Corpus linguistics and language documentation: challenges for collaboration

Christopher Cox

University of Alberta

Abstract

Recent literature in corpus linguistics (e.g., McEnery & Ostler 2000) and language documentation (e.g., Johnson 2004) suggests both disciplines may share natural points of interaction, having in common an interest in the construction and use of permanent collections of diverse linguistic data. Although considerable benefit might be anticipated from close collaboration between these two areas, divergences in their respective purposes, practices, and products may render such an interaction more difficult to foster than might initially be expected.

This paper considers points of commonality and difference between corpus linguistics and language documentation in four specific areas of practice, drawing upon examples from ongoing corpus construction and language documentation efforts centered on Mennonite Plautdietsch in Canada. Given the results of this comparison, this study proposes viewing corpora as descriptive applications of language documentation, to be built directly upon the permanent documentary record. By founding corpora upon documentary materials, such an approach opens language documentation more readily to the analytical and methodological contributions of corpus linguistics, while providing a solid empirical basis for future corpus construction.

1. Introduction

In a recent commentary on responses within linguistics to issues of language endangerment, Himmelmann (2008) remarks on the concentration of one such initiative, language documentation, upon the construction of reusable, permanent collections of primary linguistic data. Paralleling other definitional statements in the current literature on documentary linguistics (e.g., Himmelmann 1998, 2006, Woodbury 2003, Austin 2010), Himmelmann observes that

> [l]anguage documentations [...] focus on observable linguistic behavior and knowledge. The goal is a lasting, multifunctional record of the linguistic practices attested at a given time in a given speech community and the knowledge speakers have about these practices. (Himmelmann 2008: 346)

Programmatic definitions such as these aimed at the goals and practices of language documentation bear notable similarity to sentiments often expressed in corpus linguistics, where a similar emphasis is placed upon multifunctional records of attested linguistic acts (cf. Leech 1992: 112, Stubbs 2001: 221). Commenting on the benefits of reusable corpus-linguistic materials in the context of language endangerment, McEnery & Ostler (2000) note that

> [...] in collecting corpus data for endangered and minority languages, researchers are not only paving the way towards better descriptions of those languages, they may also be allowing work to begin on [a] range of language technology applications [...]. Corpora are multifunctional resources, and their potential for reusability is high. Consequently, although the trend towards minority and endangered languages to be studied largely in the context of ethnolinguistics is understandable, this work, as it becomes corpus based, also opens up the prospect of reuse and retargeting of corpus data gathered by such work to language-processing applications. (McEnery & Ostler 2000: 417)

A closer review of the recent literature in both disciplines suggests such similarities between present-day corpus linguistics and language documentation to be more than skin deep. In addition to this shared interest in producing and analyzing reusable collections of linguistic data, studies in both disciplines have (i) laid claim to a common intellectual heritage, (ii) acknowledged language endangerment as a significant motivation for certain aspects of their development, and (iii) often involved critical reflection upon the empirical foundations of linguistic research. Each of these commonalities is discussed in greater detail below.

1.1 Reusable collections of linguistic data

As the introductory quotation by Himmelmann suggests, the theoretization of practices surrounding the production of permanent, multifunctional stores of primary linguistic data has represented a central issue for language documentation. This concern reflects in part a broader effort to ensure that documentary records are suitable for use in as wide a range of linguistic and socio-cultural applications as possible. Corpus linguistics intersects with language documentation in this regard inasmuch as it deals with the construction and analysis of consistent, reusable collections of linguistic data. At their core, both disciplines are rooted in a common concern for what in corpus linguistics has occasionally been termed "first-order data" (Stubbs 2001: 66), or, in language documentation, "primary data" (Himmelmann 2006: 7).

1.2 Common intellectual heritage

This concentration upon primary data compilation and analysis in both disciplines has sometimes been cited as reflecting a common intellectual heritage, as well. In particular, studies in both corpus linguistics and language documentation have laid claim to the anthropological linguistic tradition established by Franz Boas and his students. McEnery & Ostler (2000: 404) identify the “recording [of] dying languages for posterity in the form of paper-based corpora” as a component of Boasian fieldwork, establishing a clear connection between the construction of linguistic corpora and Boasian text collection. Likewise, although they are careful to distinguish between the characteristics of current language documentation and other preceding traditions, both Woodbury (2003: 35) and Himmelmann (2006: 14) similarly acknowledge the formative influence of “[l]inguistic and anthropological fieldwork in the Boasian tradition” (Himmelmann 2006: 16) in the development of documentary linguistics, particularly in its emphasis upon the recording of texts.

1.3 Language endangerment

As in the Boasian Americanist linguistic tradition, issues of language endangerment have also contributed substantially to the development of language documentation (Himmelmann 2008, Austin 2010: 14). While less central to its historical development, contemporary corpus linguistics has similarly viewed corpus development as a possible response to language endangerment. McEnery & Ostler (2000: 414) present a particularly clear statement to this effect, perceiving a critical role for the development of spoken corpora in the study and support of endangered languages, while underscoring the importance of the data-gathering required for such corpora as a component useful to both linguistic analyses and language revitalization efforts.

1.4 Empirical foundations of linguistic research

This consideration of potential applications of reusable language resources has also brought attention in both disciplines to the empirical underpinnings of linguistic research. Practitioners in both areas have repeatedly advocated replicable studies which draw upon such resources, whether in the form of corpora from the corpus-linguistic tradition, or the collections produced in language documentation.[1] Himmelmann (2006: 1) remarks on bolstering the “empirical foundations” of linguistics through greater reliance upon the primary data available in documentary collections, thus supporting greater accountability in published analyses through replication and falsification (2006: 6). These comments are essentially mirrored in Leech (1992: 112), who similarly proposes falsification as a desirable trait of linguistic models based upon open corpora.

1.5 Mutual interests of corpus linguistics and language documentation

These commonalities are not limited to a parallel history or shared positions on what constitutes desirable research practices in linguistics, but also encompass a mutual interest on the part of each discipline in the practices of the other. For corpus linguistics, the results of documentary linguistic research offer strikingly diverse, well-catalogued samples of natural language, materials which might serve as the basis for the construction of linguistic corpora. These same language materials often represent comparatively understudied languages whose integration into digital corpora and tools for corpus-based analysis presents a long-standing challenge for corpus linguistics (cf. McEnery & Ostler 2000). In both respects, the prospects for corpus-linguistic involvement in the development of documentary linguistic materials appears promising, presenting an opportunity to extend the scope of corpus-linguistic methods, tools, and standards to include languages whose typological features may not have been addressed in corpus construction to date.

Similarly, language documentation might seek to benefit from closer engagement with developments in corpus and computational linguistics. From the perspective of documentary linguistics, corpus linguistics might be seen as potentially offering another methodological perspective upon the documentary record, as well as another set of tools for this record's analysis. Several decades of research in corpus linguistics on collocation and corpus-based lexicography, for instance, present no shortage of elaborated methods which might be harnessed in the production of documentation-based dictionaries, one of the more common descriptive applications of language documentation (Teubert 2001, Heid 2008). Corpus-linguistic procedures may present new methods of opening the documentary record to analysis, and thus encourage further reuse of the documentary record in a wider range of descriptive applications.

