Tools for Digital Humanities: Enabling Access to the Old Occitan Romance of Flamenca

Olga Scrivner
Indiana University
Bloomington, IN, USA
obscrivn@indiana.edu

Sandra Kübler
Indiana University
Bloomington, IN, USA
skuebler@indiana.edu

Abstract

Accessing historical texts is often a challenge because readers either do not know the historical language, or they are challenged by the technological hurdle when such texts are available digitally. Merging corpus linguistic methods and digital technology can provide novel ways of representing historical texts digitally and providing a simpler access. In this paper, we describe a multi-dimensional parallel Old Occitan-English corpus, in which word alignment serves as the basis for search capabilities as well as for the transfer of annotations. We show how parallel alignment can help overcome some challenges of historical manuscripts. Furthermore, we apply a resource-light method of building an emotion annotation via parallel alignment, thus showing that such annotations are possible without speaking the historical language. Finally, using visualization tools, such as ANNIS and GoogleViz, we demonstrate how the emotion analysis can be queried and visualized dynamically in our parallel corpus, thus showing that such information can be made accessible with low technological barriers.

1 Introduction

In the past, historical documents and manuscripts were studied exclusively by a manual, paper-based approach. This limited the access to such documents to scholars who, on the one hand, know the historical variety of the language and, on the other hand, had access to the manuscripts. However, recent achievements in corpus linguistics have introduced new methods and tools for digitization and text-processing. Similarly, the progress in digital technology has created novel ways of data visualization and data interpretation. Finally, "by accessing linguistic annotation, we can extend the range of phenomena that can be found" (Kübler and Zinsmeister, 2014). That is, a digital corpus enriched with linguistic annotation, such as syntactic, pragmatic, and semantic annotation, can amplify our understanding of the literary or historical work. With such a corpus, a researcher can overcome many of the challenges provided by the historical nature of the manuscripts as well as by the technological barrier provided by many available query tools. For example, one of the challenges in working with historical texts consists in the variation in spelling or in lexical variation. In such cases, instead of performing a direct lexical search, researchers can access this information via linguistic annotation if lemma information is available.

However, there are challenges for which standard linguistic annotations are not useful. First, searching in text is usually restricted to known phenomena or explicit occurrences of data. For example, some phenomena may involve a variation between explicit and implicit tokens, e.g., null subjects or zero relative clauses. While some corpora allow for a query of null occurrences, such as in the MCVF (Martineau et al., 2010), most of the monolingual corpora do not provide such an annotation. Furthermore, it is essential for the researcher to be aware of all possible forms and contexts for queries, which is a difficult task in monolingual corpora. That is, there is always a possibility that "some relevant form will be overlooked because it has never been studied"

(Enrique-Arias, 2013, 107). For example, consider the study of Medieval Spanish discourse markers, e.g., *he*, *evás* 'behold'. Enrique-Arias (2013) shows that only by using a parallel corpus is he able to observe unexpected linguistic structures conveying the same discourse function, e.g., *sepas que* 'know that' and *cata que* 'see that'. Finally, monolingual corpora are usually accessible only to an audience with prior knowledge of a given historical language, leaving a large public aside.

We propose to address these challenges by using a parallel corpus. In our case, the two documents consist of the original, historical text and its translation into modern English. Until recently, parallel corpora have been almost exclusively used in the fields of machine translation, bilingual lexicography, translator training, and the study of language specific translational phenomena (McEnery and Xiao, 2007). With the increased availability of historical parallel corpora, we have seen the emergence of their use in historical linguistics (Koolen et al., 2006; Zeldes, 2007; Petrova and Solf, 2009; Enrique-Arias, 2012; Enrique-Arias, 2013).

In this paper, we introduce a parallel annotated Old Occitan-English corpus. We show how the alignment with modern English makes this historical corpus more accessible and how the word alignment can facilitate the cross-language transfer of emotion annotations 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 the literary text for technically less inclined readers.

