Speeding up corpus development for linguistic research: language documentation and acquisition in Romansh Tuatschin

Géraldine Walther
University of Zürich
Plattenstrasse 54
8032 Zürich, Switzerland
geraldine.walther@uzh.ch

Benoît Sagot
Inria
2 rue Simone Iff
75 012 Paris, France
benoit.sagot@inria.fr

Abstract

In this paper, we present ongoing work for developing language resources and basic NLP tools for an undocumented variety of Romansh, in the context of a language documentation and language acquisition project. Our tools are designed to improve the speed and reliability of corpus annotations for noisy data involving large amounts of code-switching, occurrences of child speech and orthographic noise. Being able to increase the efficiency of language resource development for language documentation and acquisition research also constitutes a step towards solving the data sparsity issues with which researchers have been struggling.

1 Introduction

Contemporary linguistic research relies more and more heavily on the exploration of statistical patterns in language. The non-categorical distribution and variety of linguistic units has become a focus for understanding complex variations across patterns, such as dative alternations (Bresnan et al., 2007) or cases of optionality (Wasow et al., 2011). Studies like these require the availability of large consistently annotated corpora.

For lesser or undescribed languages however, such resources are not readily available. What is worse, the current rate of language extinction could lead to the disappearance of 20-90% of today's spoken languages by the end of the 21st century (Krauss, 1992). Documenting endangered languages will allow us to preserve traces of the current language diversity, a part of the world's cultural heritage. Building reliable linguistic resources will allow us to study them according to fast evolving research standards.

Similarly, studies on language acquisition are based on recorded, transcribed, and annotated data of parent-child interactions. Language acquisition research has produced significant databases, such as the CHILDES project (McWhinney, 2000), yet mainly manually and at enormous costs.

Relying solely on manual annotators is too costly an option. Minimising resource development costs is crucial. Manual language resource development for language documentation and language acquisition projects should be sped up, as soon as technically possible, by employing NLP tools such as spelling correction/normalisation tools and part-of-speech (POS) taggers. For instance, POS taggers used as pre-annotators have been shown to increase both annotation speed and annotation quality, even when trained on limited amounts of data (Marcus et al., 1993; Fort and Sagot, 2010).

Yet language acquisition and language documentation data presents specific challenges for automatic linguistic annotation. Firstly, such data usually consists of transcriptions of spontaneous speech. Secondly, previously undescribed languages are often not written and lack established orthographies, resulting in noisy transcriptions. Thirdly, acquisition data consists of recordings of child-parent interactions. The recorded target children's language production can differ dramatically from adult language, adding another layer of linguistic variation. Finally, as new data is usually still being collected, available raw and even more so annotated data is rare, which significantly limits the available training data for annotation tools.

In this paper, we show the interaction between manual resource development (morphological lexicon, spelling and POS-annotated corpus) and automatic tools on current annotation experiments for a language documentation and acquisition project on the undocumented and previously

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 89–94, Vancouver, BC, August 4, 2017.

non-written Romansh dialect of Tuatschin.

2 Romansh Tuatschin

The term Romansh denotes a set of Romance languages with important Germanic lexical and grammatical influence, mostly spoken in the canton of the Grisons in South-Eastern Switzerland. Although Romansh is considered one of the four official national languages of Switzerland, the term Romansh covers in fact a variety of languages and dialects with significantly differing features. The dialect we focus on in present paper corresponds to a previously undocumented dialect of the Romansh Sursilvan variety called Tuatschin. It is spoken by approximately 1,500 speakers in the Val Tujetsch area. Contrary to the neighbouring main Sursilvan dialect, which is also the main language in local schools, Tuatschin is at this point an unwritten language. It is however still natively spoken and transmitted both in the local area and within families who have left and settled in larger cities within the country. Speakers are proud of their language and culture and promote it through a local cultural association and occasional, non-normalised, publications in the local newspaper.

The development of the resources described here is part of a project which combines language documentation and acquisition research. One aim of this project is to gain better understanding of intergenerational language transmission in endangered language contexts, which might contribute to eventually slowing down, if not reversing, linguistic and cultural erosion of minority languages. The data used in this project is mostly original data from fieldwork in the Val Tujetsch, original recordings within five Tuatschin speaking families with at least one child aged between 2 and 3 years, and updated and normalised lexical data from the Tuatschin word-list by Caduff (1952).

Adult speakers of Tuatschin are usually multilingual, natively speaking at least two, if not more, Romansh dialects (Tuatschin and Sursilvan), as well as German (both High German and the local variety of Swiss German), and often a fair amount of Italian and French. Everyday conversations in the Tuatschin speaking area comprise a high amount of code-switching. Conversations between speakers of neighbouring dialects more often than not result in each speaker speaking their own variety, accommodating the other person's dialect to varying degrees. Insertion of German lexical items or full utterances is ubiquitous.

