Corpus linguistics and language documentation:
Challenges for collaboration

Christopher Cox • University of Alberta
<christopher.cox@ualberta.ca>

Introduction

- Recent literature in **corpus linguistics** (*e.g.* McEnery & Ostler, 2000) and **language documentation** (*e.g.* Johnson, 2004) suggest both disciplines share natural points of interaction:
  - **Corpus linguistics** as methodological toolkit or as distinct subdiscipline of linguistics, dealing with construction and analysis of consistent collections of linguistic data.
  - **Language documentation** as effort to produce permanent, reusable collections of diverse linguistic data, often motivated by concerns over language endangerment (cf. Himmelmann, 1998; Woodbury, 2003).

Corpus Linguistics and Language Documentation

A match made in heaven?

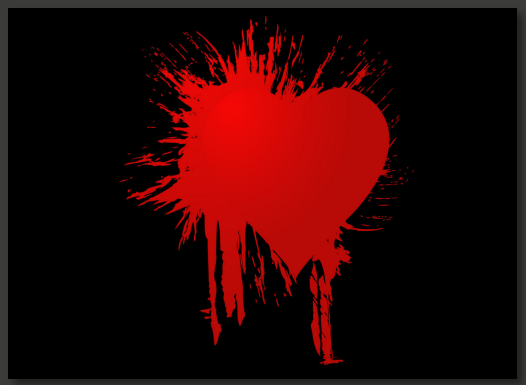
- For corpus linguistics, language documentation offers:
  - ✓ Beautifully diverse, well-catalogued samples of natural language data.
  - ✓ 'Raw material' for corpus construction for underrepresented languages – a standing challenge in corpus linguistics (cf. McEnery & Ostler, 2000)

- For language documentation, corpus linguistics offers:
  - ✓ Another methodological perspective on the documentary record, and another set of tools for its analysis.
  - ✓ Another means of rendering the documentary record accessible, both to academic and non-academic communities – a standing challenge in documentary linguistics (cf. Nathan, 2006)

Corpus Linguistics and Language Documentation

After the honeymoon...

vs.



- Although benefits are anticipated from such collaboration, interaction may not be as simple to foster as it would appear *prima facie*.
  - Strengths of such collaboration lies in distinctiveness of each discipline – distinctiveness which extends to their respective practices, purposes, and participants, and does not guarantee immediate compatibility.

Corpus Linguistics and Language Documentation

- **Purpose of paper:** Consider practical commonalities and differences between corpus linguistics and language documentation in four areas:
  1. **Relationships** between **stakeholders**;
  2. **Methods** of linguistic **sampling**;
  3. **Technologies** conventionally employed;
  4. **Treatment** of linguistic **data and metadata**;

- Examine each point through the lens of ongoing, corpus-based documentation of Mennonite *Plautdietsch* in Canada.

Mennonite Plautdietsch?

- Mennonite Plautdietsch:
  - Traditional language of the **Dutch-Russian Mennonites** (Anabaptist Christian denomination)
  - Est. **300,000 speakers** globally; varying endangerment status
- Language documentation ongoing:
  - Preservation of existing language resources, creation of new records
  - Corpus construction as parallel component of documentation

1. Stakeholder Relationships

- Stakeholder relationships in corpus construction:
  - **Linguist-driven:** decisions as to corpus composition typically made by linguists; content negotiated with copyright holders (*if any*) when and where need arises.

- Stakeholder relationships in language documentation:
  - **Product of partnership:** linguists and community members (where not one and the same) both determine the contents and composition of the corpus, either directly or indirectly.
  - Documentation often politicized (cf. Ostler, 2009), depends upon relationships of trust built upon years, even decades of interaction.

1. Stakeholder Relationships

- Stakeholder relationships in corpus construction:

Mennonite Plautdietsch:

- Some printed materials (e.g. Bible translations) available without extensive consultation
- Language documentation has generally followed from relationships with speakers and community institutions, who ultimately codetermine the final product(s) of documentation.

Documentation often politicized (cf. Ostler, 2009), depends upon relationships of trust built upon years, even decades of interaction.

2. Methods of Sampling

- Sampling in corpus linguistics:
  - Emphasis upon **balance**, **representativeness**, planned composition of corpora.

- Sampling in language documentation:
  - Generally sympathetic to issues of balance and representativeness, but rarely able to control coverage of contexts to extent desired by corpus linguists.
    - Recording communicative events in natural contexts to a certain degree **necessarily opportunistic**;
    - Contents of record codetermined by speech community.

2. Methods of Sar

- Sampling in corpus linguistics:
  - Emphasis upon **balance**, **represent**
    composition of corpora.

> "The corpus builder should retain, as target notions, representativeness and balance. While 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." (Sinclair, 2005: 9)

- Sampling in language documentation:
  - Generally sympathetic to issues of balance and representativeness, but rarely able to control coverage of contexts to extent desired by corpus linguists.
    - Recording communicative events in natural contexts to a certain degree **necessarily opportunistic**;
    - Contents of record codetermined by speech community.

