MedieKultur | Journal of media and communication research | ISSN 1901-9726

Article – Theme section

Linguistics and the digital humanities:
(computational) corpus linguistics

Kim Ebensgaard Jensen

MedieKultur 2014, 57, 115-134

Published by SMID | Society of Media researchers In Denmark | www.smid.dk

The online version of this text can be found open access at www.mediekultur.dk

Corpus linguistics has been closely intertwined with digital technology since the introduction of university computer mainframes in the 1960s. Making use of both digitized data in the form of the language corpus and computational methods of analysis involving concordancers and statistics software, corpus linguistics arguably has a place in the digital humanities. Still, it remains obscure and figures only sporadically in the literature on the digital humanities. This article provides an overview of the main principles of corpus linguistics and the role of computer technology in relation to data and method and also offers a bird's-eye view of the history of corpus linguistics with a focus on its intimate relationship with digital technology and how digital technology has impacted the very core of corpus linguistics and shaped the identity of the corpus linguist. Ultimately, the article is oriented towards an acknowledgment of corpus linguistics' alignment with the digital humanities.

Introduction

If there is one discipline within the humanities that has embraced the digital nearly since the inception of modern computer science, it must be corpus linguistics (henceforth, CL). Yet, CL remains rather obscure in most contemporary literature on the digital humanities (henceforth, DH). In DHM (2009), "automating corpus linguistics" is mentioned in pass-

ing, and no mention is made in an overview of digitisation by Burdick et al. (2012, pp. 8-9). Svensson (2010, p. 114) merely alludes to linguists who use concordancers. Hockey (2004) more generously acknowledges CL's place in the history of DH. Despite its relative obscurity in the overall sphere of the humanities, CL has been around for decades and has made leaping and creeping advances in tandem with the development of digital technology.

This state-of-affairs raises two related questions. First, one might ask how CL has evolved and why it has evolved as a digital humanistic framework whose focus is on frequency and quantification. Second, one might ask how DH and CL, as digital humanistic frameworks, relate to one another and whether or not DH and CL are relevant to one another. The present article addresses these questions, discussing CL-DH interrelations and introducing some of the principles and goals of CL; a bird's-eye view of CL as a methodological framework is also provided.

Our overview of the principles and history CL is covered in the first three sections of the present article. These should clearly show how CL is inextricably tied in with computer technology in terms of collection and analysis of data, as we relate language corpora to Owens' (2011) three conceptions of data in the humanities. This should provide some answers to the first of the two above-mentioned questions. As we will see throughout this article, there are several parallels and alignments between DH and CL, but CL is still marginal in contemporary DH as it is. In the last section of the present article, we will discuss some potential reasons for this while also addressing the mutual relations of relevance between DH and CL.

Corpus linguistics: definitions and history

Kirk (1996a, pp. 250-251; see also McEnery & Hardie 2012, p. 1) offers the following insight into the position of CL within linguistics as such:

> Corpus linguistics is not like psycholinguistics, sociolinguistics, or geolinguistics, in which the theories and methods of neighboring subject areas are applied to and correlated with linguistic data, the variation in which can usefully be explained by invoking the criteria of the respective subject areas. Nor is it like phonology or lexicology, with a *-logos* 'written word' or 'knowledge' element about the sounds or vocabulary of the language. Nor is it like semantics or pragmatics, in which there is a similar focus on the *-ics* 'written word' or 'knowledge' about the formal or contextual meaning of linguistic utterances. Corpus linguistics does not align itself with any of these other *-linguistics*, *-ologies*, or *-ics*. As a methodology for research, corpus linguistics is in a class of its own. In particular, it foregrounds data and methodology, analysis, and interpretation. If the *-istics* analogy is valid at all, then corpus linguistics has a methodological bias referring to the use of corpora and computers as tools for analysis add to the use of the results as a basis for interpretation in the study of any aspect of language.

Kirk captures three very important characteristics of CL. First, it is theory-neutral to a large extent. Second, it is a corpus-based empirical methodology of language research. Third, CL

is closely associated with the use of computers as tools of analysis. McEnery & Wilson (2001, p. 1) characterize CL as "the study of language based on examples of 'real life' language use". Similarly, Biber et al. (1998, p. 4) inform us that "it is empirical, analyzing the actual patterns of use in natural texts" and that "it depends on both quantitative and qualitative analytical techniques".

Corpora are fundamental to CL as an empirical endeavor. They form the basis of analysis and provide data for hypothesis-testing, language model construction, exemplification, and empirical grounding (Kirk 1996b, pp. 252-254). In the humanities, 'corpus' is generally used with reference to a collection of 'texts' in the broadest semiotic sense. In CL, the term covers a "large and principled collection of natural texts" (Biber et al. 1998, p. 4). Thus, a corpus linguist's conception of corpora is more specific: a corpus must be designed for linguistic analysis; general text archives typically do not qualify as corpora but are seen as databases (Baker et al. 2006, p. 48; Gries 2009a, pp. 8-9).

Today, corpora are *per se* digital, but computerized language corpora were once the exception (Aston & Burnard 1998, p. 5; Svartvik 2007). 'BC' and 'neolithic corpus linguistics' are sometimes used jokingly by corpus linguists to refer to corpus-based language descriptions 'before computers' (Svartvik 2007, p. 12). Joking aside, this reflects how closely corpus linguists identify themselves with the use of computers. Corpus-based language descriptions may be traced back to the late 19th-century work of Friedrich Wilhelm Käding, who made use of an impressive collection of some 11 million words. Another (in)famous example is John Murray's work in English lexicography, which resulted in the *Oxford English Dictionary*; Murray had collected around 4 million citation slips and employed his own children to sort and alphabetize his corpus. Other notable names are Alexander J. Ellis, Otto Jespersen, Charles Carpenter Fries, Franz Boas, Edward Sapir, Leonard Bloomfield, Zellig Harris, and Archibald A. Hill.

The BC era ended in the 1960s with the introduction of computer mainframes at universities. Shortly prior to the introduction of computers, *The Survey Corpus* was initiated by the Survey of English Usage research group. Thus, a corpus consisted of manually annotated paper slips documenting instances of authentic language use. *The Survey Corpus* remains hugely important in at least two respects. First, many future spearhead figures in CL, such as Jan Svartvik, Geoffrey Leech and Sidney Greenbaum, worked on *The Survey Corpus*. Second, some of the most influential descriptive grammars of English in the 20th century were based on *The Survey Corpus*. The first machine-readable corpus (and, thus, the corpus that rocketed CL into the digital era) was W. Nelson Francis and Henry Kučera's *Brown Corpus* of written American English, which was completed in the early 1960s (Svartvik 2007, p. 14). *The Survey Corpus* was eventually also digitized and assimilated by the *London-Lund Corpus*. Both the *London-Lund Corpus* and the *Brown Corpus* consisted of 1 million words, which, in the nascence of modern computer technology, was impressive.

