Sign Language Corpus Creation:
A Digital Humanities Ethnography

Ernst D. Thoutenhoofd[1]

1. Abstract

In this paper I will develop a social science perspective on the construction of sign language corpora for digital humanities scholarship. At this point in time sign language corpora are emergent, although sign language repositories (some including metadata standardised for sign languages) do already exist. What content to include in language corpora is generally discussed with respect to the issues of representativeness and size, the available resources, the scope for comparison with existing datasets, and the linguistic reasons for building a corpus. Issues in the construction and management of language corpora are therefore mainly points of discussion within linguistic circles. This paper aims to broaden the context of language corpus creation, by taking as a starting point not corpus linguistics, but digital humanities—defined as the inclusive study of dynamic interaction between people, their heritage, institutions, and new technology.

This conceptual broadening of a language corpus to include the ceaseless making and remaking of digital artefacts by users and reproducers calls for methodical attention to the nature of corpus creation as ongoing cultural performance. With this paper I would like to propose that the complex hermeneutics of digital humanities language corpora calls for a theory of corpus linguistic practice (an account of *praxis*) and an emphasis on reflexive ethnographic methods.

2. Introduction

Contrary to popular belief, there is great variation, akin to variation within and among spoken languages, within and among naturally occurring sign languages across nations and deaf communities. Following five decades of informant-based description of sign language grammar and lexis, one of the tasks that confronts sign linguists following the spread of quantitative, empirical methods of language description in the parent-discipline of mainstream linguistics is to document the various sign languages with data quantities that will be sufficient for frequency-based types of knowledge claims; that is (in effect) to follow the mainstream linguistic community in re-grouping *vis à vis* new technology supported empiricism.

This projected effort presents the small sign linguistic community with considerable linguistic, technological and organisational challenges (Morford and MacFarlane 2003). It is thought that establishing large sign language corpora will involve creating movie-based multimedia datasets, with associated standards for coding schemes and metadata description (Crasborn and Hancke 2003; Johnston and Crasborn 2006). In this corpus-design type, a sign language corpus is a multimedia

[1] Virtual Knowledge Studio for the Humanities and Social Sciences, Royal Netherlands Academy of Arts and Sciences
e-mail: ernst.thoutenhoofd@vks.knaw.nl

repository that contains digital movies of signing Deaf people, along with annotations that facilitate searching through the data, and which are placed in layered annotation tiers. Such corpora are currently being developed for the future purpose of empirical linguistic analysis of national sign languages, for example with the aid of the XML-based ELAN annotation software developed by the Max Planck Institute of Psycholinguistics in Nijmegen[2], along with associated ISLE Metadata Initiative (IMDI) standards for describing multimedia and multi-modal language content (Broeder and Wittenburg 2006).

As with spoken language corpora, two main types of data are distinguished in creating a sign language corpus: movie data that are imported from existing video libraries or digital movie repositories, and movie data that are collected through interviews, dialogues between language users, or other organised forms of language elicitation. Whichever method is chosen, the recording of sign language on time-based media is critical (like the sound recordings that are needed to construct speech corpora), especially since no writing system for encoding sign language discourse as text is in widespread use. All this means that sign language corpora will consist of substantial movie-collections of particular, clearly identifiable deaf individuals. There is therefore an important sense in which future national sign language corpora – even more so where they are designed to be representative in some way – will take on the character of a 'family album' of the national deaf community at a particular time (for the significance of family albums as performance of social organisation, see Chaney 1993 chapter 4; Spence and Holland 1991; Thoutenhoofd 1998). For this reason the construction of a sign language corpus is not a matter of linguistics per sé but a matter that will interest and concern social scientists and humanities scholars across, and hopefully beyond, the fields of deaf studies and sign linguistics.

3. Digital humanities as scholarship practice

I focus on digital humanities for the specific reason that the mediating qualities of new technology (that which Hans Ulrich Gumbrecht has referred to as the 'special effects' of new technology, 2004:xv) connect with humanities and social science practice in highly complex ways (Wouters et al. 2007). A core concern in the digital or 'cyber' infrastructuring of the sciences is achieving data interoperability (Ribes et al. 2005; Edwards et al. 2007). While interoperability among language corpora may not yet be a matter of organised concern in corpus linguistics (Fry 2004), collective effort has been devoted to the development of quality standards and the standardisation of data and metadata descriptions. Given the cultural situatedness and specificity of sign language corpus construction work in combination with the historically persistent liminal status of deaf people, social science interests arise here with greater urgency. My interest is in the interplay of knowledge practices and the standardising tendencies that mark data construction practices in digital linguistic scholarship.

