O. Scrivner

TOOLS FOR DIGITAL HUMANITIES: PARALLEL CORPUS AND VISUALIZATION

Abstract. The merging of corpus linguistic methods and digital technology can provide new ways of representing medieval digital texts. In this paper, we introduce a multi-layered parallel Old Occitan-English corpus. We show how parallel alignment can help overcome some challenges associated with historical manuscripts. Furthermore, we apply a resource-light method of building an emotion annotation via parallel alignment. Finally, using visualization tools such as ANNIS and GoogleViz, we demonstrate how our parallel corpus can be queried and visualized dynamically via modern language.

Keywords. Parallel corpora, word alignment, corpus visualization, emotion annotation

1. Introduction

In the past, researchers took exclusively non-digital approaches to the study of historical documents and manuscripts. However, recent achievements in corpus linguistics have introduced state-of-the-art methods and tools for digitization and text processing. Similarly, advances in digital technology have opened up novel ways of visualizing and interpreting data. Furthermore, «by accessing linguistic annotation», we have extended «the range of phenomena that can be found» [Kübler and Zinsmeister 2014]. That is, we have learned that a digital corpus enriched with linguistic annotation, such as syntactic, pragmatic and semantic, can enhance our understanding of the literary or historical work. Finally, it has been shown that the creation of parallel corpora not only helps researchers overcome some of the challenges of historical manuscripts, e.g., differences in spelling or lexical variation, but also makes texts accessible to audiences with no prior knowledge of a given historical language. Until recently, parallel corpora have been exclusively

used in the field of machine translation, bilingual lexicography, and translator training, as well as in studies of language-specific translational phenomena. With the increase of available historical parallel corpora, e.g. original texts and their modern translations, we have seen the emergence of their use in historical linguistics [Zeldes 2007].

In this paper we introduce a parallel annotated Old Occitan-English corpus. We further show how the alignment with Modern English makes historical corpora more accessible and how word alignment can also facilitate the cross-language transfer of emotion annotation from a resource-rich modern language into a resource-poor language, such as Old Occitan. Finally, we demonstrate how emotion visualization techniques can contribute to a richer understanding of literary texts. The remainder of this paper is organized as follows: section 2 reviews the concept of a parallel corpus and its exploitation in historical studies; section 3 describes the compilation of a parallel corpus and emotion annotation; section 4 introduces queries and visualization methods via ANNIS and GoogleViz; section 5 provides an outlook on future steps for the project.

2. Using Parallel Corpora in Historical Linguistics

Parallel corpora are collections of two or more parallel texts that contain an original text (source) and its translations. The essential part of a parallel corpus is the alignment between a source text and its translation. One can identify the following schema of word alignment: i) between two single words (one-to-one), ii) between a single word and a multi-word unit (one-to-many), iii) between a multi-word unit and a single word (many-to-one) and iv) zero alignment. It is commonly agreed that there exists a link between form and meaning in any language. In a translated text, we can assume that the meaning of the translation is approximately the same as that of the original. These links can be further displayed and exploited in various ways, e.g. contrastive studies, sense disambiguation, lexicography and translation study, among others

[Lawson 2001]. Furthermore, parallel corpora can assist researchers with no prior knowledge of a historical language, as medieval documents can be queried via the modern translation rather than via the older language. Moreover, given the common assumption that «translation equivalents are likely to be inserted in the same or very similar, syntactic, semantic and pragmatic contexts», we can assess not only lexical, but also morphological variations [Enrique-Arias 2013]. That is, it is possible to i) identify forms that have never been studied, ii) find occurrences based on their textual or stylistic conventions and iii) search for null occurrences in a source language via explicit tokens in the modern language.

3. Parallel Occitan-English Corpus

In this project we focus on the Romance of Flamenca. This anonymous romance, written in the 13th century, presents an artistic amalgam of fabliau, courtly romance, troubadours' lyrics and narrative genre. The compilation and architecture of the monolingual Old Occitan Romance of Flamenca corpus has been described in [Scrivner et al. 2013]. In this paper, we focus on the augmentation of the Romance of Flamenca corpus with a parallel Old Occitan-English level. First, in the selection of source translations it was important to find the closest translation to the original poem as possible. We chose the work by Blodgett for several reasons. Blodgett «endeavored, so far as possible, to respect the loose and often convoluted syntax of the original» [Blodgett 1995: xli]. Furthermore, Blodgett followed a conservative approach and omitted lines that were suggested in earlier editions to replace lacunae in the original manuscript.