As a result, our recorded Tuatschin data comprises a high amount of German and standard Sursilvan. This is even more acute in the acquisition corpus data used within present experiments, as the children's families tend to be natively bilingual. While the children's mothers are all native Tuatschin speakers, the fathers are Sursilvan, Swiss German or Italian speakers. The children therefore produce a significant amount of mixed utterances. In addition to the high amount of code-switching due to this particular language setting, language acquisition data also comprises intrinsic noise due to the differences between child and adult language, including nonce-words or specific child-speech. The variation observed in our corpus data ranges from language and dialectal variation to adult/child register differences. Developing automated tools for such a corpus requires coping with noisy, heterogeneous corpus data.[1]

3 Developing guidelines

3.1 Orthography

The first challenge in developing linguistic resources and automated tools for a previously unwritten language like Tuatschin consists in developing an orthography that can be used for transcribing recorded data, and training future transcribers (native speakers) in using this new orthography. We wanted this orthography to also be usable by native speakers outside the project. In collaboration with two native speakers within the Val Tujetsch, we developed a new orthography for Tuatschin, which is mainly based on the orthography for the neighbouring written dialect Sursilvan, but accommodates the phonetic and morphological differences of Tuatschin, from the pronunciation of specific vowels and diphthongs to diacritic marking of infinitive forms. Once the main principles of the orthography had been established, we started training our native corpus transcribers. However, without complete resources (such as a full lexicon or grammar) at their disposal, each one of them still had their own interpretation of the overall principles for the tran-

[1] Note that while Sursilvan does have an established orthography and is used a a language of instruction in schools, there are no automatic tools available for the language. The existing online lexicon can only be queried online but is not freely available. Despite the languages' similarity there are no existing resources that could be leveraged to facilitate automatic work on Tuatschin at this point.

scription of individual words, adding an additional transcriber-related layer to the variation within the data. Developing the new orthography also required several passes, based on feedback from our transcribers and progress in our own understanding of the language data through our ongoing field work. Subsequent changes contributed to variation even within an individual transcriber's orthography, however reducing the differences between different transcribers' orthographic strategies.

3.2 Annotation tagset inventories

For our corpus annotation, we developed two separate tagsets, one for part-of-speech (POS) annotation and one for morphosyntactic features. Just as for our orthographic conventions, our tagset evolved alongside field work progress while annotation was already ongoing, adding further noise to our corpus data. This kind of noise is a common problem in language documentation projects where data collection, annotation, and analysis are conducted in parallel. Without the availability of automatic tools, it normally requires several passes of manual post-cleaning and adds to the overall cost of language resource development.

Our POS tagset comprises a fine-grained and a coarse POS inventory. The full coarse-grained inventory, used for our POS tagging experiments (see Section 5.2) is the following: ADJ, ADV(erb), COMPL(ementiser), CONJ(unction), DET(erminer), INTER(jection), N(oun), PN (proper noun), PREP(osition), PRN (pronoun), PUNCT(uation), QW (question word), SOUND, V(erb). Fine-grained tags add information such as DET_def for definite articles or PREP_loc for locative prepositions. It also comprises a specific _childspeech refinement of all POS for words that are specific to child-speech. The morphosyntactic features follow the Leipzig Glossing Rule conventions commonly used in language documentation projects and comprise distinctions for number and gender, as well as tense, mood and person.

Although we have designed this language-specific inventory as a means to better model the morphological and morphosyntactic properties of Tuatschin, we recognise the relevance and importance of the Universal Dependency (UD) initiative,[2] whose aim is to "develo[p] cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective." In the current version (2.0), several dozen languages are already covered, sometimes with more than one treebank. What is more, the UD website announces a Romansch and a (distinct) Sursilvan Romansch treebank. Our initiative is not related to these treebank development efforts—we had independently decided to develop a deterministic mapping between our language-specific label inventory and a UD-compatible annotation scheme, in order to pave the way for a future Romansch-Tuaschin UD corpus. Together with the dependency annotation of our data, this will be the focus of future work.

[2] http://universaldependencies.org

4 Manual ressource development

4.1 TuatLex

We manually developed a morphological lexicon for Tuatschin. For that, we devised an explicit grammatical description of Tuatschin nominal and verbal morphology based on our own field work data. We implemented this description within the Alexina$_{PARSLI}$ lexical framework (Sagot and Walther, 2013). In addition to the implemented morphological description, our Tuatschin lexicon, TuatLex, comprises a list of 2,176 lemmas, based on an updated version of the Tuatschin word-list by Caduff (1952) complemented with our newly collected data, among which 780 verbs, 949 nouns, and 146 adjectives. The Alexina inflection tool uses the grammar to produce 46,089 inflected entries, among which 29,361 verbal, 15,137 nominal, and 762 adjectival forms.