My digital humanities focus therefore takes account of implications of new technology, in particular the supposition that interdisciplinarity and data-accessibility imply broad opportunities for participation and collaboration (extending into data-collection, data-sharing and lay/expert collaboration on data-analysis). For my part, situating language corpora in a social structure mediated by new technology entails

[2] http://www.lat-mpi.eu/tools/elan/

prior recognition of the inherent instability of the corpus as digital object, caused by forms of plurality with respect to what and how language data come to mean, for whom, at which time, and to what ends.

Despite my broad definition of digital humanities as research that focuses on individual and institutional practice and digital phenomena, I use digital humanities as a concept for knowledge practices (and expertise claims) in the humanities that draw upon digital data-objects and technologically mediated forms of communication. More particularly, the digital humanities at this time point to various transformations in material conditions, the *infrastructure*, of humanities scholarship as scholars get to grips with information technology in the collective body of work. It seems important to consider how we (by this I refer to scholars in science and technology studies, the philosophy of science, and the sociology of knowledge) might understand the various ways in which humanities scholars individually, collectively, and more formally as disciplinarians make sense of that transformation.

The emergence of sign language corpora provides a useful and perhaps unique case study for the transformation of traditional scholarship to so-called e-research, for three reasons: (1) the clear sub-disciplinary status of corpus sign linguistics in relation to mainstream corpus linguistics as parent discipline, (2) the exemplars and associated theories about the turn to corpus-based empiricism, and corresponding work practices, that already exist in mainstream linguistics but not yet in sign linguistics, and (3) working with sign languages and sign language communities involves patently different sorts of challenges for establishing corpora, so that new solutions need to be found. The various intersections of communities and practices that are entailed in these three reasons account for my own position as a social scientist in relation to corpus sign linguistics: for me, corpus sign linguistic practice is itself an object of study.

I am aware that my definition of digital humanities contrasts with understanding the digital humanities as a field, and contrasts even more with understanding them as a discipline (Schreibman et al. 2004:xxiii). The key difference between considering digital humanities as a process of transformation in relation to an emergent infrastructure and considering digital humanities as the completed integration of technology into disciplinary organisation is that I see the digital humanities not as an expanding collection of locations where practice and computation intersect and subsequently homogenise around a unified form of knowledge valorisation. Instead I insist on creative but at the same time historically embedded forms of adjustment and resistance within diverse scholarship practices to changing social and material conditions, and therein the construction of new technology through practice, as digital humanities. In my view the process of ongoing negotiation that is implied has deeper implications for how humanities scholars understand what scholarship is and how it corresponds with the world.

My definition is not much closer to conceiving of humanities computing as a practice of representation based on a set of ontological commitments (Unsworth 2002). Similarly, Galison's work points to a dual role of computers: as substitutes for human labour, and as substitutes for nature, in the sense that many scientific techniques involve forms of modeling or simulation (Galison 1997). Galison therefore sees computers mainly as linguistic devices that interpret between human and computational forms of representation. The problem that I see with both these assessments is that all answers to the question why we worry about being able to express humanities knowledge in terms that are tractable to computation (such as the above stated need for interoperability), would lead to the upkeep of a representation or

substitution discourse. They also entail an implied divide between a universalist 'world of data' on the one hand, and particular forms of scholarship on the other. I prefer instead to tackle the question of how the connection between the world and knowledge that is mediated through computational techniques is to be construed. With that question I hope to avoid, at one and the same time, the pitfalls of a discursive analysis that leads to relativism and a Cartesian division between worldly objects and cognate subjects in constructing forms of knowledge-making practice.

At the same time all this implies a central hermeneutic dimension to digital humanities practice that points to a departure from purely interpretative traditions. Hermeneutics in this sense provide the underlying principles or 'rules' for understanding the construction of language corpora as cultural objects by corpus linguistic scholarship. The ethnographic position that is implied in that clarification will be addressed shortly, but first this hermeneutic dimension itself needs to be stated. Accounting for the hermeneutic dimension that I see will also involve discussion of epistemology. Epistemology, the study of implicit rules that evade being constituted or construed as rules but around which a community coheres through practice, is highly central to the organisation of scholarship practice (Knorr-Cetina 1999) and therefore epistemics provide both impetus for and commitment to organised action in corpus linguistics as discipline. Hermeneutics and epistemology are therefore two key elements in the construction of theory about digital humanities as new infrastructure for scholarly practice. I will turn to both before moving on to the role of ethnography in creating sign language corpora.

3. Hermeneutics in corpus linguistics

There are many accounts of the role of hermeneutics in scholarship. Because it centres on a sociological interpretation of scientific methods, a useful version for my purpose is in the revised edition of *New rules of sociological method* by the sociologist Anthony Giddens (1993). In his book Giddens makes an explicit connection between hermeneutics and the social nature of scientific practice as developed in the science and technology studies (or STS) literature. Giddens points to the ethnography of laboratory work undertaken by Karin Knorr-Cetina. A well-known STS scholar, Knorr-Cetina noted that discussions over methodological rules in social science are renewed with every appearance of a new conception of social life, whereby the only constant is the maintenance of a declared contrast with the standard set of rules that are associated with the hypothetico-deductive model of method in the natural sciences: replicable testing of hypotheses against a declared theory by empirical means. By contrast, Knorr-Cetina's own empirical observations of scientific practice led her to conclude that natural science 'is grounded in the same kind of situational logic […] which we are used to associate with the symbolic and interactional character of the social world' (Knorr-Cetina 1981:336).