Second, a completely manual word alignment is a costly and time-consuming process. On the other hand, there do not exist any automatic aligners for the Old Occitan-English pair. In addition, most of the literature on alignment methods focuses on modern languages and non-lyric genres. So we decided to apply a semi-automatic approach. We selected GIZA++ [Och et al. 2000], a

freely available automatic aligner, which allows for one-to-one and one-to-many word alignment. The output of automatic alignment was further corrected manually. The GIZA++ output before correction is illustrated in (2), and the corrected version is shown in (3):

1) *poissas lur dis tot en apert*
   1 2 3 4 5 6
2) then ({1}) he ({}) said ({3}) to ({4}) them ({2}) openly ({6})
3) then ({1}) he ({3}) said ({3}) to ({2}) them ({2}) openly ({6})

Finally, we augmented our corpus with emotion annotation. While emotion analysis constitutes an important component in the study of literary genre, narrative corpora annotated for emotional content are not very common. In recent years, however, with the increasing access to digitized books, e.g., Google Books Corpus and Project Gutenberg, there has been growing interest in applying emotion annotation to narrative stories. For example, Alm and Sproat [2005] annotate 22 of Grimm's fairy tales, demonstrating the importance of story sequences for emotional story evaluation, and Francisco et al. [2011] create a corpus of 18 English folk tales. Both corpora are built using a manual annotation. In contrast, Mohammad [2012] applies a lexicon-based method to the emotion analysis of Google Books. As Mohammad states, emotion annotation in the literary domain can be used for multiple purposes: 1) emotion search and analysis, 2) social analysis, 3) comparative analysis, 4) summarization, 5) word persuasion analysis and 6) gender analysis.

While emotion annotation can be a valuable resource in linguistics and literary studies, annotated corpora and emotion lexicons exist mainly for resource-rich languages, such as English. However, given the assumption that the translated word shares a concept or sense with the original word, word alignment can be used as a bridge for emotion transfer. In order to develop such a bridge, we first compiled a word list from the English version of

the text and removed common function words. We further used the NRC English emotion lexicon, which consists of words and their associations with 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) and positive or negative sentiment[1]. We further transferred emotion labels into the Occitan version via a word alignment. Finally, we imported our corpus into the ANNIS search engine [Zeldes et al. 2009], which makes this corpus accessible online[2]. At present, our parallel corpus contains 14 100 tokens and 1 000 aligned verse lines. Our corpus consists of several layers of annotations: i) a (morpho-)syntactic layer (part-of-speech and constituency annotations for Occitan and part-of-speech for English), ii) lemmas, iii) a discourse layer (speaker classification, e.g. king, queen, Flamenca), iv) temporal sequencing (events), v) word alignment (Occitan → English), vi) an emotion layer (joy, trust, fear, surprise, sadness, disgust, anger and anticipation) and vii) a sentiment layer (negative and positive).

4. Corpus Query and Visualization

In recent years we have seen increasing interest in visual and dynamic applications in corpus and literature studies. For example, Moretti [2005] advocates for the use of maps, trees and graphs in literary analysis. Oelke and Kokkinakis [2012] suggest person network representation, showing the relations between characters. Furthermore, Hilpert [2011] introduces dynamic motion charts into the linguistic field. These charts are common in the socio-economic statistic field and make it possible to visualize in motion the development of a phenomenon across time. As Hilpert points out, «the main purpose of producing linguistic motion charts is to give the analyst an intuitive understanding of complex linguistic development». In this paper, we also argue that the merge of novel visualization techniques and parallel corpora not only allows the

1 http://saifmohammad.com/WebPages/lexicons.html

2 www.oldoccitancorpus.org

reader to discover medieval work, but also facilitates access to rich historical and literary information for a large audience without a prior knowledge of a medieval language.