The remainder of this paper is organized as follows: Section 2 reviews the use of corpora in historical studies, and section 3 describes work on transferring annotations via alignment. Section 4 describes the textual basis of our corpus, the 13th-century Old Occitan *Romance of Flamenca*. In section 5, we explain the compilation of the parallel corpus. Section 6 describes the emotion annotation, which was carried out for English and then transferred to Old Occitan. Section 7 introduces emotion queries via ANNIS, a freely available on-line search platform, and visualization methods with motion charts in GoogleViz. Section 8 draws general conclusions and provides an outlook on future steps for the project.

2 Using Corpora in Historical Linguistics

Parallel corpora are collections of two or more texts consisting of an original text (source) and its translation(s). Generally, such parallel corpora are annotated for word *alignment* between the source text and its translation. Word alignment can be carried out automatically via tools such as GIZA++ (Och and Ney, 2000).

Monolingual historical corpora are undoubtedly valuable linguistic resources. It is not uncommon, however, to encounter different spellings and other lexical variations in historical texts. Not knowing an exact spelling or just a simple language barrier may hinder data collection. Parallel corpora can assist in such situations through Historic Document Retrieval (Koolen et al., 2006), which allows researchers to query via the modern translation rather than via the older language. Given the common assumption that "translation equivalents are likely to be inserted in the same or very similar, syntactic, semantic and pragmatic contexts" (Enrique-Arias, 2013, 114), we can assess not only lexical, but also morphological variations. That is, it is possible to a) identify forms that have never been studied and b) find occurrences based on their textual or stylistic conventions. For example, using the parallel Bible corpus of Old Spanish and its English translation, Sánchez López (To Appear) is able to identify a new form, *salvante*, that has never been reported.

Similarly, Enrique-Arias (2012) examines discourse markers, possessive structures and clitics in the Latin Bible and its medieval Spanish translation. In addition, the author is able to observe stylistic variation in choices made by translators depending on Bible sub-genres such as narrative or poetry.

In recent years, there has also been increasing interest in the correspondence between translation and language change. In this view, translation is seen as a "means of tracking change" (Beeching, 2013). Various studies have demonstrated the feasibility of parallel corpora in studies of semantic and (morpho-)syntactic change. For example, Beeching (2013) examines the evolution of the French expression *quand-même* using monolingual and parallel corpora. Similarly, Zeldes (2007) looks at the parallel corpus of Bible translations in two different stages of the same language, Old and Modern Polish. The author is able to detect various changes in nominal affixes. Another interesting approach is suggested by Petrova and Solf (2009), who investigate the influence of Latin in historical documents. When we deal with historical documents that are translations of Latin or other languages, it is hard to assess which phenomena are target language specific or introduced via the translation from the source language. Petrova and Solf (2009) show that this issue can be resolved with a parallel diachronic corpus: They analyze a change in word order in the Old High German translation of the Bible and its original Latin version. Given the assumption that any word order deviation in the translation can be viewed as evidence for Old High German syntax, their investigation is restricted to cases where word order in the translation differs from its original. The results reveal that in contrast to Latin, preverbal objects in subordinate clauses in Old High German convey *given* information (explicitly mentioned in the previous context), whereas postverbal objects carry *new* information.

Finally, linguists and computational linguistics can benefit from parallel corpora in studies of implicit constituents, e.g. zero anaphora. Not all corpora are annotated for implicit occurrences or the omission of certain elements in a sentence. This is a common challenge in the fields, where zero anaphora resolution is necessary, e.g., automatic summarization, machine translation, and studies of syntactic variation. With a parallel corpus, it is possible to search for the explicit form in a translated text and then observe the use or omission of that form in the original text. For instance, in his study of Biblia Medieval, a parallel corpus of Old Spanish Bible translations, Enrique-Arias (2013) analyzes discourse markers and observes instances of zero-marking in Old Spanish by searching for explicit Hebrew markers. Furthermore, such corpora can be used to investigate the convergence universal in machine translation, a correlation between zero anaphora and the degree of text explicitation (Rello and Ilisei, 2009).