CL has evolved in tandem with computer technology. Computers not only allowed for storage and the processing of increasingly massive amounts of data, but they also enabled

increasingly complex quantitative analysis, which is integral to the study of language use. While early corpus analysis consisted of word counting which required huge amounts of processing by building-sized computers in nearly inaccessible computer labs in university basements, corpus linguists can now perform advanced statistical analyses on their laptops at home or in their offices, using platforms such as *R* (Gries 2009a,b), *Python* (Bird et al. 2009), or *Perl* (Hammond 2003).

Just as digital humanists face adversity in the form of traditionalist views in their respective departments, CL was generally shunned by the linguistics establishment until the late 1980s. This was because generative linguistics, which was dominant in that period, rejected the evidentiary value of corpora, adopting instead native speaker intuition as the only true source of reliable language data, and CL was simply considered a waste of time (Francis 1982, pp. 7-8). Since generative linguistics was dominant in the 1960s-1980s, parallels may be drawn between the hostility towards CL and an established academic elite's animosity towards researchers who operate at the fringes of the field in question. This is not unlike contemporary digital humanist Dan Edelstein's experience of how his seniors perceive his digital mapping of the exchange and flow of knowledge in the Enlightenment "as whimsical, the result of playing with technological toys" (Cohen 2010). In the 1960s-1980s, when digital analytical techniques were rather primitive, doing computer-aided linguistic analysis was genuinely frowned upon by the academic establishment and viewed merely as the work of a spanner-wielding handyman as opposed to the proper academic work of the rationalist elite (Svartvik 2007, pp. 19-20; see also Fillmore 1992, p. 35). Indeed, back then, "corpus work was, indeed, little else but donkey work" (Leech 1991, p. 25).

The data dimension

As Kirk (1996b, p. 251) points out, "[t]here are two dimensions to corpus linguistics as a field of scholarly research: data and methodology". The data dimension covers the compilation of a corpus and the data included in it, while the methodology dimension covers the use of such data in linguistic analysis.

Corpus data fall under all three possible treatments of data within the humanities: data as constructed artifacts, data as interpretable texts, and data as processable information (Owens 2011). Regarding data as constructed artifacts, Owens (2011, p. 6) informs us that "[t]he production of a data set requires choices about what and how to collect and how to encode the information", which applies to the process of corpus compilation in the selection and subsequent encoding of texts for inclusion. In essence, a corpus is a constructed artifact that captures not the direct instances of language use but processed representations thereof.

This awareness is reflected in, for instance, Kirk's (1999, p. 35) description of a corpus as a mirror of language use. Owens (2011, p. 6) also writes:

> Now, when data is transformed into evidence, when we isolate or distill the features of a data set, or when we generate a visualization or present the results of a statistical procedure, we are not presenting the artifact. These are abstractions. The data itself has an artifactual quality to it. What one researcher considers noise, or something to be discounted in a dataset, may provide essential evidence for another.

This relates to the analysis dimension of CL, as quantitative analysis of a given phenomenon in the corpus represents, in the form of frequencies, an artifactual rendering of its behavior within the corpus. This is a fact that most corpus linguists acknowledge, and few corpus linguists would uncritically see corpus data as corresponding directly to the truth. Moreover, results of corpus analyses are only acceptable as representations of usage patterns if shown to be statistically significant. As for data as processing, Owens (2011, p. 7) writes: "Data can be processed by computers. We can visualize it. We can manipulate it. We can pivot and change our perspective on it. Doing so can help us see things differently." This is primarily relevant to the analysis dimension, since the data that such processing amounts to is identical to the representations of usage patterns mentioned above.

While these two ways of understanding data are readily applicable to CL, the application to corpus data of Owens' (2011, p. 7) take on data as interpretable texts may require some explication:

> As a species of human-made artifact, we can think of data sets as having the same characteristics as texts. Data is created for an audience. Humanists can, and should interpret data as an authored work and the intentions of the author are worth consideration and exploration. At the same time, the audience of data also is relevant. Employing a reader-response theory approach to data would require attention to how a given set of data is actually used, understood, and interpreted by various audiences.

There is no denying that linguistic data are human-produced, and few corpus linguists would deny that language is a means of interpersonal communication that involves senders and recipients. Senders are aware of recipients, and senders' utterances and texts are shaped by various intended effects on the recipients. Moreover, utterances and texts are decoded and interpreted by their recipients in accordance with a number of contextual, cultural, cognitive, and communicative factors. In that sense, corpus data are "authored" and "audience"-directed simply because they represent communicative situations and the interlocutors therein. This is just the reality of language. Many spoken language corpora are even digitally encoded with information on interlocutors. The spoken component of the *British National Corpus* contains sociolinguistic information on speakers, such as gender, age, social class, and dialect (Aston & Burnard 1998, pp. 112-128); such information allows for studies such as the investigation by McEnery et al. (2000) of gender-targeted verbal abuse. There is another layer to this, which relates to the point of corpora being artiacts themselves: a corpus is compiled by someone, and the texts within have been carefully selected to serve the purpose of the corpus. When all is said and done, the computational

or manual analysis of corpus-date involves interpretation on the part of the compiler and the user.

Let us return to the evolution of the notion of a corpus. Sinclair (1996) defines a corpus as "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language", which is contrasted with the notion of a computer corpus: "A computer corpus is a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance". Likewise, in Leech (1987), the terms 'computer corpus' and 'machine-readable corpus' figure in contrast with 'corpus'. In more recent definitions, this distinction has vanished, and corpora are *per se* electronic and machine-readable now (Baker et al. 2006, p. 48). Gries (2009a, p. 7-8) points out that

> nowadays virtually all corpora are stored in the form of plain ASCII or Unicode text files that can be loaded, manipulated, and processed platform independently. This does not mean, however, that corpus linguists only deal with raw text files – quite the contrary: some corpora are shipped with sophisticated software that makes it possible to look for precisely defined syntactic and/or lexical patterns. It does mean, however, that you would have a hard time finding corpora on paper, in the form of punch cards, or digitally in HTML or Microsoft Word document format; the current standard is text files with XML annotation.