This conclusion is partly based on the finding that what counts as a relevant observation to evaluate a theory can only be determined by calling on certain assumptions about the relationship between knowledge and reality. In particular, the natural sciences are characterised by a popularly supported and internally unchallenged assumption that a key difference exists between the social and natural world. While the social world is deemed to be symbolically encoded, the natural world is considered to not constitute itself as meaningful in any way; nature (also in the form of the original Latin meaning of data, 'that is given') does not talk back to

the scientist. In the philosophy of science this conception of nature as constituting uninterpreted matter is known as the single hermeneutic of scientific practice.

The social sciences and humanities, by contrast, tend towards a double hermeneutic. To be able to describe what a person is doing in any given context – such as speak, or listen – implies understanding what that person knows and applies in the constitution of their activities (Giddens 1993:13). It is therefore not at all unreasonable to assume that both humanities and social science methods are grounded in language, since language would constitute the sole means through which people can collectively construct and understand the symbolic nature of the world. Interpretation, the methodological stance that is associated with this double hermeneutic, is thereby essentially linguistic in nature and constitutes a paradigmatic form of humanities praxis.

Much of empirical humanities and social science methodology therefore involves language as data: discourse analysis, ethnography, participant observation, interviews and surveys, and of course language corpora, all collect information on the symbolic, collective construction of the world through language. However, as scholars in STS are wont to point out, a double hermeneutic that involves language applies not only to a collective understanding of the social world, but equally holds in collective understanding of the natural world (e.g. Gilbert and Mulkay 1984). In other words, nature is itself also socially constructed and grounded, in the language and practice of the natural sciences. The net result is a hermeneutic circle in scholarship, whereby to understand the collective construction of both natural and social phenomena in science and scholarship requires understanding their social construction in the language and practice of academic effort – a theory of practice.

To my mind, the hermeneutic dimension of corpus linguistic scholarship therefore includes at least the following three aspects, which would need to be accounted for in a sociology of corpus linguistic knowledge: (1) at the conceptual level of research practice, the reconstruction of attested language as data implies a particular hermeneutic stance on the relationship between mass data as the object of linguistic scholarship, and language as a social but supposedly naturally occurring practice; (2) at the level of method, the turn to empiricism in corpus linguistics implies a hermeneutic circle, whereby corpus linguists construct empirical data, through data collection, annotation and coding, as a subset of natural, unmediated references to language as praxis (the notion of attested forms), whereas language praxis is reconstructed through analyses of attested forms as individual language manifestations; and (3) at the latter meta-level of interpretation – a scientific reconstruction of language praxis – a double-hermeneutic applies: a reflexive position is taken on how corpus linguists construct what knowledge of language as social practice pertains from the user's perspective (e.g. Austin 1972, Searle 1969, Halliday 1978). This hermeneutic dimension in corpus linguistics points to the complex social reconstruction of practice in corpus linguistic language.

4. Epistemics in corpus linguistics

A theory of practice is an account wherein scholars may account for the different orders of knowledge that derive from the systematic and institutionalised analysis (and reproduction, Foucault 1970) of social phenomena, the immediate character of the world, and social order. In order to be able to discriminate at a conceptual, abstract level between individual, collective, and institutional forms of knowledge, the French

sociologist Pierre Bourdieu used phenomenological knowledge as a concept to refer to the sorts of everyday understanding that help individuals through daily life – the truth of primary experience – whereas he referred to the reinterpretation of that primary experience (both the practice of science and the science of practice) as dialectics (Bourdieu 1977:3). The intermediate, collective form of knowledge is termed objectivist, by which Bourdieu meant to refer to a common-sense integration of phenomenological knowledge in shared understandings associated with truths of sorts.

What prompts Giddens, Bourdieu and other sociologists in setting up these social knowledge hierarchies is a concern with the character of agency, structure, and material and digital knowledge-objects in the constitution of, for example, corpus linguistics. For Giddens this understanding calls for paying attention to logical and empirical elements that constitute a double hermeneutic, in which the logical implies social science concepts that are themselves constitutively produced and appropriated in everyday action (social science knowledge is often conceived to reflect 'common sense' knowledge), while the empirical of a double hermeneutic addresses institutional reflexivity, the many ways in which our knowledge is incorporated into, and reflects, social order. Bourdieu might refer to this latter reflexive character of the double hermeneutic as *habitus*, a structured system of durable dispositions that produce a regular and recognisable patterning of social practice (Bourdieu 1977:72).