More generally, corpus-linguistic methods present another means of rendering the documentary record accessible to *both* academic and non-academic communities. Providing even a simple, concordance-based search facility for a documentary collection – a search interface comparable to that of an internet search engine, and thus perhaps less daunting to less technically-inclined users – may be of use to many individuals interested in exploring the contents of a documentary collection. This kind of interface clearly does not prevent other methods of access, or obviate the need for discussion among those parties engaged in documentation to determine the appropriateness of such a view of the documentary materials involved. Nevertheless, corpus search interfaces offer another way of sharing documentary collections with a general audience, an issue which remains a standing challenge for documentary linguistics (cf. Nathan 2006).

1.6 Convergence and divergence

A gradual convergence of both disciplines in matters of corpus construction thus presents an opportunity to encourage greater interdisciplinary collaboration from which both disciplines may benefit. Yet, such an interaction may not be as simple to foster at it would seem at first glance. The anticipated strengths of such a collaboration lie in the distinctive contributions of each discipline – in the degree to which the contributions of the one might both meet the needs of and present instructive challenges to the other. Such distinctiveness might also be expected to extend to the respective practices and purposes of each discipline, however, and thus not guarantee immediate compatibility of interests in either respect.

The following sections examine such commonalities and differences between corpus linguistics and language documentation through the lens of ongoing corpus-based documentation of Canadian Mennonite Plautdietsch. Mennonite Plautdietsch (ISO 639-3: pdt) is the traditional, West Germanic language of the Dutch-Russian Mennonites, an Anabaptist Christian denomination. The repeated persecution, emigration, and exile of Mennonite Plautdietsch speakers over the course of several centuries has contributed to the exceptional character of the language, having been separated from the larger continental West Germanic dialect continuum during this time, as well as to the unusually broad distribution of present Mennonite Plautdietsch speech communities over four continents and more than a dozen countries (Epp 1993). Although estimates of the total population of Mennonite Plautdietsch speakers are difficult due to this geographical dispersion, Epp (1993) and Gordon (2005) place the worldwide speaker population at approximately 300,000, with the Canadian population contributing substantially to that number. The endangerment of most Canadian varieties of Mennonite Plautdietsch and their unique status as 'parent' dialects to several currently thriving speech communities in Latin America partly motivates efforts to develop permanent linguistic records useful to Plautdietsch speech communities and to researchers. An element of this ongoing work involves not only the preservation of existing language records and the creation of new language resources, but also corpus construction as a parallel component of documentation.

The aim of this paper is thus twofold, seeking to offer a general comparison of current practices in language documentation and corpus construction and their apparent compatibility in several areas, while grounding its conclusions in the specific experiences of language documentation and corpus construction in the case of Canadian Mennonite Plautdietsch. As such, this study does not claim to present an exhaustive or authoritative comparison of both disciplines. Differing perspectives may well exist on the relative compatibility of both areas, and other recommendations might justifiably be made on the basis of differing experiences in resolving potential points of difference between the practices of the two fields. The conclusions of this study are based upon consideration of issues which appear relevant with Mennonite Plautdietsch, and invite further consideration in the light of other corpus construction and language documentation efforts.

2. Corpus linguistics and language documentation

The concentration upon corpus linguistics and language documentation in the following sections calls for more thorough consideration of what is understood by both terms. Following Himmelmann (1998, 2006, 2008) and Woodbury (2003), this study takes language documentation to be an independent discipline within linguistics, concentrating upon the theoretization and practice of developing and analyzing permanent, multipurpose collections of linguistic data. As such, it is distinct from, though not entirely independent of, linguistic description in the form of lexicography, grammatography, and other aspects of linguistic inquiry which make use of the contents of the permanent documentary record. By comparison, the focus of most of the discussion which takes place here under the label of 'corpus linguistics' is upon corpus building and the development of facilities for the analysis of corpus resources – that is, upon the process and products of corpus construction and the accompanying technical infrastructure, rather than upon any particular methodological standpoint on the 'proper' interpretation of the resulting corpora. This paper thus attempts to remain essentially neutral in the debate over the status of corpus linguistics as a distinct area of research within linguistics, or as a methodology of potential service to other linguistic subdisciplines.

The discussion that follows concentrates upon four specific areas of comparison, focusing on the practices of language documentation and corpus construction as they relate to the relationships between project stakeholders, methods of linguistic sampling, conventional technologies, and the treatment of data and metadata in both disciplines. Each of these areas is considered in turn, beginning in each case with a comparison of perspectives from both disciplines and concluding with comments on how the issues thus raised are reflected in the documentation of Canadian Mennonite Plautdietsch.

2.1 Stakeholder relationships

Common to both corpus construction and language documentation is the involvement of multiple parties in the development process. Even where the teams responsible for development are relatively small, efforts are commonly still made to develop a final resource amenable to the audiences and end-uses envisioned by the developers. This study refers to these audiences, both those directly involved in development and those providing guidance and evaluation to the development process, as *stakeholders* – individuals or groups having an interest in the development or results of development of the language resources under consideration.

Within the context of corpus construction, decisions concerning the overall composition of a corpus and the eventual selection of its contents are often the result of careful corpus planning. This process serves to determine the distribution of corpus documents over demographic, geographical, social, and textual categories of interest, and thus informs the later choice of language materials which might fill these categories (Sinclair 2005). More often than not, these corpus planning decisions are made in accordance with the linguistic interests of the developers and the intended final uses of the corpus, rather than beginning with an assessment of the language materials on hand and proceeding from there to determine categorical boundaries. Where corpus planning suggests that access may be required to materials which are not available to or cannot be easily produced by the corpus development team, terms of use are typically negotiated with the copyright holders of such materials when and where the need arises.[2]

The relationship which exists between corpus developers, copyright holders, and the intended end-users of the corpus is and was typically one of business, being characterized by relatively short-term interactions centered around the formal negotiation of terms of access and redistribution acceptable to each stakeholder. In such circumstances, it is rare that a publisher holding the rights to materials sought after in building a new corpus will express any immediate interest in the interface which will be applied to the final corpus, for instance, or raise concerns over the application of part-of-speech tagging or orthographic normalization to the texts provided, as long as the agreed-upon terms of use, redistribution, and access are strictly maintained. It is rarer still for individual authors represented by that publisher to have much say in the ultimate representation of their works in the corpus. As corpus elements, these authors' texts are treated first and foremost as linguistic data, rather than as works intended for the aesthetic or artistic appreciation of a general audience. In the end, the decisions which define the composition and contents of a corpus are typically made by linguists for linguists, with negotiation of access to the relevant texts being negotiated essentially 'on demand' with the relevant stakeholders. These stakeholders' interest in the resulting compilation of linguistic data only infrequently extends beyond the protection of their asserted intellectual property rights.

This situation stands in contrast to the relationships which commonly exist between stakeholders in language documentation. Here, planning often emerges as a product of active partnership between linguists and community members (where these are not already one and the same), with both groups determining the composition and contents of the collection, either directly or indirectly through their interactions. In such cases, stakeholders representing the language community and the academic community participate in shaping the ultimate form of the documentary record through long-term participation and negotiation (Leonard & Haynes 2010). This interactive process of definition seeks to develop common goals and outcomes which meet the needs of all stakeholders involved, and is thus not entirely under the direction of any linguists involved, as is often the case in corpus construction.