The logic behind this development is simple: since corpora are now digital by default, and virtually all previous paper corpora have been digitized, it does not make sense to distinguish between corpora and computer corpora.

Corpus compilation is typically a five-step process (Kennedy 1998, pp. 70-85):

1. corpus design
2. planning the storage system
3. obtaining copyright for the material included in the corpus
4. capturing the text to be included in the corpus
5. markup/annotation of the corpus

Corpora are, thus, principled collections of texts that document naturally occurring spoken or written language. 'Principled' entails that the texts in a corpus are not random but are carefully selected so that they represent, in a balanced fashion, language use in the type(s) of communicative situation to be represented in the corpus. 'Naturally occurring' means that the texts are empirically collected from their natural environments. Text archives may in some cases serve as corpora insofar as they happen to be representative of language use within whatever domains they are associated with. In fact, in the 1990s and early 2000s, newspaper CDs were not uncommonly used as corpora, and Minugh (2000, p. 60) points out in his study of idioms in newspaper English that newspaper CDs are "all heterogeneous, in the sense that they represent a well-defined genre".

A distinction is made between general corpora, representing language use generally – examples are the *British National Corpus* (Aston & Burnard 1998) and *Corpus of Contemporary American English* (Davies 2014) – and specific corpora, representing language use in specific domains (Gries 2009a, p. 9) – examples are the *TIME Magazine Corpus* and *Corpus of American Soap Operas* (Davies 2014), and the *Chat Corpus*, *Beauty Blog*, *Europarl*, and *Supreme Court Dialogs* (Bick 2014a), as well as the *Carnegie Mellon University Recipes Database* (Tasse & Smith 2008). A good example of a study based on specific corpora is Ooi (2000), which reveals a number of interesting uniquely Singaporean-Malaysian cultural concepts reflected in collocations, such as 'killer litter', 'urine detector', and 'weekend car', drawing on a corpus built from selected Singapore- and Malaysia-based newspaper archives. In the absence of a corpus featuring spontaneous spoken discourse in the Tyneside dialect of English, de Lopez (2013) constructed a corpus based on dialogs from the contemporary reality show *Geordie Shore* and the 1980s drama series *Auf Wiedersehn, Pet* and then analyzed the use of clause final 'man' in the Tyneside dialect, falsifying previous hypotheses and identifying a range of previously unattested communicative functions.

Planning storage systems is becoming increasingly easy with the development of digital storage technology. While early electronic corpora were stored on magnetic tape, the primary storage medium in the 1990s and early 2000s was the CD-ROM. Today, a corpus can be stored in text-files in a zipped folder on a hard disk or via a cloud service. A corpus, even a very large one, may also be stored on an Internet server and accessed via its own webpage, ensuring greater public accessibility. This is compatible with DH's quest for open access to data and metadata in general (DHM 2009, pp. 3, 10; MDH 2011, point 9) and DH's more general utopian core:

> Digital Humanities have a **utopian core** shaped by its genealogical descent from the counterculture-cyberculture intertwinglings of the 60s and 70s. This is why it affirms the value of the open, the infinite, the expansive, the university/museum/archive/library without walls, the democratization of culture and scholarship, even as it affirms the value of large-scale statistically grounded methods (such as cultural analytics) that collapse the boundaries between the humanities and the social and natural sciences. This is also why it believes that copyright and IP standards must be freed from the stranglehold of Capital, including the capital possessed by heirs who live parasitically off of the achievements of their deceased predecessors. (DHM 2009, p. 3; see also MDH 2011, points 7-8).

It should be mentioned that corpus compilers have faced restrictions in terms of copyright issues. In the past, access to corpora typically had to be purchased. Recently, ways of providing corpora that observe copyright laws and still, to varying extents, appeal to the democratization of knowledge have been adopted, and both on- and offline corpora are increasingly becoming freely accessible. In some cases, the potential user is required to plead non-commercial use of the corpus. Another model provides free, but limited, access to the corpus for unregistered users and unlimited, or less limited, access for registered users.

The fourth step is where digitization really comes into play because it is here that the texts in question are transferred from their original formats into a digital format. Originally, text capture manually typing in the texts. With the arrival of image scanning technology and, subsequently, image-to-text conversion, text capture became much easier. This, along with advances in data storage technology, has played a pivotal role in the massive increase in corpus sizes to the point that corpora of less than 100,000 words have been termed mini-corpora (Kirk 1996b, p. 251). Even more recently, corpora also include language data from the Internet, which is, of course, already in digital form; two examples are *Global Web-Based English* and *TIME Magazine Corpus* (Davies 2014). In such cases, text capture consists of harvesting the Internet for texts. With spoken data, the custom has been to include them in corpora in transcribed form. Due to a lack of speech-to-text technology, this has largely been done manually. While digital capture of spoken data is still at an early stage, technological advances have been made. Kirk (1996a, p. 263) notes that "actual speech can nowadays be provided in a digitised fashion, as a kind of voice mail (if this is still a rare occurrence, it will only become more common in the future)". With today's development of SoundCloud and similar technology that allows the embedding of sound files, actual speech can be distributed in digital form beyond the restrictions of voicemail. Skype technology, which has already seen use in CL, makes possible recording and speech-to-text conversion via speech recognition software and will probably play a revolutionizing role in the future. Moreover, software such as the CLAN system whose multimodal nature enables real-time linking of transcription to both audio and video files (MacWhinney 2014). Recently, the auditory data in the spoken component of the *British National Corpus* were included into the corpus, so that it is now possible to access, via its XIARA concordancer, both sound and transcription.