A final point that is important in Giddens' interpretation is that natural science's deep-seated belief in a single hermeneutic (a direct relationship between knowledge claims and natural phenomena) rests on a social order that is favourable to institutional empirical validation of knowledge claims. So hermeneutics becomes more visible in epistemological particularities, as concerns with what knowledge is and how it is constructed and validated, expressed in practice. I will therefore suppose that epistemology is where hermeneutic dimensions most strongly connect with practice, and give some examples of this.

First, the scientific empirical approaches that have entered disciplines such as linguistics (Hajič 2004) and philology (Talstra 1980; van Peursen 2002) through computational techniques trouble the traditional double hermeneutic dimension of scholarship in these disciplines, because large-scale collected data especially may be assumed to stand in an unmediated, direct relationship to the object of study. In corpus linguistics epistemology, attested forms are considered elements of a larger entity, so that reference may be made to a language 'as a whole' (e.g. Beaugrande 2002). This epistemology points to a single hermeneutic dimension in the construction of language as a singular entity, whereby at the very least the community of linguists would understand a language to be, or buy into it being, an object that exists in a definite form outwith the collective knowledge that applies to it; indeed exactly what would hold for mass among the community of physicists. This construction may lead to speculations of how large a language is relative to the size of a given corpus, and to analyses whereby the corpus may be considered a sufficient evidence-base with respect to knowledge claims (e.g. Widdowson 2000; Stubbs 2001).

The invocation of a single hermeneutic points to the fact that some of the techniques and practices in humanities computing originated in the natural sciences – such as distributed data-hosting, modeling, simulation and visualisation techniques, collaboratories, and things like grid-computing and the use of global positioning data, which is now also used in dialect mapping. Although linguistics (Busa 2004) and social sciences (Agar 2006) may perhaps be considered early adopters, the association with a single hermeneutic remains visible in both positivist and empirical orientations

in corpus construction. With respect to the size of a language for example, there are clear epistemic parallels in the consideration of invariant mass in physics, which refers to a quantity that also does not depend on an observer or an inertial frame used for observation, and to which we also have access exclusively in 'attested form'.

Second, the exact inverse is also at play. The socio-cultural implications of new technology challenge the strict separation of the laboratory from the outside world, and the notion of generally accessible sign language corpora that contain elicited discourse by deaf people recorded in a studio serves as a good example of the emergence of various kinds of contact zone between academic and non-academic fields of cultural performance and production. Across all forms of scholarship, one consequence of this blurring between the university and the worlds it studies seems to be that theory itself becomes a form of situated action (Gilman 2004), involving constant hermeneutic adjustment alongside opportunistic epistemic drift between *Erklärende* (empirical) and *Verstehende* (interpretative) methodologies. Within these various contact zones, the justification of expertise claims and the ordering of knowledge claims into coherent narratives and perspectives that are '*from* somewhere' become much more of a process of delicate, temporally unstable series of adjustments among all participants to a corpus as multiple digital object.

This phenomenon has been beautifully described in a study by the Dutch philosopher Annemarie Mol. Following a drawn-out period of ethnographic fieldwork, the multiple simultaneous conceptualisations and enactments in practice of a particular disease (arterosclerosis) by patients and various kinds of health and medical practitioners leads her to the construction of the human body as a single entity that is nevertheless at one and the same time multiple (Mol 2002). Here we are at the point of digress from epistemology into ontology, the study of existence or being in metaphysics, where the question becomes through what sort of practice the claim that a language corpus makes reference to 'a language' is performed, and to what ends. The study also points towards the central role of ethnographic methodology in affording the promise of integrating multiple perspectives and multiple practices into a narrative form that is, if perhaps not exactly holistic, then at least plural and dynamic with respect to positions taken by actors towards a knowledge object.

Third, the attention in science and technology studies has turned towards analyses of computing in the humanities and social sciences, pointing to the routinisation of scientific work and an associated division of labour caused by increasing dependency on technologists, who provide academic expertise that in itself has little connection with interpretative scholarship (Agar 2006). Forms of humanities computing that ostensibly aim at interpretative grounding may therefore nevertheless risk letting empiricism in through contributions from the back of the shop. In this sense, the conceptual turf that STS scholars seem to occupy can be thought of as the introduction of a circular, double hermeneutic into areas of scientific knowledge production that might under normal circumstances default (in practice if not also in principle) to a single hermeneutic in which both a self-evident differentiation and a stable relationship are assumed to hold between data and subject matter. Hence the tendency of STS to add additional layers of meaning to the *modus operandi* of science. While this is certainly not without value for a theory of corpus linguistic practice, if STS boils down to the reasonable insistence that (contrary to Bourdieu's more generally stated claim based on analysis of Kabyle kinship structure 1977:36) not all is implied when the principles of production are extracted from the final product, then the programme of STS is itself exhausted at the point of turning that

insistence onto itself (Fuller 2000); this is merely accentuated when a key STS scholar points to sociology for overstating 'the social' as focus of analysis (Latour 2005) at the same time as overstating the relevance of networks in social theory. So what is needed is a form of complexity analysis that avoids circularity, such as perhaps the application of complexity theory in the social sciences (Byrne 1998).