Interactions between stakeholders in language documentation may resemble those found in corpus construction, involving short-term, business-like negotiation of access and distribution rights with the holders of intellectual property rights. In many cases, however, language documentation proceeds on a more familiar and longer-term basis, depending critically upon relationships of trust built between individuals in communities over the course of years and even decades of interaction. Such relationships are often essential in situations where language documentation is perceived to be a politicized activity (cf. Ostler 2009), whether this politicization stems from previous interactions between the stakeholder communities or from attitudes towards the language(s) and/or language materials involved. Interactions between stakeholders under such conditions, particularly in determining appropriate access and reuse conditions, are arguably of a different quality than the comparatively less politicized negotiations between corpus linguists and copyright holders. The relative disinterest on the part of the latter copyright holders in the ultimate representation of the provided language 'data' would be unexpected in the kinds of language documentation situations described above. Where individual language materials bear considerable cultural, historical, or personal significance, their intermediate treatment and final presentation in language documentation may require more extensive collective deliberation between stakeholders.

Language documentation involving Canadian Mennonite Plautdietsch faces many of the issues noted above in both corpus linguistic and language documentation stakeholder relationships. As in corpus linguistics, some existing language materials have been made available for language documentation and corpus construction without an extended period of consultation. Bible translation agencies, first-language literacy initiatives, and radio programs in particular have often been willing to share their language materials with a wider audience, in accordance with their respective mandates. Even in these cases, however, there is often the concomitant expectation that the contributed resources will be used in a respectful manner and with due acknowledgement, which in turn requires a mutual understanding of what proper reuse entails. In one instance, a series of discussions concerning historical Plautdietsch language materials appearing in a prominent Mennonite periodical concluded with the blessing of the editor for the documentation project to "do whatever you like" with the sources – as long as everyone was clear on what was meant by 'whatever you like'!

More often, language documentation has followed from long-term relationships with community members and institutions who have an expressed interest in Mennonite Plautdietsch, and whose contributions and guidance codetermine the final products of documentation. Several prominent Mennonite Plautdietsch authors, including the lexicographer Herman Rempel and scholar Reuben Epp, have made portions of their respective Mennonite Plautdietsch literary estates generally available under Creative Commons agreements, having been encouraged to consider doing so in consultation with close colleagues and friends. In contrast to many other minority language communities in Canada, where language documentation and description efforts have typically involved the presence of interested 'outsiders' to the speech community, the vast majority of academic linguistic description and language documentation concerned with Canadian Mennonite Plautdietsch has been undertaken by members of Mennonite Plautdietsch-speaking communities themselves. All dictionaries of Mennonite Plautdietsch published to date have been compiled by native speakers of the language, for instance. This situation has perhaps contributed to a less politicized view of language documentation within Canadian Mennonite communities, where it has generally been seen as bringing welcome, if somewhat unexpected attention to the language. As in the case of the periodical discussions mentioned above, however, the familiar, community-internal nature of most linguistic scholarship serves to underscore the expectation of a thorough understanding of what constitutes ‘appropriate’ materials and practices for language documentation.

2.2 Methods of sampling

A concern shared by both corpus linguistics and language documentation centers on determining adequate coverage of linguistic contexts in the collection under development. This includes deciding what should be part of the final collection and in what proportion; when collection activity itself should cease for the sake of processing, release, and archival; and the degree of first-hand involvement of the collators of linguistic materials in the linguistic acts being recorded. Each of these aspects is considered below before returning to review the situation with Canadian Mennonite Plautdietsch.

2.2.1 Representativeness and balance

As noted in the introduction, corpus linguistics has long discussed what materials should comprise a corpus and in what proportion, such that the resulting resource might be taken to be a reasonable sample of the group of speakers and speech practices under consideration. Typically, corpus construction seeks to achieve *balance* (representation of variation across the types of corpus sources sampled that is proportional in some way to their occurrence in the larger, unsampled population) and *representativeness* (coverage of all those types of corpus sources deemed relevant to the reproduction of the linguistic practices or knowledge to be modeled by the corpus). Both balance and representativeness are often taken in corpus development as “target notions”, in the terms of Sinclair (2005: 9), who acknowledges that “(w)hile these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.”

The lack of representativeness or balance may limit the research questions which may be asked of a corpus, as well as the degree to which generalizations derived from such a corpus can be applied to other situations (cf. Biber, Conrad, & Reppen 1998: 246). A study of the language used in English prose fiction, for example, would find it challenging to argue convincingly that its conclusions about that genre are in fact of broader scope if it used a corpus of science fiction novels alone. Likewise, psycholinguistic research drawing upon corpus samples for general frequency measures of words as possible response predictors might justifiably express concern if the contents of a given corpus are skewed in the direction of a particular demographic, geographical, social, or textual category. These ‘distortions’ might further exacerbate problems caused by sample non-randomness when deriving reliable frequency estimates from corpora (cf. Evert 2006). Such concerns commonly manifest themselves in corpus linguistics through an emphasis upon careful design, planning, and selection in corpus construction to ensure that balance and representativeness are maintained throughout.

Discussions of documentary linguistic practice are generally sympathetic to issues of balance and representativeness in sampling, given an overarching concern for broad and balanced coverage of linguistic practices and metalinguistic knowledge in permanent documentary collections. Only rarely, however, are practitioners of language documentation able to exercise control over the planned coverage of sampled contexts to the extent commonly expected by corpus linguists to ensure representativeness and balance. There are several reasons why this may be the case:

Naturalness and opportunistic recording. When language documentation aims to represent naturally-occurring (i.e., unstaged, non-elicited) language events within a speech community, the recording of a wide range of communicative events in natural contexts is, by necessity, partially opportunistic (cf. Woodbury 2003: 39). Woodbury (2003: 47) identifies the opportunistic nature of this aspect of documentary linguistics as a hallmark of "good corpus production." Individuals involved in language documentation can decide what to record and what to leave unrecorded, or aim to be present at particular events or in particular situations in order to shape the balance of the documentary collection. Even then, however, it is still largely impossible for these individuals to know in advance what precisely will come out of different, unscripted contexts, let alone attempt to control *a priori* for coverage of discourse categories to the same extent as in corpus linguistics.

Issues of subsampling. Another way of achieving corpus-like balance of materials within a documentary collection involves *subsampling*, or selecting a subset of the available records to produce more 'balanced' or 'representative' collection. While potentially effective, such efforts may not be met with equal acceptance by all stakeholders in documentation. Materials may have cultural or personal significance to individual stakeholders which extend beyond these materials' membership to particular corpus categories. It may be difficult to justify leaving out one speaker or author's contributions to documentation while including another's, simply because one particular corpus design category has already been saturated. Likewise, the decision to keep only a snippet of two thousand words from a sacred text may prove highly controversial, even though it follows a corpus design practice established in the influential Brown family of corpora (cf. Francis & Kučera (1964), though note the more recent recommendation of Sinclair (2005: 7) to have corpus documents represent whole sources).

