Continuous variation in computational morphology - the example of Swiss German

Yves Scherrer

Continuous variation in computational morphology
The example of Swiss German

Yves Scherrer
ALPAGE, Université Paris 7 Diderot & INRIA
Rue Albert Einstein, F-75013 Paris
yves.scherrer@inria.fr

Keywords

Continuous variation, Finite-state technology, Rule-based systems, Dialects, Swiss German

1 Introduction

Most work in natural language processing is geared towards written, standardized language varieties. This focus is generally justified on practical grounds of data availability and socio-economical relevance, but does not always reflect the linguistic reality of sub-standard varieties. In this paper, we aim at the computational description of the morphology of a language with continuous internal variation, as it is encountered in most dialect landscapes. The work presented here is applied to Swiss German dialects; these dialects are well documented through dialectological research and are among the most lively ones in Europe in terms of social acceptance and media exposure.

Our work is based on previous research in generative dialectology and computational linguistics (Section 2), but extends this work in several respects (Sections 3 and 4). We propose a finite-state model that contains georeferenced transformation rules for morphological, but also lexical and phonetic phenomena. We will focus on two aspects of this model: its theoretical value as a computationally effective description of continuous linguistic variation, and its practical value as a word-level machine translation system from Standard German into the various Swiss German dialects. We evaluate the model on the latter aspect (Section 5).

2 Previous work

In the 1970s and 1980s, **generative dialectology** was introduced as a rule-based approach to dialectological description (Veith, 1970, 1982). Its main idea is to derive multiple dialect systems from a single **reference system** with the help of transformation rules. For example, the following rule indicates that the lexeme *Töpfer* 'potter' of the reference system B (Standard German) is realized as *Häfner* in the dialect systems D_{33333} to D_{46999} (Veith, 1982, 280) :

(1) $\#\text{Töpfer}\#_B \rightarrow \#\text{Häfner}\#_{D_{33333-46999}}$

A single rule base thus allows to fully describe a large number of dialect systems, like for instance the 50 000 varieties of Wenker and Wrede's survey of German dialects. Besides lexical transformations like the one in (1), Veith (1982) analyzes sound change in terms of phonetic feature transformations and uses morphosyntactic features in morphological transformation rules. These rules apply in cascade. The order of the rules – a thorny issue in rule-based NLP systems – receives a precise interpretation here: it corresponds to the historical order in which dialect changes have propagated.

Vaillant (2008) adopts a similar approach to build a multi-dialectal grammar of five French-based Creole languages of the West-Atlantic area: each syntactic rule contains a numeric parameter that specifies in which dialect(s) it is valid. Vaillant (2008) thus conceives the different dialects as discrete entities which can be clearly distinguished.

While this view is justified for Caribbean creoles spoken on different islands, it cannot be maintained for dialect areas that lack major topographical and political borders, such as German-speaking Switzerland.

Both proposals make use of transformation rules which are highly ambiguous, but can be disambiguated on the basis of numeric dialect markers. These transformation rules may be called **georeferenced**, in the sense that each dialect marker links to a pair of geographic coordinates that can be grounded on a map.

3 Our model

Our work differs from the above-mentioned research in one crucial aspect: we maintain that dialectal variation is continuous in many cases, and that discrete dialect markers (e.g., in the form of integer numbers, as above) mask parts of the dialectal variation. More crucially, Veith (1982)'s proposal is restricted to the locations in which the dialect survey has been conducted, and is not able to predict the dialectal system of non-surveyed areas. In summary, instead of subscripting the rules with discrete dialect markers, we prefer to subscript them with a probabilistic map. This allows to handle transition zones in which several variants are accepted. The following example illustrates this approach:

(2) *immer* [Standard German] → *geng* 

The Standard German word *immer* 'always' is transformed to Swiss German *geng* in the black area on the map, corresponding roughly to the Cantons of Berne and Fribourg.

In contrast to Veith (1982), we provide a full implementation of our model. Three questions related to the implementation arise: (i) Where do we get the probability maps from? (ii) How are the rules implemented? (iii) How can we measure the performance of the resulting system?

Since we have already described the map acquisition process in detail elsewhere (Scherrer and Rambow, 2010a,b), we focus on questions (ii) and (iii) in the following sections.

4 Implementation

Rule-based computational descriptions of morphologies are usually implemented with the help of finite-state machines (Hopcroft et al., 2006; Beesley and Karttunen, 2003). These methods have become popular in computational phonology and morphology because of their highly optimized algorithms and the availability of several toolkits. Most of these toolkits also define specific notational shortcuts to facilitate the efficient writing of linguistic rules. Our implementation is based on the XFST toolkit (Beesley and Karttunen, 2003).