And fourth, the organisation of technology-bearing labour processes itself produces representations of technical product outcomes that have been shown to be used to justify investments in technological infrastructure (Vann and Bowker 2006). Attention to how commitment to future product development is enacted in effort and encoded in epistemic discourse can also help us understand how the call for 'interoperability' ties local or national research projects into the will to produce global research infrastructures. Within corpus linguistics, such a pragmatic politics of representation connects for example with developments at the European level such as the CLARIN[3] initiative and global ones such as ISLE[4] and DOBES[5]. But it would also seem to tie in with frequently made calls for the recording and archiving of disappearing or changing linguistic phenomena, in an effort to keep a technical representation of those phenomena in store. This form of argument derives from epistemological claims about the value of linguistic research within a historical framework of language and public interest, and is often made in particular with reference to minority languages (DOBES, McEnery 2001) including therefore most notably sign languages (van Hout 2006).

With respect to all four of my examples, my proposal is that a theory of corpus sign linguistic practice calls for an ethnographic approach. By ethnography I refer to a methodology that focuses on the plural and dynamic construction of knowledge objects, and on the representation of practices practiced; a methodology that fits constructivist and discursive academic traditions, that is itself grounded in social science and humanities scholarship, that is open to plurality of knowledge constructions and agency that changes nature according to the perspective that is taken, and that has demonstrable historical connections with linguistics as fieldwork in a tradition that contrasted most sharply with theoretical or 'homework' linguistics (Beaugrande 1996, 2002; Jenudd and Neustupny 1991). This last deep-rooted division in belief in specific types of scholarly practice stands in a causal relationship with ongoing transformation in corpus linguistics as digital humanities.

Although it would seem an apparent double contrast (of opposing means mobilised to opposite ends), in my mind the empiricism that attaches – I will call it the will to naturalise – to the construction of very large datasets, at the same time re-introduces ethnography as alive to the dimension of interpretative reflexivity, the will to socialise, that marks the humanities. What my hermeneutic interpretation calls for is understanding how field relations entwine actors and corpus, material and digital, object and agency (create a 'mess'), as a constant moving towards hermeneutic circularity within theory and data that is disrupted only by ongoing articulation of corpus linguistics practice as praxis.

[3] CLARIN: Common Language Resources and Technology Infrastructure, see www.mpi.nl/clarin/

[4] ISLE: International Standards for Language Engineering, see www.ilc.cnr.it/EAGLES/isle/ISLE_Home_Page.htm

[5] DOBES: *Dokumentation Bedrohter Sprachen* (Documentation of Endangered Languages, see www.mpi.nl/DOBES/

5. Ethnography of corpus linguistics

With respect to how corpus linguistics is organising itself in relation to new technology, the STS scholar Jennifer Fry (2004, 2006a) has undertaken an in-depth analysis of the appropriation of new technology within the disciplinary culture of corpus linguistics, and the extent to which the discipline coheres around infrastructural resources in support of corpus work. The focus is not on technology but on the reporting of developments in the field by corpus linguistic scholars. The study therefore has a distinctly ethnographic character, including interviews with corpus linguists and a discourse analysis of public documents.

On the basis of those data and similar data relating to a comparison-discipline (in this case high energy physics), corpus linguistics is categorised according to a classificatory scheme developed by the sociologist Whitley (1984). The classification aims at ascertaining levels of 'dependency' among scholars, and levels of 'uncertainty' in their practice. In particular, greater strategic dependence among community members and less technical uncertainty in the work to be undertaken collectively accords in Fry's analysis with coherence around a unifying infrastructural arrangement with respect to new technology. This actually amounts to the hypothesis that unified goals and universal means correlate with strong infrastructural organisation, which is of course paradigmatic of a rationalist engineering and natural science conception of how cyberinfrastructure and e-science tie up the means and ends of new technology. Fry concludes that low levels of strategic dependency (scholars pursuing individual research targets) and high levels of uncertainty (i.e., a spread of techniques and technologies) explain the 'limited success within the corpus-based linguistics community in developing field-wide social and technical standards and protocols for computer-mediated collaborative work' (Fry 2004:316).