4.2 Manual corpus annotation

After devising the tagset, we trained advanced linguistic undergraduate students to manually annotate our corpus data. The students were no native speakers of Romansh. They were asked to annotate each token for (fine-grained) POS and morphosyntactic information and to indicate an English translation for each token's base form.[3] Using the WebAnno annotation tool (Eckart de Castilho et al., 2014), annotators took approximately 15 to 25 hours for annotating files containing typically 1,000 to 2,000 words. Difficulties for annotators mainly came from orthographic vari-

[3] For example, for the Tuatschin token *nûrsas*, they would have provided the following annotations: POS = N, Morphosyn = pl, and the English translation 'EWE'.

ation, variation in the representation of word tokens,[4] and dialectal variation.[5] Weekly meetings between two trained linguists, among whom one Sursilvan native speaker, and all annotators were organised to compare notes, discuss recurrent difficulties, and adapt the tagsets whenever necessary.[6]

4.3 Manual corpus normalisation

In order to simplify the annotation task, we set up a procedure for systematic orthographic correction. We first asked one of the native speakers, who had been involved in the development of the orthographic conventions, to manually correct already transcribed corpus data for orthographic errors and individual-word-based code-switching in a separate *normalised* tier within the corpus.[7] We then manually introduced an additional tier indicating for each 'normalised' token whether it had been corrected for orthography, code-switching, child-speech, or actual pronunciation errors. This intermediate layer, in addition to being useful for subsequent acquisition or code-switching studies, is meant to help the development of an automatic spell-checker. Aside from variation in the usage of diacritics, some of the most frequent errors in the transcribed data involved the amalgamation of words meant to be written as separate tokens.[8]

5 Automation and tools

5.1 Tokenisation and orthographic correction

In order to speed up the manual development of orthographically normalised corpora based on new transcriptions, but also to prepare the fully automatic processing of non-annotated text, we first developed an automatic tokenisation and spelling standardisation/correction tool. It is implemented in the form of a Tuatschin-specific parametrisation of the SXPipe shallow processing architecture (Sagot and Boullier, 2008).

Note that the spelling standardisation/correction tool is not meant to be used as a standalone tool. It has been designed and developed only for speeding up future corpus development and for serving as a cleaning step before applying the POS tagger, whose results are obviously better on (even only partially) normalised data.

Our spelling standardisation/correction tool relies on a list of deterministic rewriting patterns that we automatically extracted and manually selected based on the data described in Section 4.3. More precisely, we applied standard alignment techniques for extracting n-to-m correspondences between raw and normalised text. Among the extracted rules, 695 were deterministic, which means that there was a unique output in the corpus for a given input. Out of these 695 candidate rules, we manually selected 603, whose systematic application could not result in any over-correction.[9] Several others are non-deterministic. However, a careful manual examination of these candidate ambiguous rules showed that most of the ambiguity is an artifact of the rule extraction process, and that true ambiguities can be resolved in context.[10] As a result, contextual information was included in our standardisation/correction patterns, thus resulting in a fully deterministic standardisation/correction module.

5.2 POS tagger

In order to assess and improve the quality of the POS annotation, but also to have a POS pre-annotation tool for speeding up future manual annotation, we trained the MElt POS tagger (Denis and Sagot, 2012) on the manually POS-annotated data available so far, using coarse POS to reduce sparsity (see Section 3.2). The 2,571 already annotated sentences, containing 9,927 tokens, were divided in training, development and test sets by randomly selecting sentences with re-

[4] E.g. *vegni* instead of *vegn i* 'it comes' (lit. 'comes it').

[5] For example, unexpected morphological forms that prevented the recognition of inflected forms such as verbs ending in Sursilvan *-el* instead of Tuatschin *-a* in the first person singular.

[6] The project being mainly a documentation project on a previously undescribed language, data collection, annotation, and data analysis are currently being carried out in parallel. Data annotation in particular has been performed as a collaborative task rather than as a task conducted by individual annotators, that would have to be evaluated for inter-annotator agreement.

[7] The purpose of this normalised layer is solely to help automatic (and, to a lesser extent, manual) annotation of the data. It is not meant to replace the original transcription layer, which remains the relevant layer for subsequent linguistic analysis.

[8] Cf. the *vegni*/*vegn i* example from previous footnote.

[9] A few examples: *stù*→*stu*, *sèl*→*sè'l*, *schia*→*schéia*.