First, alignment can be used to search for variation in name spelling. For example, by querying for the English word *flamenca*, we find spelling variants in the Occitan version, e.g. *flamencha*, as illustrated in Figure 1.

Fig. 1. Word Alignment: Sample of Results for the English word *flamenca*

Second, we can examine a morpho-syntactic variation, e.g. null subjects. For example, we can search for all English pronouns that are linked to Occitan verbs, as compared to English explicit pronominal subjects that are linked to Occitan pronouns. Furthermore, we can search for socio-linguistic and semantic contexts. To illustrate this example in greater detail, our query spans several layers of annotation: emotion (*joy*), speaker (*Father*), event (*First News*) and token alignment. The sample of results is illustrated in Figure 2. In addition, the overall emotional content can be assessed by querying for *emotion! = None* and performing the frequency analysis.

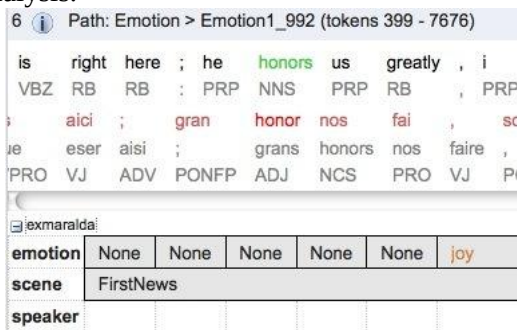

Fig. 2. Search for *Joy* with the Speaker *Father*

Finally, our emotion analysis can be assessed on a dynamic time plot, by using R [R Development Core Team, 2008] and the GoogleViz package[1]. We converted our data into R data frame format and produced a motion chart, which is available for local downloads. At present, the chart allows for the capture of changes in emotion across time. In the future we plan to add discourse and token information.

5. Conclusion

We have presented a novel approach to analyzing Old Occitan medieval texts via parallel word alignment. Using ANNIS we have shown how our multi-layered corpus can be queried via a user-friendly query builder. Finally, we have presented a motion chart, which allows for the dynamic analysis and tracing of data. Our future goal is to continue developing visual annotation, facilitating access to medieval documents for a large audience.

References

1. *Alm C., Sproat R.* (2005), Emotional sequencing and development in fairy tales. Affective Computing and Intelligent Interaction, Tao J., Tan T., Picard R. (eds.), Vol. 3784, pp. 668–674. Springer.

2. *Enrique-Arias A.* (2013), On the usefulness of using parallel texts in diachronic investigations: Insights from a parallel corpus of Spanish medieval Bible translations. New Methods in Historical Corpora. Durrell P., Scheible M., Whitt S., Bennett R. (eds.), pp. 105–116.

3. *Francisco V., Hervás R., Peinado F., Gervás P.* (2011), EmoTales: Creating a corpus of folk tales with emotional annotations. Language Resources and Evaluation, 46, pp. 341–381.

4. *Hilpert M.* (2011), Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate

[1] http://cran.r-project.org/web/packages/googleVis/index.html

data from diachronic corpora. International Journal of Corpus Linguistics, 16, pp. 435–461.

5. *Kübler S., Zinsmeister H.* (2014), Corpus Linguistics and Linguistically Annotated Corpora. Bloomsbury Academic.

6. *Lawson A.* (2001), Collecting, aligning and analyzing parallel corpora. Small Corpus Studies and ELT. Theory and practice, pp. 279–309. Philadelphia: John Benjamin's.

7. Kübler S., Zinsmeister H. (2014), Corpus Linguistics and Linguistically Annotated Corpora. Bloomsbury Academic.

8. *Scrivner O., Kübler S., Vance B., Beuerlein E.* (2013), Le Roman de Flamenca: An Annotated Corpus of Old Occitan. Proceedings of the Third Workshop on Annotation of Corpora for Research in Humanities, Mambrini F., Passarotti M., Sporleder C. (eds.), pp. 85–96.

9. *Zeldes A., Ritz J., Lüdeling A., Chiarcos C.* (2009), ANNIS: A Search Tool for Multi-Layer Annotated Corpora. Proceedings of Corpus Linguistics.

Olga Scrivner
Indiana University (USA).
E-mail: **obscrivn@indiana.edu**