The last step, markup/annotation, is hugely important. Text archives are annotated mainly in terms of biographical, bibliographical, and historical information. By contrast, corpora are often very richly annotated, featuring the above-mentioned types of information and various linguistic and extralinguistic information. This is because linguistic analysis differs in some respects from other types of analysis within the humanities in terms of focal areas and the nature of inquiry. Moreover, because CL focuses on patterns of language use, descriptions must account for contextual features, requiring such features to be encoded into the corpus. According to Ide (2004), "a distinction is made between the *primary data* (i.e., the text without additional linguistic information) and their annotations." Within the primary data, metadata and elements such as text structure, paragraphs, titles, headings, various orthographic features, and footnotes are annotated. Non-primary data encoding covers annotation for additional linguistic information as well as other extralinguistic information. Here is a list of the information types included in annotation systems in contemporary corpora (Aston & Burnard 1998, pp. 33-36; McEnery & Wilson 2001, pp. 32-69; Ide 2004):

- Part-of-speech annotation: most contemporary corpora are annotated in terms of part-of-speech so that each unit in the corpus is tagged with a code that indicates its word class membership. Such annotation is typically done automatically, and a number of considerably precise part-of-speech taggers are available for this task, such as the *Stanford PoS Tagger* (SNLPG 2014), the *UCREL CLAWS Tagger* (Garside 1987, UCREL 2014), or the *VISL* parser Bick (2014b).
- Syntactic annotation: some corpora are annotated in terms of syntactic function and structure. A number of parsers are capable of doing this automatically, but the margin of error is typically larger than with part-of-speech tagging. It should be mentioned that a number of part-of-speech taggers also specify some morpho-syntactic. An example of a quite powerful syntactic tagger is the *VISL* parser (Bick 2014b), which draws on constraint grammar (Karlsson et al. 1995).
- Semantic annotation: this type of annotation "can be taken to mean any kind of annotation that adds information about the meaning of elements in a text" (Ide 2004) and covers specification of both grammatical and lexical semantics. Lexical semantics is particularly challenging and, in its current state, automatic tagging in this area amounts to crude dictionary-based sense tagging.
- Discourse-pragmatic annotation: some corpora are annotated in terms of various aspects of discourse-pragmatics. Ide (2004) mentions topic identification as information that may be annotated for automatically thanks to topic detection/tracking technology. Ide (2004) also mentions annotation for co-reference and discourse structure; this sort of annotation is still primarily done manually due to limitations in digital technology in this area. An example of a corpus with discourse-pragmatic mark-up is Kirk's (2013) *SPICE-Ireland* corpus of spoken Irish English, which is annotated for speech-act functions, utterances, discourse markers, quotatives, and prosody.
- Annotation of spoken language: corpora may also be annotated for features of spoken language. Such annotation may include information on prosody – as in the *SPICE-Ireland* corpus (Kirk 2013) – or pronunciation in transcribed form.
- Problem-oriented annotation: this is the annotation of a corpus in terms of information specific to the analyst's task at hand (Baker et al. 2006, pp. 134-135).

There are a number of possibilities for experimentation with non-automatic techniques from contemporary DH here. Automated annotation is a goal in CL, because manual annotation of large corpora is extremely time-consuming. This goal was set in a time when online networking and mass collaboration were impossible. Nowadays, instantaneous online mass collaboration is possible in the form of expert networks and crowd science, and it might be worth considering whether network or crowd annotation is an alternative to automatic tagging of semantic and discourse-pragmatic information (e.g., Fromheide et al. 2014). Of course, such an approach would reintroduce the human error factor into the process, but it would also include the dynamism of context-awareness that humans have

and computers lack. It should be mentioned that, when it comes to annotation, CL has always been closely allied with structural computational linguistics to the point that the two are sometimes conflated. With its focus on developing computational models of morphosyntax that enable automatic analysis, structural computational linguistics is invaluable to CL's quest for automatic annotation. Conversely, CL provides structural computational linguistics with attested naturally occurring language which can serve as input for the development of realistic language computational models. Crowd sourcing could potentially strengthen this alliance in the following way: having the crowd re-evaluate automatic morphosyntactic tagging and, then, feeding this information back into the computational language model would improve the model considerably and, thus, allow for more precise automatic morphosyntactic annotation. This procedure is essentially a crowd science version of manual benchmarking in which a human evaluates and corrects analytical output by an automatic parser; Hovy et al. (2014) are currently experimenting with crowd sourcing in relation to morphosyntactic tagging.

Arguably, building a corpus is inherently a DH project. It involves digital archiving of a body of human-made artifacts (i.e., the texts and the usage-events therein) that are processed and interpreted via a plethora of digital methods. Corpus-building is typically also a collaborative effort involving computational linguists and corpus linguists, and it is not uncommon that more organic network-like processes are involved in the processing of the data in the corpus. For instance, a tagset developed for one corpus might be adopted and adapted by the compilers of a different corpus – an example of this is the application of the UCREL CLAWS 7 tagset in the *Corpus of Historical American English*, the *Corpus of Contemporary American English*, or the *TIME Magazine Corpus* (Davies 2014). In addition, many specialized corpora are often compiled from existing general corpora, as is the case of the *Lancaster Corpus of Abuse* (McEnery et al. 2000) and the *VU Amsterdam Metaphor Corpus Online* (Steen et al. 2012). Both include data from the *British National Corpus*. This is a good example of digital corpus data being adopted and adapted to special purposes and undergoing further interpretation in the process.

There is one point at which CL and DH diverge in connection with data. The latter tends to operate with unstructured data in the form of "raw" text whose only structure is often that individual texts are stored in their own documents. By contrast, corpus linguists operate with highly structured data. Since the analytical techniques of CL require structured data, they may not be readily applicable to the digital humanist's unstructured data, and text-mining or data-mining techniques are considered more appropriate in this respect. However, corpora also start out as "raw", unstructured data before they undergo the process outlined by Kennedy (1998, pp. 70-85). It is true that corpora are structured from the start due to the principled selection of texts. I would contend that a DH project – in particular, an archival one – must be principled in its selection of data simply because the selected data must be in line with the overall purpose of the project. For instance, a recently initiated digitization of all Norwegian texts ever produced (Elmelund 2014) is

inherently principled in that it only includes Norwegian texts. In that sense, many DH projects share some fundamental starting points with most corpus compilation projects. While unstructured data are widespread in DH, one might speculate that data structuralization may to some extent be desirable to digital humanists and that techniques akin to those of CL may be deployed. One might object that the digital humanist's unstructured data sets are so massive that structuralization becomes too challenging. However, structuralization of digital data would allow data-retrieval techniques associated with CL alongside those that are more typical of DH, such text- and data-mining. One goal, which I would consider a very important one for digital humanists and corpus linguists to collaborate on in this respect, would be the development of methods of structuralization that both digital humanists and corpus linguists could benefit from. For instance, in addition to part-of-speech and syntactic annotation, annotation in terms of sentiment, text structure, topic information and a host of other features would be as useful to linguists and discourse analysts as to literary critics and cultural analysts.

