Creating a part-of-speech tagged and parsed corpus of historical code-switching data

Mareike Keller
Institute for English Linguistics
University of Mannheim
markelle@mail.uni-mannheim.de

Sarah Schulz
Institute for Natural Language Processing (IMS)
University of Stuttgart
schulzsh@ims.uni-stuttgart.de

As has been pointed out in several studies (Wenzel 1994, Schendl & Wright 2011, Jefferson & Putter 2013), historical mixed texts are an interesting, yet still widely unexplored, source of information concerning language use in multilingual societies of medieval Europe. One reason for the dearth of linguistically oriented studies in the field is the lack of digital corpora. This contribution presents an approach to the creation of a part-of-speech tagged and parsed corpus of historical code-switching. The aim of the project is to create gold standard data for the systematic testing of existing code-switching theories, and to provide guidelines for the annotation of historical mixed language texts.

To set the scene we will briefly address automatized language recognition (cf. Solorio et al. 2008) and the choice of an appropriate tagset, small enough to make cross-linguistic generalizations yet at the same time detailed enough to provide enough information to the user (cf. Petrov et al. 2012; Jamatia et al. 2015). Then we will zoom in on theoretical considerations involved in the selection of a parser. This includes the choice of a theoretical model of grammar and its implications for possible analyses that then can(not) be carried out on the corpus. Even though the theoretical code-switching model we adopt is based on constituency, dependency-based parsers have clearly yielded superior results so far. If time permits we will also address the pros and cons of either using two parsers, one for each language, or one joined parser which then has to be designed to capture the grammatical structures of more than one language at a time (cf. Goyal et al. 2003).

To illustrate our points we present results from a collection of macaronic sermons (Horner 2006, Middle English and Latin, ca. 100,000 words), a text genre containing diverse code-switching and language mixing structures which is thus highly informative both for multilingualism research and for computational linguistics. In the following excerpt Middle English word forms appear in roman, Latin word forms in italic. Non-distinctive elements at switch-points are marked by an underscore:

> *Sepe* men fallen in despaire *pro* losse of good and dignite. *Si vis nauigare* sauelich <u>in</u> *istis maribus et euadere hec pericula, oportet* strike seil <u>in</u> *primo mari* and cast ankur <u>in</u> *secundo*. *Tunc pro combinacione partium et processu sermonis, dico primo*: strike sail in þe perlus see of welth and prosperite þat þi schip ouerturne not with þe wynde of pride and vanite. Cast ankur in God *dum nauigas mare aduersitatis* þat þou periche not be despair for losse of good or dignite. Qwo so hath sailed both þes sees and an biden bittur stormys, *experiencia docet eum* to telle of many perels. (Horner 2006, p. 521)

In conclusion we will address some problematic items like particles, adverbs and idioms which are well-known challenges to computational linguistics, and point out possible solutions, balancing our linguistic ambition against the potential and limitations of automatic language processing.

References:

Goyal, P., Mital, M.R., Mukerjee, A., Raina, A.M., Sharma, D., Shukla, P. & K.V. Saarthaka (2003). A Bilingual Parser for Hindi, English and code-switching structures. In *Proceedings*

of the 11th Conference of the European Chapter of the Association for Computational Linguistics, April 12-17, 2003.

Horner, P. (2006). *A macaronic sermon collection from late medieval England*. Toronto.

Jamatia, A., Gambäck, B. & A. Das (2015). Part-of-Speech Tagging for Code-Mixed English-Hindi Twitter and Facebook Chat Messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

Jefferson, J.A. & A. Putter (eds.) (2013). *Multilingualism in medieval England (c. 1066-1520): Sources and analysis*. Turnhout.

Keller, M. (forthcoming). Code-switched adjectives and adverbs in macaronic sermons. In C. Delesse & E. Louviot (eds.), *Studies in Language Variation and Change 2*. Newcastle upon Tyne.

Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford.

Myers-Scotton, C. (2008). Language contact: Why outsider system morphemes resist transfer. *Journal of Language Contact* Thema 2: 21-41.

Petrov, S., Das, D. & R. McDonald (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Schendl, H. & L. Wright (eds.) (2011). *Code-switching in early English*. Berlin, Boston.

Schulz, S. & M. Keller (submitted): Code-switching ubique est - Language identification and part-of-speech tagging for historical mixed text. *In Proceedings of Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. Berlin.

Solorio, T. & Y. Liu (2008). Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 973-981.

Vyas, Y., Gella, S., Sharma, J., Bali, K. & M. Choudhury (2014). POS Tagging of English-Hindi Code-Mixed Social Media Content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wenzel, S. (1994). *Macaronic Sermons: Bilingualism and preaching in late-medieval England*. Ann Arbor.