We argue that the addition of a translation into a modern language is a simple and intuitive way of giving access to historical texts. This is a useful tool not only for historians and historical linguists but also for a lay audience since the translation provides access to the meaning without introducing additional hurdles such as by a semantic annotation. In section 4, we will present our corpus of choice, the *Romance of Flamenca*, an Old Occitan poem, along with the parallel version that includes a modern English translation. Then, we will present an approach to annotate this corpus for emotions, using non-experts working on the English part and then transferring the annotation to the source language.

3 Using Alignment Methods for Annotating Resource-Poor Languages

Linguistic annotation often involves a great amount of manual labor, which is often not feasible for low-resourced languages. Instead, we can use a method from computational linguistics, namely cross-language transfer, as proposed by Yarowsky and Ngai (2001). This method does not involve any resources in the target language, neither training data, a large lexicon, nor time-consuming manual annotation.

Cross-language transfer has been previously applied to languages with parallel corpora and bilingual lexica (Yarowsky and Ngai, 2001; Hwa et al., 2005). This approach uses parallel text and word alignment to transfer the annotation from one language to the next. Yarowsky and Ngai (2001) show the transfer of a coarse grained POS tagset and base noun phrases from English to French. Yarowsky et al. (2001) extend the approach to Spanish and Czech, and they include named entity tagging. Hwa et al. (2005) use a similar approach to transfer dependency annotations from English to Chinese. Snyder and Barzilay (2008) extend the approach to unsupervised annotation of morphology in Semitic languages via a hierarchical Bayesian network. And Snyder et al. (2009) extend the framework to include multiple source languages.

In previous work (Scrivner and Kübler, 2012), we used another cross-language transfer method, based on the work by Feldman and Hana (2010), to annotate the Flamenca corpus with POS and syntactic annotations. This method does not require parallel texts, it rather uses resources such as lexical or POS taggers from closely related languages. Our annotations were based on resources from Old French (Martineau et al., 2010) and modern Catalan (Civit et al., 2006).

In machine translation, the transfer of sentiment analysis is common in machine translation.Kim et al. (2010) use a machine translation system to map subjectivity lexica from English to other languages. In word sense disambiguation, word alignment has been used as a bridge (Tufis, 2007) based on the assumption that the translated word shares the same sense with the original word. A similar method was used for sentiment transfer from a resource-rich language to a resource-poor language (Mihalcea et al., 2007).

4 Romance of Flamenca

Medieval Provençal literature is well known for its lyric poetry of troubadours. There remains, however, a small number of non-lyric provençal texts, such as *Romance of Flamenca*, *Girart de Rossilho* and *Daurel et Beto*, that have not received much attention. In this project we focus on *Romance of Flamenca*, which can be faithfully described as "the one universally acknowledged masterpiece of Old Occitan narrative" (Fleischmann, 1995). This anonymous romance, written in the 13th century, presents an artistic amalgam of fabliau, courtly romance, troubadours lyrics, and narrative genre. The uniqueness of this "first modern novel" is also seen in its use of setting, adventures, and character portrayal (Blodgett, 1995). The narrator virtuously depicts *Archambaut*'s transmutation from a chivalrous knight to an unbearably jealous husband who locks his beautiful wife *Flamenca* in a tower, as well as *Guilhem*'s ingenious conspiracy to liberate *Flamenca*. Furthermore, this prose in verse played an influential role in the development of French literature. The potential value of this historical resource, however, is limited by the lack of an accessible digital format and linguistic annotation.

There are no known records of *Romance of Flamenca* before the late 18th century, when it was seized from a private collection during the French Revolution and placed later in the library of Carcassonne (Blodgett, 1995). The manuscript came in 139 folios with 8095 octosyllabic verses but missing the beginning and end. It was first edited by Raynouard in 1834. Since then, the manuscript has been edited multiple times (Meyer, 1865, 1901; Gschwind 1976; Huchet, 1982). While Gschwind's edition (1976) remains "the most useful edition", which provides a more accurate interpretation (Blodgett, 1995; Carbonero, 2010), we have chosen the second edition by Meyer for two reasons: a) The edition has no copyright restriction and b) this edition is available in a scanned image format provided by Google[1].