The methodology dimension

As seen above, analysis enters the picture already in the corpus compilation process in the annotation phase. Before looking at the use of computational tools in such analysis, we should address the empiricism and objectivity of corpus analysis.

As Ide (2004) points out, "[a] corpus provides a bank of *samples* that enable the development of numerical language models, and thus the use of corpora goes hand in hand with empirical methods." CL has to be empirical *per se* because one cannot really describe language use without looking at language in use, and rationalist introspection is of limited value in this respect. As Berez & Gries (2009, pp. 158-159) point out,

> the corpus output from a particular search expression together constitute an objective database of a kind that made-up sentences or judgments often do not. More pointedly, made-up sentences or introspective judgments involve non-objective (i) data gathering, (ii) classification, and (iii) interpretative processes on the part of the researcher. Corpus data, on the other hand at least allow for an objective and replicable data-gathering process; given replicable retrieval operation, the nature, scope, an idea underlying the classification of examples can be made very explicit.

In a similar vein, Kirk (1996b, pp. 253-254) argues that, among the scientific strengths of corpus-based models or theories of language, we find falsifiability, replicability, and objectivity. Moreover, "the richness and diversity of naturally occurring data often forces the researcher to take a broader range of facts into consideration" (Berez & Gries 2009, p. 158), because the language system transcends the language competence of the individual native speaker. This does not mean that introspection is totally absent from CL. Introspection often figures in the construction of hypotheses to be tested against corpus data. Moreover,

while "[w]ith a corpus-based methodology, subjectivity is controlled" (Kirk 1996b, p. 254), Berez & Gries (2009, p. 158) remind us that "the analysis of corpus data requires classificatory decisions which are not always entirely objective – no corpus linguist would deny this fact, just as no scientist would deny that some degree of intuition plays a role in nearly any study". However, CL is ultimately an empirical discipline and not a rationalist one.

Kirk (1996b, p. 252) provides this description of the role of computers in CL as a methodology:

> The methodology of corpus linguistics as a branch of linguistic enquiry is inseparable from the computer's resources not only to store data but to sort, manipulate, calculate, and transform them and to present them in a wide range of different output formats, all dumb ways characteristic of the machine itself; moreover, the computer is increasingly being used to store annotations of the data in the form of encodings representing analyses, categorizations, and interpretations of the data and to manipulate those – and thus to behave in a seemingly intelligent way.

The application of dumb and intelligent computational processes is at play in the annotation stage of corpus compilation and in the use of corpus-date in linguistic analysis. This is one of the areas in which the distinction between the data dimension and the methodology dimension is blurred, and another reflection of the artifactual nature of language corpora. At this point, we should be reminded that

> [i]n much the same way that encoding a text is an interpretive act, so are creating, manipulating, transferring, exploring, and otherwise making use of data sets. Therefore, data is an artifact or a text that can hold the same potential evidentiary value as any other kind of artifact. That is, scholars can uncover information, facts, figures, perspectives, meanings, and traces of thoughts and ideas through the analysis, interpretation, exploration, and engagement with data, which in turn can be deployed as evidence to support all manner of claims and arguments. (Owens 2011, p. 7)

Corpus data hold evidentiary value, endowing corpus methodology with a powerful potential to test hypotheses and to produce strong descriptions, theories, and models of language.

It is generally accepted in contemporary CL that the usage-patterns, or discursive behavior, of a linguistic unit can be identified by investigating its association patterns, which are "the systematic ways in which linguistic features are used in association with other linguistic features" (Biber et al. 1998, p. 5). Given that corpus linguists identify and investigate patterns, quantification is important. The argument is simple: a pattern is based on regularity, and regularity is reflected in frequency. Thus, CL involves both qualitative analysis and quantitative analysis in order to identify and categorize linguistic features and patterns of use.

Given the importance of association patterns, one of the central digital tools for the analysis of corpus data is the concordancer. Concordancers automatically produce lists of keywords, or key expressions, that the user instructs the concordancer to find in the corpus in question. The list, or concordance, features the immediate context of each instance of the keyword in question – this is referred to as KWIC (*key*word *i*n *c*ontext); most concordancers also allow for including more context as an option.

The evolution of concordancers has been crucial to the development of CL itself. McEnery & Hardie (2012: 37-48) conceptualize this development in terms of generations:

- First generation concordancers: the earliest concordancers produced only KWIC lists and were restricted to the mainframe systems of the institutions in which they were developed and often required immense amounts of computational power.
- Second generation concordancers: this category comprises the first concordancers available for use on a PC. Second generation concordancers were not designed for operating with annotations and, due to the limited computational power of early PCs, they would cause the computer to run out of memory if the datasets were too large. They were nonetheless of tremendous importance to the growth of CL, because one was no longer dependent on access to a university mainframe. McEnery & Hardie (2012, p. 39) link the introduction of second generation concordancers to the CL boom of the late 1980s.
- Third generation concordancers: operable on PCs, these concordancers may, thanks to the introduction of Unicode, be applied to more than just one language, and they are able to operate with annotation (originally in SGML but now primarily in XML) and often provide word lists and statistics in addition to KWICs.
- Fourth generation concordancers: these concordancers are very similar to third generation ones. However, while third generation concordancers are installed and run on a PC and applied to corpora stored in the PC in text file format, fourth generation concordancers are online and applied to online corpora, thus providing worldwide access to the corpora in question.

Since statistical analysis is much more cost-efficient to carry out computationally than manually, it was inevitable that corpus linguists should embrace computational statistical tools. The simplest form of statistical corpus analysis is the frequency list, which is a list of words or expressions accompanied by their frequencies of occurrence within the corpus in question. Most third and fourth generation concordancers can generate frequency lists. Such lists are useful in providing an idea of the texture of the corpus, but, beyond that, they are too rude and simplistic.

More sophisticated processing in the form of statistical analysis is needed for its evidentiary value to be really compelling. As a first step, the frequency data should be tested and verified as statistically significant and not coincidental. Such testing may itself be a compli-

cated affair and ranges from difficult to impossible to do manually. Fortunately, computers can do this sort of work for us, and a multitude of software is available, including purchasable statistics packages such as *SSPS*, spreadsheets, easily accessible online problem-oriented calculators like the *Easy Chi-square Calculator*, and complex and sophisticated open source platforms such as *Python* and *R*. The latter is increasingly popular among corpus linguists due to its flexibility (it allows users to write their own scripts through the *Rgui* code editor or various *R*-oriented IDEs such as, for instance, *Tinn-R*, *RStudio*, or *RKWard*) and accessibility. *R* is free and generally nurtures an open source culture in its users in which data are shared and scripts are exchanged in a network-like fashion, so that researchers facilitate each other's work; not only is *R* gaining popularity among corpus linguists (Gries 2009a,b), it has also made its way into digital literary text analysis (Lockers 2014). The former has also been deployed in the name of CL due to the availability of libraries such as *pandas* and *NymPy*. The increasing popularity of open source tools and willingness to share scripts and software (including certain concordancers), as well as corpora that are not bound up in copyright restrictions, reflect a culture among corpus-linguists that is similar the utopian core of DH (DHM 2009, p. 3).