However, the results that obtain would seem to depend on the type of comparison that is made, and the categorisation that is imposed prior to analysis reduces disciplinary differences in for example epistemics to a narrow subset of formalised variables. What results is not an analysis grounded in exploration but the testing of a generally stated hypothesis (Whitley's) in discourse, which would seem contrary to an ethnographic privileging of social phenomena over theory, and unlikely to fail. Stated in STS terms, the case entails a bias with respect to the symmetry principle in the methodological rules of STS. Another aspect of the analysis is that it utilises actor-network theory framework of interpretation, which does not take account of structure beyond the explicit network that is described (because it treats the network itself as sufficient to a social analysis). While the research is ethnographic with respect to the collection of evidence that is circulating 'in the field' of corpus linguistics, the analysis does not aim to provide a holistic account of language corpora as part of a wider field of cultural production. Yet an approach that transcends the more strictly academic nature of corpus linguistics as professional practice is necessary if an account of the 'cultural identity' (Fry 2004:303) of the corpus linguistic community is attempted that would locate it with reference to those whose language the corpus contains.

In a subsequently published element of the study Fry focussed on the concept of websphere (Schneider and Foot 2002), a hyperlink analysis of in- and out-links to and from key websites in the discipline that are taken to give an indication of the qualities of the information network and geographic reach associated with that website. This study focuses on the website of the Corpus of Spoken Dutch (CGN), a national corpus project based at Nijmegen University in the Netherlands. Here the

analysis focuses more on Bourdieu's highly specific definition of a 'field' as fluid entities that are structured by an internal dynamics of competition (Bourdieu 1988), so that the already mentioned focus on the academic organisation of practice is retained. The concept of websphere as developed in Steve Schneider and Kirstin Foot's analysis of website activity during the 2000 American national election does begin to address the role of new technology in transforming cultural practice and the conduct of research in a digital world, and that element is also visible in Jennifer Fry's analysis. Fry's analysis ends with the comment that the combination of ethnographic methods (interviews and discourse analysis) and web crawling technology enabled 'a particular augmentation of the websphere concept that made traces of the intellectual and social identity of academic specialist fields on the web more visible' (Fry 2006b).

A quick sift through the in- and out-links associated with two different digital sign language objects may help to clarify the considerations so far. In June 2007 I compared two websites that are based on the public availability of sign language, the academic-oriented ECHO pilot corpus of Dutch, Swedish, German and British sign language movies based at the University of Nijmegen in the Netherlands[6], and the 'MobileSign'[7] website that is part of the Centre for Deaf Studies at Bristol University in the UK, a website that connects a digital corpus of practically oriented British Sign Language/English translations to users' mobile phones. A simple count of in-links to the two websites via Google connected the ECHO corpus with four in-links: a contribution made to one of the ECHO workshops, a link to a Ministry of Education public information resource called *Kennislink* (Knowledge link), a linguistics discussion list, and the commercial portal called Answers.com, where an explanation was offered of the IMDI metadata scheme used with the corpus.

The MobileSign website produced 313 in-links, the vast majority of which were sign movies that were connected to the website via in-links, which demonstrates one of the weaknesses of this search-type. However, a substantial number of remaining in-links tended to be associated with tech-sites that found interest in the innovatory character of the website as a call-up dictionary of sign language, while other in-links were from community-based pages and blogs with URLs such as 'Stone Deaf Pilots', 'Deafbiz-com' and 'Taubenslag.de'.

The more recent webpage associated with a current nationally funded sign language corpus project, the *Corpus Nederlandse Gebarentaal*[8] also based at Nijmegen University produced twenty-nine in-links that distributed functionally as news-reporting, academic institutions, the national funding council, and a few community blogs. Since the corpus itself is not yet online, the website only contains information about the project at this stage. What this absolutely off the cuff hyperlink analysis points to is that a predictable association exists between the content of a website, its subject matter, its connection of audience and function, and its outward reception. I am much less convinced that hyperlink analyses of this sort can stand as indicators of any form of cultural or social association in anything more than a sense that points to a superficially imposed (that is, numerical) ordering of meaning-associations standing in for a dialectical account of the performance of meaning in practice. But that it will be immensely valuable to develop sophistication in ways to operationalise notions such as websphere, in order to account for the special effects of

[6] www.let.kun.nl/sign-lang/echo

[7] www.mobilesign.org

[8] http://www.let.kun.nl/sign-lang/corpusngt/public/home.html

new technology on the cultural production and performance of digital objects such as sign language corpora, there I have no doubt.

6. Conclusion

A social theory of corpus linguistic practice that pays attention to the special contribution of sign language corpora would in my framework of analysis entail three objectives, since the participatory nature of sign language corpora calls (1) for attention to both the academic institutional context (as mediated by the new technology that is used for corpus construction and academic practice) but also, and perhaps more importantly, (2) to the cultural reconstruction and appropriation of the corpus as digital cultural artefact in contemporary deaf culture, and in particular with respect to those deaf individuals who contribute to the corpus content; and (3), both these contexts involve access and contact in a substantial part through digital means and to sign language corpora as digital object, and so the ethnography would equally be strengthened by working with ideas that are being elaborated with respect to digital environments through virtual ethnography (Beaulieu 2004, Hine 2000); these include ideas about ethnographic work on new scholarship infrastructures (Hine 2006).