5 Parallel Occitan-English Corpus

The compilation and architecture of the monolingual Old Occitan corpus *Romance of Flamenca* along with the morpho-syntactic and syntactic annotations has been described by Scrivner and Kübler (2012) and Scrivner et al. (2013). In this paper, we augment the monolingual version into the *Romance of Flamenca* corpus with a parallel Old Occitan-English level. As described in section 2, we regard monolingual and parallel corpora as complementary resources. That is, our new corpus conforms to the traditions of a conventional multi-layered monolingual corpus, and at the same time, it offers the research advantages of a parallel bilingual corpus.

In our task, we have made various methodological decisions related to translation and alignment. First, in the selection of a translation of the source, it was important to find the most faithful translation to the original poem. While free translations have their own merits, they pose a great challenge to the alignment task. Our choice fell to the work by Blodgett (1995) for several reasons. Blodgett "endeavored, so far as possible, to respect the loose and often convoluted syntax of the original" (Blodgett, 1995). In addition, the author was able to add lines from the manuscript that were missing in the previous editions. Finally, Blodgett (1995) followed a "conservative approach" and omitted lines that were suggested earlier to replace lacunae in the original. This conservative approach is necessary for ensuring the accurate line alignment of verses.

In a second step, we provided word alignment. This is a challenging task, partly because of the non-standardized spelling in the Old Occitan source, but also because the amount of aligned text is rather small for standard unsupervised approaches. An ad-

[1] https://archive.org/details/leromandeflamen00meyegoog

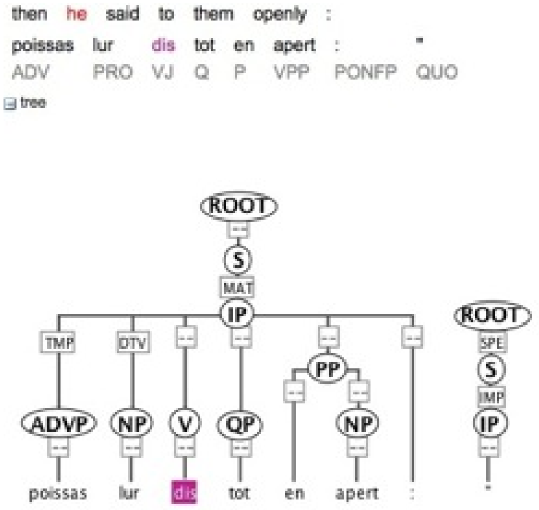

then he said to them openly :
poissas lur dis tot en apert : "
ADV PRO VJ Q P VPP PONFP QUO
tree
ROOT
S
MAT
IP
TMP
DTV
PP
ADVP
NP
V
QP
NP
ROOT
SPE
S
IMP
IP
poissas
lur
dis
tot
en
apert
:
"

Figure 1: Word alignment: sample of results for the null pronoun in Old Occitan.

ditional challenge results from the verse structure of the poem, which necessitates deviations in the translation. This genre is prone to various stylistic word orders, as compared to political or historical narratives. In addition, sentence boundaries in Occitan do not always correspond to those in the English translation.

If we followed common practice in automatic alignment and chose sentences as the basic text units for the automatic alignment, this very likely would result in many mis-alignments. As a result, we decided to split data by lines, instead of sentences. We performed the line alignment by means of NATools[2]. NATools is a Perl package for processing parallel corpora. This package helps with the alignment of two files on a sentence level and with extracting a probability lexicon.

For word alignment, after some experimentation, we decided to use a fully unsupervised approach since there do not exist any automatic aligners for the Old Occitan-English pair. We chose GIZA++ (Och and Ney, 2000), a freely available automatic aligner, which allows for one-to-one and one-to-many word alignment. In addition, when using GIZA++, we can make use of our extracted probability dictionary. The output of the automatic alignment was then corrected manually. Below, we show an example. In (1), we show the original sentence with the word index as well as the English translation. The GIZA++ output before correction is illustrated in (2), and the corrected version is shown in (3). In both versions, the numbers in parentheses behind a word indicate which word in the original is aligned with this word in the translation.