Collaborative determination of documentary coverage. Language documentation is commonly undertaken with a commitment to share and develop documentary records with all stakeholders wherever ethically possible within the framework of a given project. Broad and balanced coverage may be among the goals of participating stakeholders, but are almost certainly not the only guiding principles. No less determinant are the goals of language community members themselves, who may define appropriate balance of coverage differently from possible stakeholders in the academic linguistic community. As Himmelmann (2006: 4) observes, the degree to which broad documentary coverage might be achieved is, to a significant extent, determined by relationships with and within the language community, as well as by the larger social setting in which documentation takes place. What comes to be part of the documentary record is, in many cases, not only the result of *a priori* planning to address concerns over balance and representation, but also of ongoing discussion of the priorities of all parties – each of whom may favor increased representation of certain contexts or genres deemed to be of greater interest or intrinsic importance.

Language endangerment. Where documentation takes place within the context of language endangerment, in which a language of wider communication or higher overt prestige has begun to occupy domains of traditional language use, it may be difficult to find instances of all genres of interest still being practiced in their original settings, if that is goal of balanced sampling. Section 2.2.3 revisits this point in greater detail.

2.2.2 Project completion

As noted above, it is not uncommon in corpus linguistics to set explicit goals for corpus size and composition as part of an initial corpus design, and, thus, to have some notion of when the corpus is ready for release. In his recommendations concerning corpus development, redistribution, and preservation, Wynne (2005b: 72) explicitly reminds corpus builders to know when to stop developing their corpora, both to avoid the "danger of excessive perfectionism" and to permit corpus use and reuse by a wider audience. Wynne is careful to make an exception for monitor corpora, where changes over time enter into the corpus in a principled way, although still stressing the importance of predictable and well-documented management of such a resource for it to be of practical use.

By contrast, language documentation may not be 'predictable' in the same sense, with predetermined goals for corpus size and coverage being established before any public release is possible. Although clearly dependent upon the details of individual documentation projects, in cases where data gathering is in part opportunistic, development often follows a more gradual and cumulative process. Documentary materials often become available to more project stakeholders as their annotation progresses. Major language documentation projects funded by multi-year grants may have end-dates and definite goals for the production of particular language materials associated with them, but this certainly does not imply that the development of the documentary record is completed when the grant funding comes to an end, or that additional materials or annotations might not be added in the future. The ongoing nature of language documentation can make a single release event difficult to accommodate, although milestones may certainly be set in the data-gathering or annotation processes (cf. Woodbury 2003: 47).

2.2.3 Involvement in language production

Linguistic sampling in both corpus linguistics and language documentation may also differ in the degree to which individuals are presumed to be involved in both the planning of the collection and in the recorded linguistic events themselves. It has been common practice in much of corpus linguistics to rely upon existing, naturally-occurring examples of language as the basis for corpus construction. This practice of adopting existing materials for secondary uses in corpora is reflected in Stubbs's (2001: 221) insistence that "corpus data [be] part of natural language use and not produced for the purposes of linguistic analysis." Even if sampling procedures may be influenced by the interests of the researchers involved in corpus compilation, the corpus compilers are themselves not represented in the corpus *per se*, thus effecting a division between observation and analysis.

This ideal of a clean separation between data gathering and interpretation is not always met in corpora containing spoken language intended to be broadly representative of a given speech community. Certain important genres (e.g., private conversations) are often poorly represented in generally-available sources of language data, and therefore require more active data-gathering paradigms to represent adequately in corpora. The influence of the researchers involved may be mitigated somewhat by providing recording equipment to friends-of-friends or to hired 'recruits,' who then perform the required recording. A similar process is described as part of the construction of the spoken component of the British National Corpus, as outlined in Crowdy (1993) and Aston & Burnard (1998: 32-33). Even still, one might recognize potential issues posed by the direct involvement of such delegates in producing language intended for inclusion in a corpus. This solution does not wholly address the possibility of recordings being influenced by speakers' knowledge of the recordings' intended purposes (if this information is communicated in advance as part of informed consent), or by speakers' awareness of the presence of recording equipment. Although some corpus construction projects have attempted to circumvent these issues by initially concealing recording equipment and later seeking informed consent (e.g., Torreira & Ernestus 2010), in many cases, spoken corpus construction must still contend with manifestations of the Observer's Paradox.

Issues such as these are of no lesser relevance in language documentation, where observation and recording may inadvertently introduce bias into the collection. Most definitions of language documentation are careful to cast a wide net, making room for existing materials not produced explicitly for documentation to serve as welcome parts of the documentary record. This may potentially lower the risk of pervasive bias going undetected. Nevertheless, language documentation efforts appear on the whole hesitant to rule out more active investigations of language, which may include documenting interactions involving members of the documentation team. This may be partially for reasons of practicality acknowledged in the corpus-linguistic literature: since language materials are often not available in such abundance that corpus construction can proceed through secondary sampling alone, additional recording are sometimes required (cf. McEnery & Ostler 2000: 414).

Such procedures may well run afoul of Stubbs' (2001: 221) stated concentration of corpus linguistics upon "attested texts – real acts of communication used in a discourse community." However, focusing on only those communicative acts taking place currently in the language of interest may inadvertently underestimate the degree of linguistic knowledge still present but rarely enacted within the speech community, particularly in cases of recent language shift. Moreover, documentary linguists have repeatedly insisted that elicitation not be ruled out as a component of language documentation alongside naturally-occurring language (cf. Woodbury 2003: 42). Rather, elicitation presents another tool for investigating aspects of linguistic and metalinguistic knowledge not easily approached by observing naturally-occurring language alone – albeit with the understanding that "the whole elicitation procedure is to be considered a special, somewhat artificial communicative event and is to be documented as such" (Himmelmann 1998: 13).

2.2.4 Mennonite Plautdietsch

Each of the aforementioned issues bears upon documentation and corpus construction in the case of Canadian Mennonite Plautdietsch, as well. General concerns over balance and representativeness in the resources under development must be weighed in the context of both specific Mennonite community interests in language documentation, and in the light of an ongoing shift to English in most Canadian Mennonite communities. Collaborative documentation of a broad range of linguistic practices is certainly of general interest, perhaps particularly given a growing awareness within Canadian Mennonite communities of the gradual decline of Mennonite Plautdietsch from domains of language usage outside of the home (cf. Reimer, Reimer & Thiessen 1983: 1).

In the Saskatchewan Mennonite communities with which the present documentation has been most closely involved, recent public events centered around Mennonite Plautdietsch have drawn audiences of hundreds of speakers, who often offer comments on the importance of the language as a vital element of Russian Mennonite history and culture. Interest in language documentation in these communities has commonly reflected this pervasive association between Mennonite Plautdietsch language use and local historical-cultural knowledge, as well as the perceived importance of the preservation of such knowledge. As might be anticipated from this framing of the consequences of language loss, community recommendations for areas of documentation have often prominently favored Mennonite oral history, to be conducted among the eldest generation of speakers. This interest is certainly not an impediment to documentation, and presents in itself many chances for a richer documentary collection to be developed. It also presents an opportunity, however, to discuss the priorities of documentation more generally and to balance historically-oriented research with other kinds of recordings (e.g., involving the younger generations of speakers in non-interview contexts) to produce a collection which is both diverse and of cultural interest.