2. Methods of Sar

- Sampling in corpus linguistics:
  - Emphasis upon **balance**, **represent**
    composition of corpora.

"The corpus builder should retain, as target notions, representativeness and balance. While 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." (Sinclair, 2005: 9)

- Sampling in language documentation:
  - to issues of balance and
    t rarely able to control coverage of
    red by corpus linguists.
  - cative events in natural contexts to a
    arily opportunistic;
  - odetermined by speech community.

"In most if not all documentation settings, the range of items that can be documented will be determined to a significant degree by factors that are specific to the given setting, most importantly, the availability of speakers who are willing and able to participate in the documentation effort." (Himmelmann, 2006:4)

2. Methods of Sa

- Sampling in corpus linguistics:

"The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable

Mennonite Plautdietsch:

- Collaboration with Mennonite individuals and institutions far from prevents broad, balanced coverage – but favours documentation of areas of community interest (*e.g.* oral history).
- Sub-sampling potentially offensive – try to justify paring someone's grandfather's life story down to a 2,000-word chunk for the sake of 'balance.'

"I
se
b
to
that
most importantly, the availability
speakers who are willing and able to participate in the documentation effort." (Himmelmann, 2006:4)

arily opportunistic;

odetermined by speech community.

3. Conventional Technologies

- Technologies typical in corpus construction:
  - ✓ **XML-based data and metadata** (*e.g.* XCES, TEI) – though SGML still common
  - ✓ **Time-aligned transcription** of audiovisual materials (*e.g.* using CLAN, as with SBCSAE; using Praat, as with SCOTS)
  - ✓ Distribution via LDC, TalkBank, or individual websites
  - ✓ **Semi-automated annotation tools** for tagging, lemmatization, etc.

- Technologies typical in language documentation:
  - ✓ **XML-based data and metadata** (*e.g.* OLAC, IMDI) – though plain-text records still common (*e.g.* Toolbox databases)
  - ✓ **Time-aligned transcription** of audiovisual materials (*e.g.* using Transcriber, ELAN, or EXMaRALDA)
  - ✓ Distribution via **language archives** (DoBeS, AILLA) and individual websites, though paper copies, CDs, and DVDs also common
  - x **Annotation rarely automated** – corpus tools seldom amenable to changing, provisional analyses (cf. Bird, 2009)

3. Conventional Technologies

- Technologies typical in corpus construction:

Mennonite Plautdietsch:

- Commitment to long-term preservation motivates archival, creation of community-friendly interfaces to records *(including paper!)*
- Between two worlds: choice between tools and formats of corpus linguistics (powerful, but not always user-friendly) and those of language documentation (user-friendly, but not always powerful)

though paper copies, CDs, and DVDs also common

x **Annotation rarely automated** – corpus tools seldom amenable to changing, provisional analyses (cf. Bird, 2009)

4. Treatment of Data and Metadata

- In corpus construction:
  - Data typically **plentiful** (multi-million word corpora)
  - **Highly normative:** distills complexity, heterogeneity of available sources to more tractable, easily comparable corpus samples

- In language documentation:
  - Data comparatively **scarce** (records potentially irreplaceable)
  - Aim to **preserve sources in full diversity**, without reduction in internal complexity, producing permanent collections of featurally-rich 'raw' data

4. Treatment of Data and Metadata

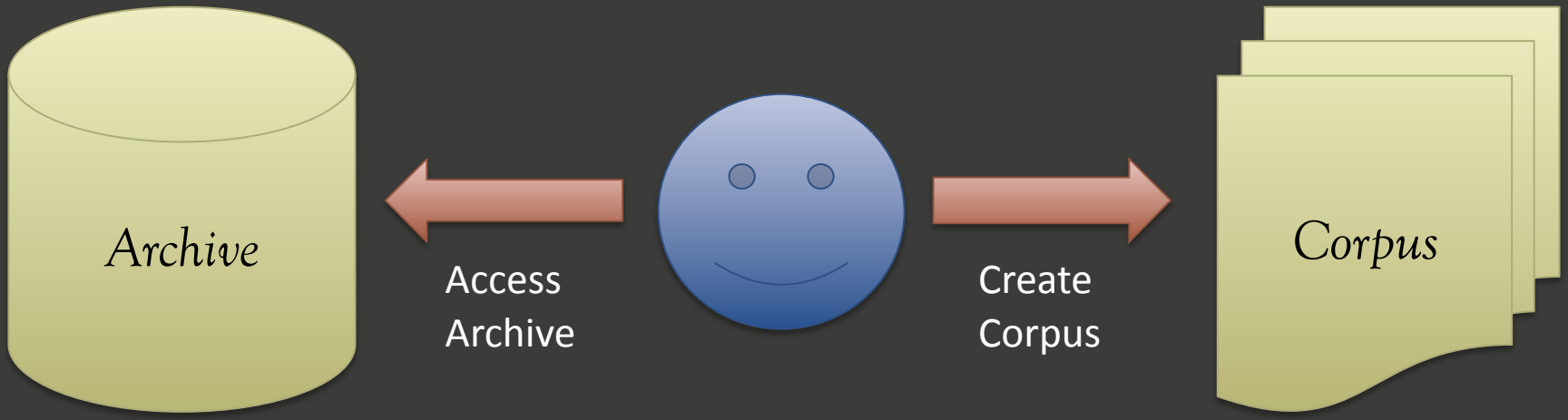
- In corpus construction:

Mennonite Plautdietsch:

- Normalization problematic, especially with controversial issue of orthography: whose standard do you normalize *to*?
- Might instead choose to lemmatize orthographically-diverse sources – but merely displaces the problem to the representation of lemmata, and still prevents reliable non-lemma searches of corpus.

featurally-rich "raw" data

4. Treatment of Data and Metadata



- Tension between reductionist pressures of corpus construction and preservationist tendencies of language documentation risks producing a gulf between disciplines:
  - Where the link between corpus documents and archival sources is not explicit, the creation of a corpus does little to augment the documentary record – a temporary exploitation, rather than a lasting enrichment of those sources.

Corpora and language documentation

- Building corpora on the basis of language documentation:

1. View **corpora as *applications* of language documentation**, built upon permanent documentary records which are referenced wherever possible throughout the corpus.
   - Back-and-forth between corpus documents and archival sources both enriches the documentary record with a corpus interpretation, permits corpus data to be checked against their sources.
   - Normalization considerably less controversial when sources retrievable at any point!

Corpora and language documentation

- Building corpora on the basis of language documentation:

2. **Balance and representativeness** may be **difficult to achieve** in a corpus based on opportunistic, community-partnered documentation – but is arguably not impossible.
   - Although openness to *all* contributions important, documentary linguists can nevertheless advocate broad and balanced coverage for reasons expounded in corpus linguistic literature.
   - Can also create a balanced subcorpus which draws on the documentary record selectively – so long as the whole records remain associated.

Corpora and language documentation

- Building corpora on the basis of language documentation:

3. Technologies and standards for corpus construction and language documentation are similar, though rarely entirely compatible.
   - "The beauty of standards is that there are so many to choose from."

4. Corpus builders in documentary contexts must come to terms with 'politicization' of their work, and with the requirements of community stakeholders.
   - Somewhat atypical in corpus construction, but potentially beneficial: close relationship with stakeholders helps avoid missteps which might otherwise hinder a project.

Conclusion

- Commonality between corpus linguistics and language documentation runs deep, both in technology and data:
  - **Commonality**, however, and **not uniformity**, either in purpose, process, or product; would be a mistake to take these disciplines to be immediately compatible.
  - Bridging the gap between corpus tools and documentary data, and between corpus documents and corresponding archival sources technically feasible, albeit not trivial.
  - Precisely in this divergence between disciplines, however, that we might expect the greatest reward – that the contributions of each might ultimately be greater than the sum of their parts.

Thanks!

References

1. Bird, Steven (2009). "Natural language processing and linguistic fieldwork." *Computational Linguistics* 35(3).469-474.
2. Himmelmann, Nikolaus P. (1998). "Documentary and descriptive linguistics." *Linguistics* 36: 161-195.
3. Himmelmann, Nikolaus P. (2006). "Language documentation: What is it and what is it good for?" *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1-30. Berlin / New York: Mouton de Gruyter.
4. Johnson, Heidi (2004). "Language documentation and archiving, or how to build a better corpus." *Language Documentation and Description. Volume 2*, ed. by Peter Austin, 140- 153. London: School of Oriental and African Studies.
5. McEnery, Tony and Nick Ostler (2000). "A new agenda for corpus linguistics – working with all of the world's languages." *Literary and Linguistic Computing* 15.403-18.
6. Nathan, David (2006). "Thick interfaces: mobilizing language documentation with multimedia." *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 363-380. Berlin / New York: Mouton de Gruyter.
7. Ostler, Nicholas (2009). "Corpora of less studied languages." *Corpus Linguistics: An International Handbook*, ed. by Anke Lüdeling and Merja Kytö, 457-483. Berlin / New York: Walter de Gruyter.
8. Sinclair, John (2005). "Corpus and text – basic principles." *Developing Linguistic Corpora: A Guide to Good Practice*, ed. by Martin Wynne, 1-16. Oxford: Oxbow Books.
9. Woodbury, Anthony (2003). "Defining documentary linguistics." *Language Documentation and Description. Volume 1*, ed. by Peter Austin, 35-51. London: School of Oriental and African Studies.