Corpus linguistics and the digital humanities

Until now, I have treated DH as a unified entity but it actually seems to me that there are at least two conceptions of DH. In one perspective, DH includes most frameworks in the humanities in which the digital is strongly integrated. In the other, DH is more strictly defined almost as a specific type of delimited research program. In this strict view, DH is perceived in terms of a wave metaphor:

> The first wave of digital humanities work was quantitative, mobilizing the search and retrieval powers of the database, automating corpus linguistics, stacking hypercards into critical arrays. The second wave is **qualitative**, **interpretive**, **experiential**, **emotive**, **generative** in character. (DMH 2009, p. 2 – boldface in original; see also Burdick et al. 2012, p. 8)

Here, the central concepts of DH research are "subjectivity, ambiguity, contingency, observer-dependent variables in the production of knowledge" (DMH 2009, p. 2), and it seems that certain dimensions are deemed obligatory, such as collaborative participation, crowd involvement, transdisciplinarity, and mass dissemination (see Burdick et al. 2012, pp. 61-73). The two conceptions share the utopian core (DHM 2009, p. 3) and the focus on the digital. It is very likely that the loose and strict versions form a continuum. However, the relevance of CL differs greatly in the two conceptions.

Given the placement of CL in the first DH wave, CL would not be terribly relevant to strict DH. This owes to the logic of the wave metaphor: when a wave breaks on the shore, it may leave some traces, but it ultimately ceases to exist, and its traces are washed away by the next wave. In this metaphor, CL has played out its role in DH, since a second – qualita-

tive and experiential – wave is currently rolling in. This easily explains the obscurity of CL in many contemporary DH circles. At this point, I must raise some criticism of the strict conception of DH. First, the wave metaphor is unfortunate, because it imposes an imprecise structure upon the state-of-affairs in the humanities. According to this metaphor, quantifying digital approaches within the humanities should be non-existent, or at least stagnant. This is not the case. As demonstrated in the previous sections, CL still alive and kicking, and it continues to undergo development and expansion in tandem with the evolution of digital technology. Lockers (2014) provides a good example of a similar situation within quantitative literary analysis. Second, strict DH is, ironically, exclusionist. The irony lies in the fact that strict DH embraces transdisciplinarity and simultaneously excludes approaches associated with the first wave as well as approaches that do not fit wholly into the ideals of the second wave.

In the loose conception of DH, CL and its quantitative framework remain relevant, because the digital is essential in CL. However, the loose conception may also be criticized, as it does not encourage transdisciplinarity and may even be argued to nurture stagnation. If conceptualized merely as covering humanistic discipline where in which the digital is central, DH would, perhaps, be too unfocused even to be considered a movement within the humanities, and there would be a lack of goals towards which to progress; ultimately, the relevance of CL – or any other framework – to DH would itself be irrelevant.

Arguably, the optimal form of DH would fall somewhere between the loose and the strict conception, as it would encourage progression and transdisciplinarity without distinguishing between a first or second or third wave, thus allowing all digital disciplines and frameworks within the humanities to thrive in their own right. It seems to me that MDH (2011) is closer to such a version of DH than is DHM (2009).

It should be mentioned that the door is not completely closed to first wave approaches in DH:

> Such a crudely drawn dichotomy [the wave metaphor – KEJ] does not exclude the emotional, even sublime potentiality of the quantitative any more than it excludes embeddings of quantitative analysis within qualitative frameworks. Rather it imagines new couplings and scalings that are facilitated both by new models of research practice and by the availability of new tools and technologies. (DMH 2009, p. 2).

Techniques from CL have indeed started to make their way into other humanistic disciplines. In some cases, these disciplines have already gone digital while, in other cases, it seems that the CL contributions currently serve a basic role in propelling such disciplines or, at least, areas within such disciplines into the digital world. Corpus linguists and computational linguists have also begun to experiment with DH techniques, and one might hope that a fruitful mutual exchange of ideas and experiences will take place between CL and DH.

Below are some examples of transdisciplinary frameworks and approaches that involve techniques and insights associated with CL, illustrating CL's potential for transdisciplinary digital collaboration with other fields in DH and the humanities more generally.

The *Dansk Sprog- og Stilhistorisk Database*, which is a collection of Danish texts in the period 1500-1700, was compiled using CL corpus design techniques and principles and has spawned research into linguistic, genre, and social dimensions in Danish Renaissance songs (e.g., Ruus 2000, 2007; Duncker 2009).

The members of the *Parsing Low-Resource Languages and Domains* project have experimented with crowdsourcing via *Crowdflower* as a method of annotation of data from Twitter (Fromheide et al. 2014; Hovy et al. 2014) as well as with sentiment analysis of data from YouTube (Uryupina et al. 2014), thus both applying and experimenting with techniques typical of CL and operating on data not typically associated with CL but rather with media studies.

Corpus-assisted discourse studies, or CADS, (e.g., Hardt-Mautner 1995; Partington 2004; Gabrielatos & Baker 2008) combines quantitative techniques of CL, such as statistical overviews, with qualitative methods from discourse analysis, such as close-reading of coherent segments of text. Unlike traditional CL, CADS typically focuses of specialized corpora containing texts within a specific discourse genre. This hybrid approach allows for patterning of features of specific discourse types and, at the same time, a closer understanding of the textual and social processes involved.

Similarly, CL techniques have made their way into stylistics, resulting in a branch of stylistics called corpus stylistics in which specialized corpora of text belonging to a literary genre or composed by the same literary author are investigated for stylistic patterns, thus combining the stylistician's interest in literary effects with the corpus linguist's interest in patterns of language use (e.g., Mahlberg 2009).