Since these three objectives also call attention to the politics of the representation of action, addressing the evolution of corpus sign linguistics as a digital humanities field implies a politicised ethnography that, as Teun Zuiderent-Jerak has put it in an ethnographic account of intervention in a health policy implementation project, recognises that any research site, such as a sign language corpus that is located at one and the same time within a scholarly community, within a language community, in a socio-cultural and political context and on a website, has a pluralistic dimension that enables actors to develop different 'views from somewhere' (2002:59; see also Mol 2002 who calls for a political ontology of practice) and through which plural constructions of a language corpus co-exist as single cultural product.

References

Agar, J. (2006) What difference did computers make?, in *Social Studies of Science.* 36(6): 869–907.

Austin, J.L. (1971) *How to do things with words*. Oxford, England: Oxford University Press.

Beaugrande, R. de (1996) The 'pragmatics' of doing language science: The 'warrant' for large-corpus linguistics, in *Journal of Pragmatics*. 25: 503–535.

Beaugrande, R. de (2002) Descriptive linguistics at the Millenium: Corpus data as authentic language. *Journal of Language and Linguistics*. 1(2): 91–131.

Beaulieu, A. (2004) Mediating ethnography: Objectivity and the making of ethnographies of the internet. *Social Epistemology*. 18(2–3):139–163.

Bourdieu, P. (1977) *Outline of a theory of practiceI* (transl. Nice, R.). Cambridge, England: Cambridge University Press.

Bourdieu, P. (1988) *Homo Academicus* (transl. Collier, P.). Stanford, US: Stanford University Press.

Broeder, D. and Wittenburg, P. (2006) The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*. 1(2): 119–132.

Busa, R.A. (2004) Foreword: Perspectives on the Digital Humanities, in Schreibman, S., Siemens, R. and Unsworth, J. (eds), *A companion to digital humanities*. Malden, USA: Blackwell. xvi–xxii.

Byrne, D. (1998) *Complexity theory and the social sciences: An introduction.* London, England: Routledge.

Chaney, D. (1993) *Fictions of collective life: Public drama in late modern culture.* London, England: Routledge.

Crasborn, O. and Hancke, T. (2003) Metadata for sign language corpora. Background document for the ECHO workshop held 8–9 May, Nijmegen, The Netherlands. http://www.let.kun.nl/sign-lang/echo/

Edwards, P., Jackson, S.J., Bowker, G.C. and Knobel, C.P. (2007) Understanding infrastructure: Dynamics, tensions, and design. Report of a workshop on 'History and theory of infrastructure: Lessons for new scientific cyberinfrastructures.' http://www.si.umich.edu/cyber-infrastructure/UnderstandingInfrastructure_FinalReport25jan07.pdf

Foucault, M. (1970) *The order of things: An archaeology of the human sciences*. London, UK: Pantheon.

Fry, J. (2004) The cultural shaping of ICTs within academic fields: Corpus-based linguistics as a case study. *Literature and Linguistic Computing*. 19(3): 303–319.

Fry, J. (2006a) Coordination and control of research practice across scientific fields: Implications for a differentiated e-science, in Hine, C. (ed.) *New infrastructures for knowledge production: Understanding e-science*. London, England: Information Science Publishing.

Fry, J. (2006b) Studying the scholarly web: How disciplinary culture shapes online representations. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*. 10(1), paper 2.

Fuller, S. (2000) *Thomas Kuhn: A philosophical history for our times*. Chicago, USA: Chicago University Press.

Galison, P. (1997) *Image and logic: A material culture of microphysics.* Chicago, US: University of Chicago Press.

Giddens, A. (1993) *New rules of sociological method (second edition).* Stanford, US: Stanford University Press.

Gilbert, N. and Mulkay, M. (1984) *Opening pandora's box: A sociological analysis of scientists' discourse.* Cambridge, England: Cambridge University Press.

Gilman, S.L. (2004) Collaboration, the economy, and the future of the humanities. *Critical Inquiry*. 30: 384–390.

Gumbrecht, H.U. (2004) *Production of presence: What meaning cannot convey*. Stanford, US: Stanford University Press.

Hajič, J. (2004) Linguistics meets exact sciences, in Schreibman, S., Siemens, R. and Unsworth, J. (eds) *A companion to digital humanities*. Malden, USA: Blackwell. 79–87.

Halliday, M.A.K. (1978) *Language as social semiotic: The social interpretation of language and meaning.* London, England: Edward Arnold.