Given the progress of language shift in many Mennonite communities, it is not immediately clear what 'representativeness' in the corpus-linguistic sense should entail for practical purposes. Devoting primary attention to Mennonite Plautdietsch among those languages maintained within traditionally multilingual Russian Mennonite communities in the first place implies a demographic skew towards the older cohort of speakers who have maintained the greatest degree of Plautdietsch fluency, and the concomitant underrepresentation of child-directed speech, the language of younger individuals who have maintained only a passive knowledge of the language, and of semi-speakers more generally. Strictly speaking, achieving demographic 'balance' between age-based cohorts in language materials collected may not be impossible, but is almost certainly not realistic in reflecting the current state of language vitality in the community – and, thus, not representative of Plautdietsch language use in the contexts where it remains practiced in these Canadian Mennonite communities.

In a similar way, language shift results in a higher representation within documentation of those domains where Plautdietsch language use remains comparatively robust (namely in private conversations between family members and friends), and lesser representation of other, traditional domains still familiar to those generations of speakers who have experienced this language shift first-hand. Documentary representation of recent Plautdietsch-language church services, for instance, a domain in which a stage-wise transition from Mennonite Standard German to Mennonite Plautdietsch to English had already run its course in many Canadian Mennonite communities several decades ago (cf. Doell 1987: 60, Regehr 1996: 312-315), is necessarily sparse as a consequence of this shift, rather than of documentary inattention. Plautdietsch-language church services may thus be conspicuously absent from the present-day record – but it is also not realistic to leave unrepresented individuals' knowledge of language use in these domains, which are still well within the abilities of many speakers. It would arguably be remiss to neglect these culturally important contexts of language use for reasons of corpus balance alone. While historical recordings of language use in these and other contexts might assist in providing diachronic examples of Plautdietsch in domains which have since shifted to English, setting 'balance' and 'representativeness' as general guidelines does not provide a simple solution to the pervasive problem of wishing to record now-uncommon linguistic practices. Likewise, the separation between involvement in corpus construction and documentary recording advocated by Stubbs (2001) does not afford the opportunity to pursue investigation of these domains through more active means of inquiry. In practice, documentation has involved 'corpus planners' participating in language recording events, with this involvement forming part of the metadata later associated with such sessions.

2.3 Conventional technologies

The range of technologies conventionally employed in corpus linguistics and language documentation represents the area of greatest similarity between these two disciplines, sharing commonalities in name, at least, if not always in practice. Both disciplines make heavy use of digital technologies, although both still commonly deal with 'legacy' materials (and often in quite different fashions). Table 1 summarizes these comparisons, which are taken up individually in more detail in the sections that follow.

Table 1: Technologies in corpus linguistics and language documentation

	corpus linguistics	*language documentation*
Representation of data and metadata	XML-based data and metadata (e.g., XCES, TEI) typical, although SGML and plain-text still common	XML-based data and metadata (e.g., OLAC, IMDI) typical, although plain-text records still common (e.g., Toolbox databases)
Treatment of audiovisual materials	Time-aligned transcription of audiovisual materials (e.g., with CLAN, Praat)	Time-aligned transcription of audiovisual materials (e.g., with Transcriber, ELAN, EXMaRALDA)
Distribution of language materials	Distribution via the Linguistic Data Consortium, TalkBank, individual websites	Distribution via language archives (e.g., DoBeS, ELAR, AILLA), individual websites, although paper copies still common
Annotation tools	Semi-automated annotation tools for tagging, lemmatization, etc.	Annotation rarely automated; tools seldom amenable to changing, provisional analyses

2.3.1 Representations of data and metadata

In both disciplines, XML-based standards for representing textual data and accompanying metadata are essentially ubiquitous. In the case of corpus linguistics, it is common to deal with SGML and plain-text sources for earlier corpora (e.g., the British National Corpus). Similarly, plain-text, semi-structured records are still exceedingly common in language documentation, particularly in the form of Shoebox or Toolbox lexical databases. Somewhat less commonality is noted in the choice of particular XML-based standards by each discipline, however. Whereas corpus linguistics has concentrated upon XCES (Ide, Bonhomme & Romary 2000), TEI (TEI Consortium 2010), LAF/GRaF (Ide, Romary, & de la Clergerie 2004, Ide & Suderman 2007), and related formats for data and metadata (where not inventing custom schemas for individual projects), documentary linguistics has largely focused on other families of standards, such as OLAC (Bird & Simons 2001) and IMDI (Broeder et al. 2001) (where any formal standard is in use, rather than free-form prose text).

2.3.2 Time-aligned transcription of audiovisual materials

Time-aligned transcription has become common practice in both disciplines, not just in areas concerned with multimodality in corpus and computational linguistics, or gesture and sign language in language documentation. As with standards for representing data and metadata, similar general practices might be identified, but less agreement on transcription tools or associated standards for storing the resulting data. Corpora with significant audiovisual components, such as the Santa Barbara Corpus of Spoken American English (Chafe, Du Bois, & Thompson 1991) or SCOTS (Douglas 2003), have commonly relied upon tools such as CLAN or Praat for producing time-aligned annotations, while language documentation projects have seen a greater adoption of Transcriber, ELAN, and EXMaRALDA for similar purposes (e.g., Johnston 2010).

2.3.3 Distribution of language materials

Both corpus linguistics and language documentation have made substantial use in recent years of computational infrastructures developed to support the preservation and distribution of digital language materials. In this respect, the lines between the two disciplines are blurred: language documentation materials are sometimes hosted with distribution services more typically used for corpora (e.g., the Linguistic Data Consortium (www.ldc.upenn.edu), TalkBank (talkbank.org)), and corpora are being deposited in archives largely centered around language documentation (e.g., DoBeS (www.mpi.nl/DOBES), AILLA (www.ailla.utexas.org), ELAR (elar.soas.ac.uk)). While rarer in corpus linguistics, it is not unusual in language documentation for materials to be made available in a variety of ‘popular’ presentation formats to interested parties, be they consumer video DVDs or audio CDs, or even on paper.

2.4 Annotation tools

Perhaps the single largest gap between the current technical practices of corpus linguistics and language documentation lies in their approaches to annotation. Corpus linguistics is often able to avail itself of computational tools for automated or semi-automated annotation tasks, such as the assignment of part-of-speech tags, lemmata, or syntactic parses. This is seldom the case for language documentation, where even the most basic annotation tasks are rarely automated to a comparable extent. This presents a significant challenge not only for language documentation efforts, where annotation levels comparable to those found in existing corpora may involve considerably more time and expense to attain, but also for corpus and computational linguistics, which face non-trivial problems in adapting existing tools and techniques to the analysis of lesser-studied languages and to changing, provisional analyses (cf. Bird 2009).

2.4.1 Mennonite Plautdietsch

Mennonite Plautdietsch documentation and corpus construction finds itself between two worlds with regards to conventional technology. Its commitment to the long-term preservation of and access to collected language materials presents motivation to participate in digital archiving and the creation of multiple interfaces to the documentary record. These 'interfaces' also assist in producing printed versions of these materials for situations where digital access does not meet the needs of stakeholders. Considering the emphasis placed by community stakeholders upon oral history with older speakers in documentation (cf. Section 2.2.4), and the general importance of documentary materials being reviewed by and returned permanently to their contributors, it would not be reasonable to expect that all Mennonite elders would feel comfortable reviewing their contributions in digital form alone, even if help were offered in doing so. As well, some contributors to documentation may have decided not to make use of digital technologies in their daily lives, either as a matter of personal preference or religious conviction. Although this issue has not arisen so far in documentation, it is certainly possible within the Mennonite communities involved. Being able to provide printed copies of digital materials to such contributors (and, where appropriate, to locally-accessible, community-run archives and libraries) may help address this possible problem. Documentary linguistic presentation tools such as CuPED (Cox & Berez 2009) provide assistance in this regard.