(1) index: 1 2 3 4 5 6
OO: poissas lur dis tot en apert
ME: then he said to them openly

(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)

This example shows that in (2), the subject pronoun *he* is not aligned with any word in Old Occitan. This is to be expected since Old Occitan is a pro-drop language, and the pronoun is not expressed overtly. However, during our manual correction, we align the pronoun *he* with the verb *dis*, a standard treatment for null subject pronouns. In a final step, we combine lines with corrected word alignment to form a sentence in order to merge this parallel alignment with our monolingual Old Occitan corpus, as illustrated in Figure 1.

This example also shows that such an alignment can be used to search the Old Occitan corpus via the modern English annotated translation. For example, the query with an explicit English pronoun (PRP)

[2] http://linguateca.di.uminho.pt/natools/

Storm & Storm (1987)	Shaver et al. (1987)	Plutchik (1980)
sadness	sadness	sadness
anger	anger	anger
fear	fear	fear
happiness	happiness	joy
love	affection	disgust
disgust		trust
anxiety		surprise
contentment		anticipation
hostility		
liking		
pride		
shame		

Table 1: Basic emotion models.

aligned to the Occitan verb (VJ) allows to find null occurrences of subject pronouns in Old Occitan.

At present, our parallel corpus contains 14 100 tokens and 1 000 aligned verse lines. Our corpus is further converted to PAULA XML (Dipper, 2005) and imported into the ANNIS search engine (Zeldes et al., 2009), which makes this corpus accessible online[3].

6 Emotion Annotation Transfer

While emotion analysis constitutes an important component in literary analysis, narrative corpora annotated for emotional content are not very common. In contrast, there is a large body of work on emotion and sentiment analysis of non-literary resources, such as blog posts, news and tweets (see (Liu, 2012; Pang and Lee, 2008) for overviews). However, despite the advances in the automatic annotation, the manual annotation of emotions remains a difficult task. On the one hand, the definition of emotion remains a controversial issue as there is still no clear distinction between emotions, attitude, personality, and mood. Various models of emotion clusters have been proposed, as illustrated in Table 1, but no clear standard has emerged so far.

On the other hand, the assignment of emotion is often a subjective decision. While many emotions can be identified through contextual and linguistic cues, e.g., lexical, semantic, or discourse cues, it has been shown that human annotators often assign a different label for the same emotional context (Alm and Sproat, 2005, 670). Finally, available annotated resources are domain specific, e.g., movie reviews and poll opinions, which makes it difficult to adapt to a narrative genre. As Francisco et al. (2011) point out, "the complexity of the emotional information involved in narrative texts is much higher than in those domains".

In recent years, however, with the increasing access to digitized books, such as the Google Books Corpus and Project Gutenberg, there has been growing interest in applying emotion annotation for narrative stories. For example, Alm and Sproat (2005) annotate 22 Grimms' fairy tales and demonstrate 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. He creates an emotion association lexicon with 14 200 word types, which determines the ratio of emotional terms in text. In addition, Mohammad shows effective visualization methods that trace emotions and compare emotional content in a large collections of texts.

As we have seen, the emotion annotation can be a valuable resource in linguistics and literary studies. However, annotated corpora and emotion lexica exist mainly for resource-rich languages, such as English. Annotating verses in Old Occitan manually for emotional content is a tedious task and requires an expert in the language. Thus neither a manual annotation of the text nor the creation of an emotion lexicon in the source language is viable. However, we can use a resource-poor approach from NLP, namely cross-language transfer (see section 3, which allows us to take advantage of English resources in combination with the word alignment. I.e., we can annotate emotions in English, which is easy enough to do given a lexicon and an undergraduate student. In a second step, we transfer the annotation via the word alignments to the source language.

