Automatic discovery of phonological and morphological features in dialect corpora with orthographic normalization

Yves Scherrer, Department of Informatics, University of Oslo, and Department of Digital Humanities, University of Helsinki
yves.scherrer@ifi.uio.no

Although dialectological research increasingly focuses on corpus-based approaches, dialect corpora (typically, transcribed interviews) lend themselves well to quantitative studies than atlases because the different interviews are not directly comparable: if informant A does not use word *x*, it may just be that A chose to talk about topics that did not require the use of word *x*, and not that *x* does not exist in A's dialect. Therefore, comparability between interviews needs to be reinstated, and the methods to do so depend on the investigated level of variation.

In this paper, we propose to use orthographic normalization to provide phonological and morphological comparability in a dialect corpus. Normalization is the annotation of every dialectal word with a canonical word form, for example the standardized spelling of the word, as illustrated in the following example from the Swiss German *ArchiMob* corpus (Scherrer et al. 2019):

Transcription:	jaa	de	het	me	no	gluegt	tänkt	dasch	ez	de	genneraal
Normalization:	ja	dann	hat	man	noch	gelugt	gedacht	das ist	jetzt	der	general
Gloss:	yes	then	has	one	again	looked	thought	this is	now	the	general

For our analysis, we align the transcribed words and their normalized counterparts on the character level, yielding correspondences between pairs of characters and pairs of character n-grams. The frequency distributions of these correspondences vary across dialects and thus can serve as a basis for comparisons. For example, in some Swiss German dialects, /l/ becomes /u/ in certain phonological contexts. In order to define the geographical area in which this /l/-vocalization occurs, it is not sufficient to compute the relative frequency of /u/ in each text, because /u/ also occurs in other, irrelevant phonological contexts. Normalization allows us to define phonological contexts easily and hence to restrict our search to those occurrences of /u/ that are aligned with normalized /l/.

The success of this analysis depends essentially on two factors: the character alignment method and the automatic discovery of dialectologically relevant alignments. For the former, we rely on alignment methods used in statistical machine translation (Koehn et al. 2003; Tiedemann 2009; Scherrer 2023). For the latter, we will demonstrate several methods and apply them to large corpora of Swiss German, Finnish (Institute for the Languages of Finland 2014) and Norwegian (Johannessen et al. 2009) dialects.

References

J. B. Johannessen, J. J. Priestley, K. Hagen, T. A. Åfarli & Ø. A. Vangsnes (2009): *The Nordic Dialect Corpus – an advanced research tool*. In: Proceedings of NODALIDA 2009, 73-80.

Institute for the Languages of Finland (2014): *Samples of Spoken Finnish (speech corpus)*. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2020112937

P. Koehn, F. J. Och, D. Marcu (2003): *Statistical phrase-based translation*. In: Proceedings of HLT-NAACL 2003, 127-133.

Y. Scherrer, T. Samardžić & E. Glaser (2019). *Digitising Swiss German - How to process and study a polycentric spoken language*. In: Language Resources and Evaluation. 53(4).

Y. Scherrer (2023): Character alignment methods for dialect-to-standard normalization. In Proceedings of SIGMORPHON 2023, 110-116.

J. Tiedemann (2009). *Character-based PSMT for closely related languages*. Proceedings of EAMT 2009, 12-19.