We have defined the practical goal of automatically translating words from Standard German to Swiss German. This has two consequences on the proposed finite-state system:

First, the proposed system not only contains morphological rules for affix generation, but also lexical and phonetic rules that generate the dialectally adequate roots. In this sense, it is actually more than just a morphological generator.

Second, the goal of transforming Standard German lemmas into Swiss German word forms naturally commands the choice of Standard German as the reference system. However, this choice is not accurate etymologically, since the Swiss German dialects did not develop out of Standard German, but rather of their common ancestor, Middle High German. In any case, this decision is a practical one that does not challenge the theoretical underpinnings of the proposed system.

An example of a morphological affixation rule is displayed below. It assumes that the input consists of the word stem and a set of morphosyntactic features that have been determined beforehand (by analyzing the Standard German source word). The rule adds the inflectional ending *-i* or *-0* (no ending) to weak singular adjectives:

(3) `define adj-2-flex [ ADJA [Nom | Acc] Sg Gender Degree Weak -> [ 0 | i ]];`

This example merely specifies two dialectal variants, but does not indicate their respective geographical areas of validity. As the finite-state toolkits are not intended for handling geographical information, we use a special notational device, **flag diacritics** (Beesley, 1998), to indicate the file system path of the map:

(4) define adj-2-flex [ADJA [Nom | Acc] Sg Gender Degree Weak ->
[0 "@U.3-254.0@" | i "@U.3-254.i@"]];

Composition algorithms allow several rules to be applied in cascade in an efficient way. We use finite-state composition to derive all potential dialectal outcomes of a given Standard German word. The drawback of this approach is that the probability maps cannot be consulted during this step, leading for example to derivations that combine an Eastern Swiss German root with a Western Swiss German affix. To this end, a second type of composition, **map composition**, is applied. The idea is that when composing two rules, the result is valid in the area defined by the intersection of the maps associated with the two rules. Since every map corresponds to a two-dimensional matrix containing probability values, map composition amounts to computing the pointwise product (Hadamard product) of the two matrices. Derivations for which the map composition yields an empty map are discarded.

5 Evaluation

Most common machine translation metrics consider words as atomic entities. Our use of phonetic and morphological transformations would be heavily penalized by such measures, since a single transformation error is weighted just as heavily as several errors in the same word. We choose to use Longest Common Subsequence Ratio (LCSR) as evaluation measure, as tentatively put forward by Tiedemann (2009) in a similar setting. This measure counts the proportion of (ordered) identical letters, not words, between the system output and the reference translation.

We have evaluated the coverage of our model on the basis of a multi-dialectal corpus consisting of 100 sentences each in five dialects, extracted from the Swiss German Wikipedia (`als.wikipedia.org`). The dialect classification was done directly by the Wikipedia writers. These original dialect texts (OD) were then manually translated back to Standard German (SG).

The experiment is designed as follows. First, the SG texts are syntactically analyzed and automatically transformed to five Swiss German dialects, corresponding to the administrative centers of the respective dialect regions. The resulting translations (RD) can then be compared (a) to the SG version they have been created from, and (b) to the OD version (in one of the five dialects). We put forward two hypotheses in relation with these two types of comparison:

1. The OD texts are more similar to their RD counterparts than to the SG texts.
2. The OD-RD text pairs from matching dialect regions are more similar than the OD-RD text pairs from unmatching regions.

Up to now, both hypotheses could be verified on three of the five dialects. While this result is promising, it also shows that some transformation rules are not robust enough and that a few important rules are still missing.

References

Beesley, K. R. (1998). Constraining separated morphotactic dependencies in finite-state grammars. In *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*, Ankara, Turkey.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications, Stanford.

Hopcroft, J. E., Motwani, R., and Ullman, J. D. (2006). *Introduction to Automata Theory, Languages and Computation*. Pearson Addison-Wesley, Boston, 3 edition.

Scherrer, Y. and Rambow, O. (2010a). Natural language processing for the Swiss German dialect area. In *Tagungsband der Zehnten Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2010)*, Saarbrücken.

Scherrer, Y. and Rambow, O. (2010b). Word-based dialect identification with georeferenced rules. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, Cambridge, MA, USA.

Tiedemann, J. (2009). Character-based PSMT for closely related languages. In *Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009)*, pages 12 – 19, Barcelona.

Vaillant, P. (2008). Grammaires factorisées pour des dialectes apparentés. In *Actes de la 15e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2008)*, pages 159–168, Avignon.

Veith, W. H. (1970). *-Explikative +Applikative +Komputative Dialektkartographie*, volume 4 of *Germanistische Linguistik*. Olms.

Veith, W. H. (1982). Theorieansätze einer generativen Dialektologie. In Besch, W., Knoop, U., Putschke, W., and Wiegand, H. E., editors, *Dialektologie – Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, pages 277–295. De Gruyter, Berlin, New York.