www.tlg.uci.edu/> (12.01.2008).
[8] <http://www.perseus.tufts.edu/> (12.01.2008).
[9] <http://www.tei-c.org> (12.01.2008).
[10] <http://urd.let.rug.nl/tiedeman/OPUS/> (12.01.2008).
[11] <http://lse.umiacs.umd.edu/> (12.01.2008).
[12] <http://www.jstor.org/> (12.01.2008).
[13] <http://wordnet.princeton.edu/> (12.01.2008).
[14] <http://www.dwds.de/> (12.01.2008).

[15] <http://corpus.byu.edu/> (12.01.2008).
[16] <http://www.sketchengine.co.uk/> (12.01.2008).
[17] <http://www.textgrid.de/> (12.01.2008).
[18] <http://books.google.com/> (12.01.2008).
[19] <http://www.germanistik.ch/publikation.php?id=Hohe_Romane _und_blaue_Bibliotheken > (12.01.2008).
[20] <http://w3.msi.vxu.se/users/rics/TLT2003/doc/kingsbury_palmer.pdf> (12.01.2008).
[21] <http://www.digitalhumanities.org/companion/> (12.01.2008).
[22] <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>.
[23] <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/amir/pdf/MachineTranslationbetweenLanguageStages.pdf>.