Although favoring documentary linguistic annotation tools and standards generally, Mennonite Plautdietsch documentation has still sought to make use of some semi-automatic annotation techniques from corpus and computational linguistics. Cox (2010) details how part-of-speech tags were assigned to materials from the documentary collection, adapting a tag set proposed for a related language and an existing probabilistic tagger for use with Mennonite Plautdietsch. While ultimately successful, this process required several custom scripts to be written to convert between documentary and corpus linguistic data formats, a need which no 'off-the-shelf' software tool in either area was able to meet.

2.5 Treatment of data and metadata

Corpus linguistics and language documentation might be contrasted not only in the technical standards and tools prevalent in each discipline, but also in the amount and treatment of data commonly available to them. In many corpus construction projects, data are relatively plentiful, presenting ample sources out of which multi-million word corpora might be produced. Given the volume of data available, corpus construction typically achieves much of its tractability through normalization, distilling the complexity and heterogeneity of the available sources to more easily compared corpus samples. This may involve 'correcting' or standardizing spelling and punctuation, removing visual formatting from written sources, applying consistent tokenization schemes, and so on. Such processes seek to render texts more consistent, and thus more readily searched and automatically annotated by tools which make use of this uniformity.

By comparison, data resulting from language documentation are often relatively scarce. It is rare for individual documentation projects to have access to the millions of words of running text that are often available in mainstream corpus construction. More commonly, documentary records must be treated as being essentially irreplaceable, whether due to language endangerment or the degree of effort and expense which entered into their production. Unlike corpus sources, documentary materials are commonly preserved in their full diversity, producing collections of featurally-rich primary data which have not been subject to an initial stage of permanent normalization. While certainly feasible, this solution nevertheless incurs a potential loss in overall searchability for the entire documentary collection, given the diversity of its component records.

Normalization is problematic for Mennonite Plautdietsch materials, especially where orthographic conventions are concerned. Multiple distinct orthographic systems exist for representing written Mennonite Plautdietsch (cf. Nieuweboer 1998). Although several proposals have gained support across Mennonite communities, no single standard is prevalent, and the choice of orthographies remains controversial. In light of the potential benefits of consistent orthographic representation for effective search and retrieval across language records, some form of normalization to a single spelling system would seem desirable for Mennonite Plautdietsch materials. Yet, as noted above, this is not merely a technical issue. While multiple orthographies may be problematic for processing materials, it is not possible to replace one system with another without the risk of alienating the original contributors. One might instead choose to lemmatize orthographically-diverse sources, but this merely displaces the problem of orthographic choice to the representation of lemmata, and still prevents reliable non-lemmatized searches (cf. Cox 2010). No less problematic is the potential loss of association between those original sources contributed by writers and publishers to language documentation efforts and the simplified forms of these materials that would typically enter into corpus construction. Converting a series of page scans into plain text, associated with one another only in name, might risk the offence of authors who wish to see their works preserved in their full original presentation.

Social and technical issues such as these are pervasive in the present documentation, and require attention if the materials involved are to be both computationally tractable and culturally acceptable. Several methods of addressing these challenges are therefore discussed in the following section.

3. Bridging the gap

The preceding sections have attempted to offer a comparison of current practices in corpus linguistics and language documentation, concentrating upon issues pertaining to stakeholder relationships, methods of sampling, conventional technologies, and the treatment of data in both fields. This comparison finds several reasons to expect compatibilities between these two disciplines. The similar paths followed in their use of conventional technologies is encouraging, for one, and suggests a relatively close alignment of both disciplines in technical practices. Potentially more challenging are those differences between language documentation and corpus linguistics which bear upon issues of corpus planning, and in particular expectations made of corpora with respect to representativeness, balance, and the sampling of language materials. Here, the opportunistic nature of key aspects of language documentation, the specific interests of stakeholder parties in particular aspects of documentation, and the direct involvement of stakeholders in the production of documentary materials need to be taken into consideration when language documentation is revisited from a corpus linguistic perspective.

Both corpus linguistics and language documentation share at their core an interest in the compilation of collections of linguistic information. Yet, as the preceding section has noted, even this fundamental concern itself is not without significant divisions. A tension might be perceived between the reductionist pressures of corpus construction and the preservationist tendencies of language documentation, a difference stemming in part from the relative abundance and ‘replaceableness’ of the linguistic records commonly available to each discipline, as well as from the perceived acceptability of a reduction in these sources’ complexity as a precondition to their inclusion in a collection or corpus. When added to the differences in perspective noted above, this tension may risk producing a gulf between the two disciplines. In cases where the link between corpus documents and documentary sources is not made explicit, the creation of a corpus arguably does little to augment the documentary record; it presents at best a temporary exploitation, rather than a lasting enrichment of those sources from which it draws its data. Moreover, the stages of normalization commonly required of materials for their use in corpus construction render the corpus itself less than ideal as the primary documentary record. The permanent normalization of language data runs afoul of the basic commitment of language documentation to preserve language materials for future reuse – reuse which may not require, or even find acceptable, the initial simplifications undertaken as part of the normal course of corpus development.

Observations such as these motivate an alternative proposal for the general relationship between corpus construction and language documentation. At the centre of this proposal is the notion that corpora might be viewed productively as *applications* of language documentation, built upon permanent records within the documentary collection which are referenced wherever possible throughout the corpus. From the perspective of documentary linguistics, this proposal establishes the corpus as another component of the larger descriptive apparatus, i.e., as one perspective among many others possible upon the contents of the documentary collection (cf. Woodbury 2003: 42, Himmelmann 2006: 11).

In practical terms, this proposal implies that corpus development is to be largely contingent upon the contents and development of the documentary collection to which it makes reference. The corpus thus essentially grows out of permanent documentation, rather than that documentation being made to fit the Procrustean bed of a single, corpus-linguistic representation. Indeed, on this view, it is entirely possible for multiple corpora to be built out of the materials provided by a single documentary collection, with no one of them being necessarily authoritative.

Critically, explicit and detailed citation of documentary materials in their corpus counterparts might be seen as contributing a 'corpus' interpretation to those materials. If such references between documentary sources and corresponding corpus interpretations are carefully constructed, the possibility exists not only to proceed from a corpus text to its documentary source, but also from a given documentary source to its interpretation within the corpus. The bidirectionality of this relationship, permitting back-and-forth navigation between corpus texts and their documentary sources, enriches both the documentation and the corpus. The accuracy and completeness of the corpus texts, which are often themselves produced through automatic or semi-automatic means like optical character recognition, are more easily confirmed against the documentary sources from which they were derived when such navigation is facilitated. Where elements of a particular transcript within a spoken corpus appear suspect, given the predictions of a particular linguistic model or personal intuition, it is possible to check the transcript against the original recording when the spoken corpus documents make reference to the corresponding sections of the original documentary recording. Similarly, if some part of a written text in the corpus appears to be missing (not an altogether uncommon error for optical character recognition procedures), the corpus text can be held up for immediate comparison against the corresponding page image in the documentary collection.