[10] For instance, *vegnì* and all other verbs from TuatLex's inflection class `Vi` can produce ambiguous tokens such as *vegni*, which must be changed into *vegnì* (infinitive) before pronouns such as *ju*, *té*, *el*, *ella*, *el*, *i*, *nus*, *vus*, *els*, *ellas*, *ins*, but, in most other contexts, must be rewritten as *vegn i* (V+PRN) with an expletive pronoun *i*. We applied the latter to those verb tokens likely to appear with expletive subjects like *vegnì* 'to come' while inserting the infinitive diacritic on all other verb instances (such as *capì* 'understand').

spective probabilities 0.8, 0.1 and 0.1. We also extracted from the TuatLex lexicon described above a set of 19,771 unique (form, POS) pairs, to be used by MElt as a complementary source of information, in addition to the training corpus.

We trained a first version of MElt, and applied the resulting POS tagger on the training data themselves. Annotation mismatches allowed us to identify several systematic errors in the training data. Some of them came from individual errors or annotator disagreement. But most were due to changes made in the annotation guidelines while manual annotation was already ongoing, and of which some had not yet been correctly retroactively applied to already annotated data.[11] We applied a series of POS correction rules to the whole corpus (train+dev+test) as a result of this training-corpus-based study, and re-trained MElt on the corrected data. The result is a POS tagger trained on 2,062 sentences (7,901 words) with a 91.7% accuracy overall and a 65.3% accuracy on words unknown to the training corpus. Interestingly, if trained without TuatLex lexical information, accuracy figures do not drop as much as usually observed (Denis and Sagot, 2012): respectively 91.6% and 62.5%. This suggests that lexical information might not be as important for improving POS taggers for child-related speech as it is for tagging standard text, a fact likely related to the more limited vocabulary size in such corpora.

6 Conclusion

We have described ongoing efforts for developing language resources (lexicon, annotated corpus)[12] and basic NLP tools for the Romansh variety of Tuatschin, in the context of a project on language description and language acquisition dedicated to a previously non-written language. Our next step will consist in using our tools for pre-annotating new raw data, in order to speed up annotation while increasing its quality and consistency. They will also be used for creating automatically annotated data, which will complement the manually annotated corpus.

On a longer term, our tools are also meant to be used for automatically categorising tokens and sequences of tokens into occurrences of child-speech or various types of code-switching, relying on the information comprised within the intermediate tier introduced during the normalisation procedure (see Section 4.3). This intermediate layer is meant to be ultimately automatically generated by SxPipe. The richer and more accurate information that our tools will be able to provide will also facilitate subsequent quantitative linguistic studies on Romansh Tuatschin, its acquisition by children and its influences by surrounding languages, especially Sursilvan and (Swiss) German.

[11] For instance, the previously defined POS "V_particle" had been discarded at some point during corpus development, yet it still had a number of occurrences in the training data.

[12] The lexicon and tools will all be made freely available.

7 Acknowledgements

The authors gratefully acknowledge the SNF project number 159544 "The morphosyntax of agreement in Tuatschin: acquisition and contact" whose funding has been instrumental to the work presented here. We also thank our project collaborators and especially our annotators Nathalie Schweizer, Michael Redmond, and Rolf Hotz for their help in annotating the data, and Claudia Cathomas for her extremely valuable help and native expertise on Romansh data. For their help in developing a new Tuatschin orthography we would like to express our warmest thanks to Ciril Monn and Norbert Berther.

References

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, Royal Netherlands Academy of Science, Amsterdam, pages 69–94.

Léonard Caduff. 1952. *Essai sur la phonétique du parier rhétoroman de la Vallée de Tavetsch (canton des Grisons - Suisse)*. Francke, Bern.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation* 46(4):721–736.

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych, and Seid Muhie Yimam. 2014. Webanno: a flexible, web-based annotation tool for clarin. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*. CLARIN ERIC, Utrecht, Netherlands, page online. Extended abstract.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth ACL Linguistic Annotation Workshop (LAW IV)*. Uppsala, Sweden, pages 56–63.

Michael Krauss. 1992. The worlds language in crisis. *Language* (68).

Mitchell Marcus, Béatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

Bryan McWhinney. 2000. *The CHILDES Project: Tools for analyzing talk.*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.

Benoît Sagot and Pierre Boullier. 2008. SxPipe 2: architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues* 49(2):155–188.

Benoît Sagot and Géraldine Walther. 2013. Implementing a formal model of inflectional morphology. In Cerstin Mahlow and Michael Piotrowski, editors, *Proceedings of the Third International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2013)*. Humboldt-Universität, Springer-Verlag, Berlin, Germany, volume 380 of *Communications in Computer and Information Science (CCIS)*, pages 115–134.

Thomas Wasow, T. Florian Jaeger, and David Orr. 2011. Lexical variation in relativizer frequency. In H. Simon and H. Wiese, editors, *Expecting the Unexpected: Exceptions in Grammar*, de Gruyter, Berlin, pages 175–195.