Corpus-based computational dialectology with normalization

Yves Scherrer, Department of Digital Humanities, University of Helsinki
yves.scherrer@helsinki.fi

Dialectological research increasingly focuses on corpus-based approaches. Dialect corpora typically consist of transcribed interviews and aim to represent realistic, everyday speech grounded in linguistic context (Szmrecsanyi & Anderwald 2018). However, dialect corpora do not lend themselves well to quantitative studies 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.

In his seminal work on quantitative corpus-based dialectology, Szmercsanyi (2013) relies on syntactic annotation to make dialect corpora comparable and to infer geographical distributions of syntactic patterns. In this paper, we focus on the quantitative analysis of more traditionally studied linguistic levels, namely phonology and morphology. Consequently, we propose to use **orthographic normalization** rather than syntactic annotation to provide comparability. Several dialect corpora, e.g., the Swiss German *ArchiMob* corpus (Scherrer et al. 2019) or the *Samples of Spoken Finnish* collection (Institute for the Languages of Finland, 2014) include either manual or semi-automatic normalization annotations on the word level.

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 ArchiMob:[1]

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 transcribed and normalized characters and character n-grams. The frequency distributions of these correspondences vary across dialects and thus can serve as a basis for comparisons between dialects. 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/. This gives us a clearer and more accurate picture of the geographical extent of /l/-vocalization.

The success of this analysis depends essentially on two factors: the character alignment method and the automatic discovery of dialectologically relevant alignments. Ideally, the character alignment method can identify many-to-many correspondences, such as those occurring between a diphthong and a long vowel. We will apply alignment methods initially developed for phrase-based statistical machine translation (Koehn et al. 2003; Tiedemann 2009) and grapheme-to-phoneme conversion (Jiampojamarn et al. 2007) to this task. We will also test different weighting schemes to discover and extract character (n-gram) correspondences that show regional variation. These findings can then be compared with

[1] The normalization language used in ArchiMob is similar, but not identical to Standard German.

traditional atlas-based dialect classifications.

References

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

S. Jiampojamarn, G. Kondrak, T. Sherif (2007): *Applying many-to-many alignments and hidden Markov models to letter-to-phoneme conversion.* In: Proceedings of HLT-NAACL 2007, 372-379.

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).

B. Szmrecsanyi (2013). *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge University Press.

B. Szmrecsanyi & L. Anderwald (2018). "Corpus-based approaches to dialect study". In: C. Boberg, J. Nerbonne & D. Watt (eds.), *The Handbook of Dialectology*. Wiley-Blackwell.

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