The benefits of such a deeply interwoven dependency between the derived corpus and the permanent documentation upon which it is based extend to practices of normalization, as well. Recall that normalization of spelling is controversial in Mennonite Plautdietsch, given the presence of multiple competing orthographies. Much less controversial is the preservation of materials in their original orthographies, a function well handled by the documentary collection, which is typically able to curate unedited page scans of printed materials or other digital facsimiles. Where corpus texts make persistent reference to these original documentary sources, the opportunity exists for corpus normalization to be non-replacive, not supplanting the existing representations found in the documentary materials, but rather supplementing those sources with a secondary representation in another orthography. In this way, then, there is the potential to respect the orthographic choices reflected in the original documents while nevertheless permitting corpus normalization, the latter processes being rendered less controversial through the persistent availability of unaltered documentary sources.

This proposal may further allow for a more dialectic relationship to develop between corpus development and language documentation. Much as with the development of documentation-based dictionaries and grammars, the construction of a corpus on the basis of a documentary collection presents opportunities to identify potential weaknesses in coverage which might otherwise escape notice, were this corpus-centered perspective not present. Corpus linguistic concerns of balance and representativeness in the documentary materials feeding into the corpus might help inform the priorities of ongoing documentation, presenting some notion of areas in which documentation might particularly benefit from expansion. Likewise, as has already been suggested in preceding sections, the diversity of materials entering the documentary collection might present healthy challenges to corpus construction, in both the variety of sources and the typological features represented in such data.

Viewing corpora as applications of language documentation does not itself present a solution to persistent issues of representativeness and balance. These 'target notions' may remain difficult to achieve using opportunistically gathered documentary materials, although progress here is certainly not impossible. As stakeholders in the development of an important part of the documentary apparatus, corpus linguists working from documentary sources might enter more fully into discussions concerning the development of the documentary collection itself, and continue to advocate for broad and balanced coverage for much the same reasons expounded in the corpus linguistic literature. Furthermore, as in the case of normalization even where only a small portion of a culturally or personally-significant documentary source is included in a corpus for reasons of balance, references from that corpus to the corresponding documentation still permit users to return to the original source in its full and unedited form. These references then provide one possible means of addressing concerns over the editorial treatment of sensitive documentary materials in linguistic corpora (cf. Section 2.2.1). A corpus which draws selectively upon the documentary record does not present the same danger of permanently limiting reuse of the original documentary sources themselves; the whole records remain accessible in the documentary collection, however they are ultimately represented in the corpus.

For such a proposal to succeed technically, corpus construction must be able to make use of documentary sources as the basis of corpora. Documentary sources must therefore be open to reference within corpus documents, and corpus formats and tools must be able to represent and process such references. Encouragingly, as Section 2.3 notes, strong commonalities exist between the technologies currently in use for corpus linguistics and language documentation, placing a definite emphasis on consistent data and metadata, and on the use of open, often XML-based digital standards for greater interoperability. Standards for the representation of data and metadata in both disciplines differ, but might be hoped to be rendered compatible where serving a comparable function. Computational and corpus linguistic software tools may prove more difficult to adapt to typological features not resembling those found in languages for which corpora already exist, although this remains to be seen.

Lastly, as participants in language documentation, it is likely that corpus builders involved in the construction of documentation-based corpora would eventually be required to come to terms with the 'politicization' of their work, and with the requirements of equal stakeholders from outside the academic and publishing communities. This situation is somewhat atypical in corpus construction based upon other publicly-accessible documents (e.g., those harvested from the web), but arguably of potential benefit for corpus compilation in the long term. A closer, more productive relationship with other stakeholders in documentation may help avoid potential missteps, either political or technical, which would otherwise hinder such corpus construction.

4. Conclusion

The commonality observed between corpus linguistics and language documentation runs deep, proceeding from a fundamental shared interest in the construction and reuse of lasting collections of linguistic data, produced to support both empirically-grounded linguistic research and efforts to address issues of language endangerment. Notable similarities exist between both disciplines in their selection of digital technologies to address common challenges faced in developing permanent language resources. Such similarities, along with other shared traits, give reason to believe that a productive interaction may be possible between both disciplines in meeting their respective needs for increased access to documentary linguistic materials on the one hand, and increased corpus linguistic attention to a wider range of languages on the other.

It bears stressing, however, that this is commonality, and not uniformity, either in purpose, process or product; it would seem premature to take both disciplines to be immediately compatible as such. Although technically feasible, it is no trivial undertaking to address the distance which presently exists between corpus linguistic tools and typologically diverse documentary linguistic data, between corpus documents and their corresponding documentary sources, and between corpus normalization techniques and the aim for the permanent preservation of rich documentary materials. Yet, as was noted in the introduction, it is precisely in the challenge of this divergence between disciplines that one might expect to arrive at a more significant reward – that the contributions of each might ultimately be greater than the sum of their parts.

Acknowledgements

This paper has benefited substantially from discussions with John Newman and Sally Rice, as well as from the comments of two anonymous reviewers and participants in the American Association for Corpus Linguistics 2009 conference. This research has been supported financially by the Social Sciences and

Humanities Research Council of Canada (SSHRC), the Killam Trusts, and the Pierre Elliott Trudeau Foundation.

Notes

1 A terminological distinction is made throughout this paper between "corpora," which are taken here to be the products of corpus-linguistic treatments of language data, and "collections," the stores of primary data and associated metadata which arise out of documentary linguistic research. References to 'corpora' and 'collections' are shared across both disciplines, albeit not always with the same connotations. Corpus linguists have frequently distinguished between "collections" and "corpora," the latter demonstrating efforts at a balanced representation of some linguistic phenomenon of interest across a range of contexts, efforts which may or may not have been present in the compilation of a 'raw' collection (cf. Leech 1991: 10, Biber, Conrad & Reppen 1998: 246, Wynne 2005a: vii). A comparable distinction would appear to be less commonly made in documentary linguistics. Johnson (2004: 3), for instance, takes both terms to be essentially synonymous, stating that a "collection, or corpus, is the body of documentary materials created by linguists and native speakers in the course of their research." Woodbury (2003: 42) similarly refers to the "documentary corpus" when discussing primary materials assembled in the course of a documentation project, an entity which Himmelmann (2006: 10) simply terms a "corpus of primary data."

2 Indeed, for desired materials not to be freely available is far from unusual in the history of European corpus linguistics. Before the advent of the world-wide web, it was exceedingly difficult to assemble sufficient amounts of text in the desired proportions for the construction of balanced, multi-million word corpora without relying on materials "almost all in commercial hands," as Leech (1992: 10) observes.

References

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

Austin, P.K. (2010), 'Current issues in language documentation', in: P.K. Austin (ed.) *Language documentation and description. Volume 7.* London: School of Oriental and African Studies. 12-33.

Biber, D., S. Conrad & R. Reppen (1998), *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Bird, S. (2009), 'Natural language processing and linguistic fieldwork', *Computational linguistics,* 35(3): 469-474.