Finally, Hilpert's (2011, 2012, 2013) work on constructional changes in language is worth mentioning. It draws on CL, construction grammar, and historical linguistics and involves digital methods of analysis and representation of data. Quantitative at heart, Hilpert's work visualizes changes in patterns of use over a number of time periods via motion charts. In print media, motion charts are rendered in a comic-book fashion as series of graphs representing periods in time. In digital environments, they have the form of animated motion graphs, which are generated using the *googleVis* tool in *R*. Not only are *googleVis*-generated motion charts animated (Hilpert 2012), they are also interactive, so that the user may rescale or reorganize the graph, manipulate time, and select specific elements to focus on. This way, a user may be immersed in the chart and even make discoveries that the chart's creator may have missed. Such motion charts are another good example of contemporary CL being compatible with a number of core DH issues. One such issue is dissemination of research. *GoogleVis*-generated motions charts are saved in html code, which means that they may be embedded into webpages and made broadly available within and beyond the research community. This relates to two other core issues: crowd science and network col-

laboration. A multitude of users may interact with a publicly available motion chart and make observations and discoveries themselves. Such discoveries could, then, be collected by the creator of the motion chart via a feedback form embedded in the same webpage. Similarly, members of a network of professional researchers may select individual elements in the chart and each focus on describing and explaining their selected element. The efforts of each network member would, then, amount to a collectively-constructed model of language change pertaining to the constructional element(s) in question.

Concluding remarks

It should be clear from this brief account of some of the basics of CL that it is a branch of empirical linguistic research that has fully embraced digital technology as regards both data and methodology; it has evolved with digital technology and will continue to do so. Yet, CL is on the fringes of contemporary DH, which is itself currently on the fringes of the humanities.

This article has revolved around two related issues – namely, the evolution of CL itself and the relation of CL to DH. Regarding the first issue, we have seen how CL has developed in tandem with computer technology. For instance, standard corpus sizes grew as advances were made in data storage technology, and statistical methods of analysis became increasingly complex with the invention of the PC. Indeed, the digital seems integral to a corpus linguist's scientific identity, and it is interesting to note that, although corpora had been around for a long time, 'corpus linguistics' was not really used in linguistic terminology until the 1960s (Svartvik 2007, p. 15). Its fringe status in DH is, perhaps, puzzling at first sight, but this is due to a number of complexities that reside in both the quantitative nature of CL with its focus on automation and the strict conception of contemporary DH as a qualitative and experiential second wave. It is clear that, in the strict version of DH, CL in its entirety will forever be on the fringes because full inclusion would require CL to abandon its quantitative core. This is very unlikely to happen, given that CL is *per se* a framework within usage-based linguistics. However, we see that hybrid disciplines are evolving that draw on some aspects of CL and combine them with approaches that are more in line with the qualitative and experiential core of DH, such as corpus-assisted discourse studies and corpus stylistics, and there are archival projects underway that involve CL techniques that may foster both traditional linguistic analysis and more transdisciplinary, culturally oriented analysis. We also see that, within the sphere of CL and computational linguistics, DH techniques are being taken into consideration and experimentally applied. In all likelihood, such techniques will ultimately serve the purpose of quantitative analysis, but it is undeniable that a mutual exchange of a scholarly nature has begun between DH and CL.

At the end of the day, the goals of CL are not all that different from those of DH: both seek to shed light on one or more aspects of the human experience, and neither is afraid to explore the opportunities offered by digital technology.

References

Aston, G & L. Burnard (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baker, P., A. Hardie & T. McEnery (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Berez, A. & S. Th. Gries (2009). In defense of corpus-based methods: A behavioral profile analysis of polysemous get in English. In S. Moran, D.S. Tanner & M. Scanlon (Eds.), *Proceedings of the 24th Northwest Linguistics Conference* (pp. 157-166). Seattle, WA: Department of Linguistics.

Biber, C., S. Conrad & R. Reppen (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bick, E. (2014a). *CorpusEye*. http://corp.hum.sdu.dk/.

Bick, E. (2014b). *Machine Analysis*. Retrieved January 7, 2014, from http://beta.visl.sdu.dk/visl/en/parsing/automatic/.

Bird, S., E. Klein & E. Loper (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.

Burdick, A., J. Drucker, P. Lunenfeld, T. Presner & J. Schnap, (2012). *Digital_Humanities*. Cambridge, Mass: MIT Press.

Cohen, Patricia (2010 16 November). Humanities scholars embrace digital technology. *The New York Times* [online version]. Retrieved December 28, 2013, from http://www.nytimes.com/2010/11/17/arts/17digital.html.

Davies, M. (2014). *corpus.buy.edu*. http://corpus.byu.edu/.

Digital Humanities Manifesto 2.0 (= DHM) (2011). Retrieved May 1, 2013, from http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.

Duncker, D.M. (2009). Faste forbindelser som genrekonstituenter: En undersøgelse i de danske visehåndskrifter før 1591. *Danske Studier* 104(9/8), 5-37.

Elmelund, R. (2014 3 January). Det er sådan noget litterater har drømt om siden munkene. *Dagbladet Information* [online version]. Retrieved January 6, 2014, from http://www.information.dk/483617.

Fillmore, C. (1992), "Corpus linguistics" or "computer-aided armchair inguistics". In J. Svartvik (Ed.), *Directions Proceedings of Nobel Symposium 82, Stockholm 4-8* (pp. 35-60). Berlin: Mouton de Gruyter.

Francis, W.N. (1982). Problems of assembling and computerizing large corpora. In S. Johansson (Ed.), *Computer Corpora in English Language Research* (pp. 7-24). Bergen: Norwegian Computing Centre for the Humanities.

Fromheide, H.,D. Hovy & A. Søgaard (2014). Crowdsourcing and annotating NER for Twitter #drift. Paper presented at *Language Resources and Evaluation Conference*, May 26-31, 2014, Reykjavik, Iceland.

Gabrielatos, C. & P. Baker (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics* 36(1), 5-38.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-Based Approach* (pp. 30-41), London: Longman.

Gries, S. Th. (2009a). *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge.

Gries, S. Th. (2009b). *Statistics for Linguistics with R: A Practical Introduction*. Berlin: Mouton de Gruyter.

Hardt-Mautner, G. (1995) *"Only Connect." Critical Discourse Analysis and Corpus Linguistics, UCREL Technical Paper 6*. Lancaster: Lancaster University.

Hammond, M. (2003). *Programming for Linguists: Perl for Language Researchers*. Oxford: Blackwell.

Hilpert, M. (2011). Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16(4), 435-461.