Hine, C. (2000) *Virtual ethnography.* London, England: Sage.

Hine, C. (ed.) (2000) *New infrastructures for knowledge production: Understanding e-science*. London, England: Information Science Publishing.

Jernudd, B.H., and Neustupny, J.N. (1991) Multi-disciplined language planning, edited by Marshall, D.F. (Ed.), in *Language planning: Focusschrift in honor*

12

of Joshua A. Fishman on the occasion of his 65th birthday. Amsterdam, the Netherlands: John Benjamins.
Johnston, T. and Crasborn, O. (2006) The use of ELAN software annotation software in the creation of sign language corpora. Presentatoin at the 2006 E-MELD workshop on digital language documentation held 20–22 June at Michigan State University in East Lansing, Michigan, US.
Latour, B. (2005) *Reassembling the social: An introduction to actor-network theory*. New York, US: Oxford University Press.
Knorr-Cetina, K. (1981) Social and scientific method or what do we make of the disctintion between the natural and social sciences? *Philosophy of the Social Sciences.*
Knorr-Cetina, K. (1999) *Epistemic cultures. How the sciences make knowledge*. Cambridge, US: Harvard University Press.
McEnery, T. (2004) Europe's ignored languages, in Sampson, G. and McCarthy, D. (eds) *Corpus linguistics: Readings in a widening discipline.* London, England: Continuum.
Mol, A. (2002) *The body multiple: Ontology in medical practice*. Durham, US: Duke University Press.
Morford, J.P and Macfarlane, J. (2003) Frequency characteristics of American Sign Language. *Sign language Studies*. 3(2): 213–225.
Ribes, D., Baker, K.S., Millerand, F. and Bowker, G.C. (2005) Comparative interoperability project: Configurations of community, technology, organization. Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries. Denver, US.
Schneider, S. and Foot, K. (2002) Online structure for political action: Exploring presidential web sites from the 2000 American election. *Javnost (The Public)*. 9(2): 43–60.
Searle, J.R. (1969) *Speech acts: An essay in the philosophy of language.* Cambridge, England: Cambridge University Press.
Spence, J. and Holland, P. (eds) (1991) *Family snaps: The meanings of domestic photography*. London, England: Virago.
Schreibman, S., Siemens, R. and Unsworth, J. (2004) The digital humanities and humanities computing: An introduction, in Schreibman, S., Siemens, R. and Unsworth, J. (eds) *A companion to digital humanities*. Malden, USA: Blackwell. xxiii–xxvii.
Stubbs, M. (2001) Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*. 22(2): 149–172.
Talstra, E. (1980) Exegesis and the computer science: Questions for the text and questions for the computer. *Bibliotheca Orientalis.* 37: 3–4.
Thoutenhoofd, E.D. (1998) Method in a photographic enquiry of being Deaf, in *Sociological Research Online*. 3(2)
Unsworth, J. (2002) What is Humanities computing, and what it not? in Braungart, G., Eibl, K. and Jannidis, F., (eds) *Jahrbuch für Computerphilologie 4*. Paderborn, Germany: Mentis.
http://www.socresonline.org.uk/socresonline/3/2/2.html
van Hout, R. (2006) Projectdetails Corpus Nederlandse Gebarentaal (CNGT), Investeringen NWO-middelgroot.
www.nwo.nl/projecten.nsf/pages/2300132576
van Peursen, W. (2002) Morphosyntactic and Syntactic Issues in the Syriac Text of 1 Kings 1, in Cook, J. (ed.), Bible and Computer: The Stellenbosch AIBI-6

Conference. Proceedings of the Association Internationale Bible et Informatique “From Alpha to Byte”, University of Stellenbosch, 17–21 July 2000. Leiden, The Netherlands: Brill. 99–112.

Vann, K. and Bowker, G.C. (2006) Interest in Production: On the configuration of technology-bearing labours for epistemic-IT. In Hine, C. (ed.) *New Infrastructures for Knowledge Production: Understanding E-Science*. London, England: Idea Group.

Whitley, R. (1984) *The intellectual and social organization of the sciences*. Oxford, UK: Clarendon Press.

Widdowson, H.G. (2000) On the limitations of linguistics applied. *Applied Linguistics*. 21(1): 3–25.

Wouters, P., Vann, K., Shcarnhorst, A., Ratto, M., Hellsten, I., Fry, J. and Beaulieu, A. (2007) Messy shapes of knowledge: STS explores informatization, new media and academic work, Hackett, E., Amsterdamska, M., Lynch, M. and Wajcman, J. (eds) *New handbook of science, technology and society*. Cambridge, US: MIT.

Zuiderent-Jerak, T. (2002) Blurring the center: On the politics of ethnography. *Scandinavian Journal of Information Systems,* 14(2): 59–78.

14