Bird, S. & G. Simons (2001), 'The OLAC metadata set and controlled vocabularies,' in: T. Declerck, S. Krauwer & M. Rosner (eds.) *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*. Toulouse: unpublished.

Broeder, D., F. Offenga, D. Willems, & P. Wittenburg (2001), 'The IMDI metadata set, its tools and accessible linguistic databases', in: S. Bird, P. Buneman & M. Liberman (eds.) *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia: Linguistic Data Consortium. 48-55.

Chafe, W.L., J.W. Du Bois & S.A. Thompson (1991), 'Towards a new corpus of spoken American English', in: K. Aijmer & B. Altenberg (eds.) *English corpus linguistics: studies in honour of Jan Svartvik.* London / New York: Longman. 64-82.

Crowdy, S. (1993), 'Spoken corpus design', *Linguistic and literary computing,* 8(4): 259-265.

Cox, C. (2010), 'Probabilistic tagging of minority language data: a case study using Qtag', in: S.Th. Gries, S. Wulff, & M. Davies (eds.) *Corpus linguistic applications: current studies, new directions*. Amsterdam: Rodopi. 213-231.

Cox, C. & A. Berez (2009), 'Software demonstration: CuPED (Customizable Presentation of ELAN Documents'. Paper presented at First International Conference on Language Documentation and Conservation (ICLDC09), University of Hawai'i, March 12-14, 2010. Available online at: http://hdl.handle.net/10125/4969.

Doell, L. (1987), *The Bergthaler Mennonite Church of Saskatchewan: 1892-1975*. Winnipeg: CMBC Publications.

Douglas, F.M. (2003), 'The Scottish Corpus of Texts andSpeech: problems of corpus design', *Linguistic and literary computing*, 18: 23-37.

Epp, R. (1993), *The story of Low German & Plautdietsch: tracing a language across the globe*. Hillsboro, Kansas: Reader's Press.

Evert, S. (2006), 'How random is a corpus? The library metaphor', *Zeitschrift für Anglistik und Amerikanistik*, 54(2): 177-190.

Francis, W.N. & Kučera, H. (1964), *Brown corpus manual*. Providence, Rhode Island: Brown University.

Gordon, R.G. Jr. (ed.) (2005), 'Plautdietsch', in: R.G. Gordon Jr. (ed.) *Ethnologue: languages of the world.* Fifteenth edition. Dallas, Texas: SIL International.

Heid, U. (2008), 'Corpus linguistics and lexicography', in: A. Lüdeling & M. Kytö (eds.) *Corpus linguistics: an international handbook*. Berlin / New York: Walter de Gruyter. 131-153.

Himmelmann, N.P. (1998), 'Documentary and descriptive linguistics', *Linguistics*, 36: 161-195.

Himmelmann, N.P. (2006), 'Language documentation: what is it and what is it good for?', in: J. Gippert, N.P. Himmelmann & U. Mosel (eds.) *Essentials of language documentation*. Berlin / New York: Mouton de Gruyter. 1-30.

Himmelmann, N.P. (2008), 'Reproduction and preservation of linguistic knowledge: linguistics' response to language endangerment', *Annual review of anthropology*, 37: 337-50.

Ide, N., P. Bonhomme & L. Romary (2000), 'XCES: an XML-based encoding standard for linguistic corpora', in: M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhaouer (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*. Paris: European Language Resources Association.

Ide, N., L. Romary & E. de la Clergerie (2004), 'International standard for a Linguistic Annotation Framework', *Natural language engineering,* 10: 211-225.

Ide, N. & K. Suderman (2007), 'GrAF: a graph-based format for linguistic annotations', in: B. Boguraev, N. Ide, A. Meyers, S. Nariyama, M. Stede, J. Wiebe & G. Wilcock (eds.) *Proceedings of the Linguistic Annotation Workshop, Prague, June 28-29, 2007*. Stroudsburg (PA): Association for Computational Linguistics. 1-8.

Johnson, H. (2004), 'Language documentation and archiving, or how to build a better corpus', in: P. Austin (ed.) *Language documentation and description. Volume 2*. London: School of Oriental and African Studies. 140-153.

Johnston, T. (2010), 'From archive to corpus: transcription and annotation in the creation of signed language corpora', *International journal of corpus linguistics*, 15(1): 106-131.

Leech, G. (1991), 'The state of the art in corpus linguistics', in: K. Aijmer & B. Altenberg (eds.) *English corpus linguistics: studies in honour of Jan Svartvik*. London / New York: Longman. 8-29.

Leech, G. (1992), 'Corpora and theories of linguistic performance', in: J. Svartvik (ed.) *Directions in corpus linguistics: proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter. 105-125.

Leonard, W.Y. & E. Haynes (2010), 'Making "collaboration" collaborative: an examination of perspectives that frame linguistic field research', *Language documentation and conservation*, 4: 268-293. Available online at: http://hdl.handle.net/10125/4482.

McEnery, T. & N. Ostler (2000), 'A new agenda for corpus linguistics – working with all of the world's languages', *Literary and linguistic computing*, 15: 403-418.

Nathan, D. (2006), 'Thick interfaces: mobilizing language documentation with multimedia', in: J. Gippert, N.P. Himmelmann & U. Mosel (eds.) *Essentials of language documentation*. Berlin / New York: Mouton de Gruyter. 363-380.

Nieuweboer, R. (1998), *The Altai dialect of Plautdiitsch (West-Siberian Mennonite Low German)*. Doctoral thesis, Rijksuniversiteit Groningen.

Ostler, N. (2009), 'Corpora of less studied languages', in: A. Lüdeling & M. Kytö (eds.) *Corpus linguistics: an international handbook*. Berlin / New York: Walter de Gruyter. 457-483.

Regehr, T.D. (1996), *Mennonites in Canada, 1939-1970: a people transformed.* Toronto: University of Toronto Press.

Reimer, A., A. Reimer & J. Thiessen (eds.) (1983), *A sackful of Plautdietsch: a collection of Mennonite Low German stories and poems*. Winnipeg: Hyperion Press.

Sinclair, J. (2005), 'Corpus and text – basic principles', in: M. Wynne (ed.) *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. 1-16.

Stubbs, M. (2001), *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

TEI Consortium (eds.) (2010), *TEI P5: Guidelines for electronic text encoding and interchange.* Version 1.8.0. November 5, 2010. Available online at: http://www.tei-c.org/Guidelines/P5/

Teubert, W. (2001), 'Corpus linguistics and lexicography', *International journal of corpus linguistics*, 6: 125-153.

Torreira, F. & M. Ernestus (2010), 'The Nijmegen Corpus of Casual Spanish', in: N. Calzolari, B. Maegaard, J. Mariani, J. Odjik, K. Choukri, S. Piperidis, M. Rosner & D. Tapias (eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Paris: European Language Resources Association. 2981-2985.

Woodbury, A. (2003), 'Defining documentary linguistics', in: P. Austin (ed.) *Language documentation and description. Volume 1*. London: School of Oriental and African Studies. 35-51.

Wynne, M. (2005a), 'Preface', in: M. Wynne (ed.) *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. vi-vii.

Wynne, M. (2005b), 'Archiving, distribution, and preservation', in: M. Wynne (ed.) *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. 71-78.