Hilpert, M. (2012). *Motion Chart Resource Page*. Retrieved June 26, 2014, from http://members.unine.ch/martin.hilpert/motion.html.

Hilpert, M. (2013). *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.

Hockey, S. (2004) History of humanities computing. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 3-19). Oxford: Blackwell.

Hovy, D., B. Plank & A. Søgaard (2014). Experiments with a crowd-sourced re-annotation of a POS tagging dataset. Paper presented at *The 52nd Annual Meeting of the Association for Computational Linguistics*, June 22-27, 2014, Baltimore, MD, USA.

Ide, Nancy (2004). Preparation and analysis of linguistic corpora. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 289-305). Oxford: Blackwell.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Kirk, J.M. (1996a). Corpora and discourse: transcription, annotation, and presentation. In C.E. Percy, C.F. Meyer & I. Lancashire (Eds.), *Synchronic Corpus Linguistics: Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora, Toronto 1995* (pp. 263-278). Amsterdam: Rodopi.

Kirk, J.M. (1996b). Review of K. Aijmer & B. Altenberg (Eds.) (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. *Journal of English Linguistics*, 24, 250-258.

Kirk, J.M. (1999). Using SARA on the British National Corpus. *AAA: Arbeiten aus Anglistik und Amerikanistik*, 24(1), 35-51.

Kirk, J.M. (2013). *SPICE-Ireland*. Retrieved January 9, 2014, from http://www.johnmkirk.co.uk/cgi-bin/generic?instanceID=11.

Leech G. (1987). General introduction. In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-Based Approach* (pp. 1-15), London: Longman.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer and B. Altenberg (Eds.) *English Corpus Linguistics* (pp. 8-29). London: Longman.

Lockers, M.L (2014). *Text Analysis with R for Students of Literature*. Heidelberg: Springer.

de Lopez, K.L. (2013). Clause-final *man* in Tyneside English. In G. Andersen & K. Bech (Eds.), *English Corpus Linguistics: Variation in Time, Space and Genre* (pp. 113-162). Amsterdam: Rodopi.

Manifesto for the Digital Humanities (= MDH) (2011). Retrieved May 1, 2013, from http://tcp.hypotheses.org/411.

McEnery, T. & A. Hardie (2012). *Corpus Linguistics*. Cambridge: Cambridge University Press.

McEnery, T. & A. Wilson (2001). *Corpus Linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T., J.P. Baker & A. Hardie (2000). Assessing claims about language use with corpus data: Swearing and abuse. In J.M. Kirk (Ed.), *Corpora Galore: Analyses and Techniques in Describing English* (pp. 45-55). Amsterdam: Rodopi.

MacWhinney, B. (2014). *The CHILDES Project: Tools for Analyzing Talk – Electronic Version. Part 2: The CLAN Programs*. Retrieved January 6, 2014, from http://childes.psy.cmu.edu/manuals/CLAN.pdf.

Mahlberg, M. (2009). Corpus stylistics and the Pickwickian watering-pot. In P. Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 47-63). London: Continuum.

Minugh, D.C. (2000). *You people use such weird expressions*: The frequency of idioms in newspaper CDs as corpora. In J.M. Kirk (Ed.), *Corpora Galore: Analyses and Techniques in Describing English* (pp. 57-71). Amsterdam: Rodopi.

Ooi, V.B. (2000). Asian or Western realities? Collocations in Singaporean-Malaysian English. In J.M. Kirk (Ed.), *Corpora Galore: Analyses and Techniques in Describing English* (pp. 73-89). Amsterdam: Rodopi.

Owens, T. (2011). Defining data for humanists: Text, artifact, information or evidence? *Journal of Digital Humanities* 1(1), 6-8.

Partington, A. (2004). Corpora and discourse, a most congruous beast. In A. Partington, J. Morley & L. Haarman (Eds.), *Corpora and Discourse* (pp. 11-20). Bern: Peter Lang.

Ruus, H. (2007). O Danmarck, o Danmarck, o Danemarck: Gud, konge og fædreland anno 1588. In H. Blicher, M.K. Jørgensen, & M. Akhøj Nielsen (Eds.), *Tænkesedler: 20 Fortællinger af Fædrelandets Litteraturhistorie: Festskrift til Flemming Lundgreen-Nielsen* (pp. 73-87). Copenhagen: C.A. Reitzel.

Ruus, H. (2000). Visernes top-18: Populære viser i overleveringen før 1591. In F. Lundgreen-Nielsen & H. Ruus (Eds.), *Svøbt i Mår: Dansk Folkevisekultur 1550-1700*, vol. 2. (pp. 11-38). Copenhagen: C.A. Reitzel.

Sinclair, J. (1996). Preliminary recommendations on corpus typology. Retrieved January 2, 2014, from http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

The Stanford Natural Language Processing Group (= SNLPG) (2014). *Stanford Log-Linear Part-of-Speech Tagger*, v. 3.3.1. Retrieved January 7, 2014, from http://nlp.stanford.edu/downloads/tagger.shtml.

Steen, G., L. Dorst, B. Herrmann, A. Kaal & T. Krennmayr (2012). *VU Amsterdam Metaphor Corpus Online*. Retrieved January 16, 2014, from http://www2.let.vu.nl/oz/metaphorlab/metcor/search/showPage.php?page=start.

Svartvik, J. (2007). Corpus linguistics 25+ years on. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years On* (pp. 11-25). Amsterdam: Rodopi.

Svartvik, J. & R. Quirk (1980). *A Corpus of English Conversation*. Lund: CWK Gleerup.

Svensson, P. (2010). The landscape of digital humanities. *DHQ: Digital Humanities Quarterly*, 4(1).

Tasse, D. & N. Smith (2008). *Carnegie Mellon University Recipe Database.* http://www.ark.cs.cmu.edu/CURD/.

University Centre for Computer Corpus Research on Language (= UCREL) (2014). *CLAWS Part-of-Speech Tagger for English*. Retrieved January 7, 2013, from http://ucrel.lancs.ac.uk/claws/.

Uryupina, O., B. Plank, A. Severyn, A. Rotondi & A. Moschitti (2014). SenTube: A corpus for sentiment analysis on Youtube social media. Paper presented at *Language Resources and Evaluation Conference*, May 26-31, 2014, Reykjavik, Iceland.

Kim Ebensgaard Jensen, PhD
Associate Professor
Department of Culture and Global Studies
Aalborg University, Denmark
kim@cgs.aau.dk