Below, we will describe our emotion annotation transfer method. First, we compiled a word list from the English version *Flamenca* and removed common function words. We then used the NRC emotion lex-

[3] www.oldoccitancorpus.org

icon[4], which consists of words and their associations with 8 emotions as well as positive or negative sentiment. Mohammad and Yang (2011) created this lexicon by using frequent words from the 1911 *Roget Thesaurus*[5], which were annotated by 5 human annotators. For our application, we focus on the 8 emotion associations: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. Since the emotion annotation is not neutral with regard to context, several emotions can be assigned to the same token in the NRC lexicon, as shown in (4).

	abandon	fear
	abandon	sadness
(4)	lose	anger
	lose	disgust
	lose	fear

As we can see, the word *abandon* has two associations, namely with *fear* and *sadness*, and the word *lose* has three possible associations, namely *anger*, *disgust* and *fear*. During the initial label transfer from the NCR lexicon to the Flamenca lexicon, we kept multiple labels. The tokens with multiple labels were further manually checked, and only one label that would best fit the context to our knowledge was retained. For example, in case of 'abandon', we selected the emotion *sadness*, as the context describes Famenca's father before her marriage. Also the evaluation of 100 randomly selected labels revealed that some associations did not fit our text due to the difference in genre. For example, 'court' in our narrative genre represents a different semantic entity (king's court), whereas in the NCR lexicon, 'court' (a criminal case) is associated with *anger* or *fear*. We decided to leave these cases unannotated.

Finally, the emotion annotation was transferred from the English translated words to their aligned words in Old Occitan. The emotion annotation layer was further added to the main corpus and converted to the ANNIS format.

7 Corpus Query and Visualization

In this section, we will focus on how our parallel corpus can be queried for emotional content. Following the approach from Alm and Sproat (2005), who show the relevance of textual sequencing for emotional analysis, we have segmented the corpus into 10 logical event sequences, namely wedding announcement, preparation for wedding, arrival of Archambaut, marriage, departure, King's arrival, and Queen's jealousy. The corpus consists of different layers of annotations: a) (morpho-)syntactic layer (part-of-speech and constituency annotation for Occitan and part-of-speech for English), b) lemmas, c) discourse layer (speakers classification, e.g., king, queen, Flamenca), d) temporal sequencing (events), e) word alignment (Occitan → English), and f) emotion layer (joy, trust, fear, surprise, sadness, disgust, anger and anticipation). For visualization, we use the search engine ANNIS (Zeldes et al., 2009), which is designed for displaying and querying multilayered corpora. One advantage of ANNIS consists of its graphical query interface, which allows for user-friendly, graphical queries, instead of traditional textual queries, for which one needs to know the query language. To illustrate the visual query tool, we present a query in Figure 2. In this query, we search for any aligned token that expresses *joy* and is spoken by *Father*.

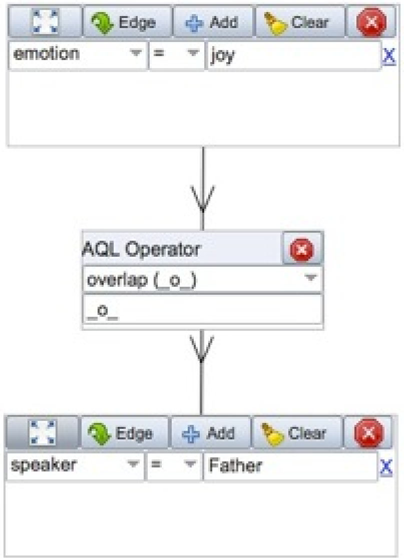
One example of a query result is illustrated in Figure 3. This example shows the Occitan token *honor* and the aligned English token *honors* that are annotated with *joy* and spoken by *Father*.

Another way of analyzing emotions more generally is to look at the overall emotional content, by querying for any words that have emotion; in other words, they are not annotated as "None" for emotion (query: $emotion! = "None"$). We can then perform a frequency analysis of the results. The frequency distribution is shown in Figure 4, where we see that *trust* and *joy* are the most common emotions.

In recent years, visual and dynamic applications in corpus and literature studies have become more important, thus showing a focus on non-technical users. For example, (Moretti, 2005) advocates the use of maps, trees, and graphs in the literary analysis. Oelke et al. (2012) suggest person network representation, showing relations between characters. In addition, they also use fingerprint plots, which allow to compare the mentioning of various categories, e.g. male, female, across several novels. Hilpert (2011) introduces dynamic motion

[4] http://saifmohammad.com/WebPages/lexicons.html

[5] The words must occur more than 120 000 times in the Google n-gram corpus.



emotion = joy

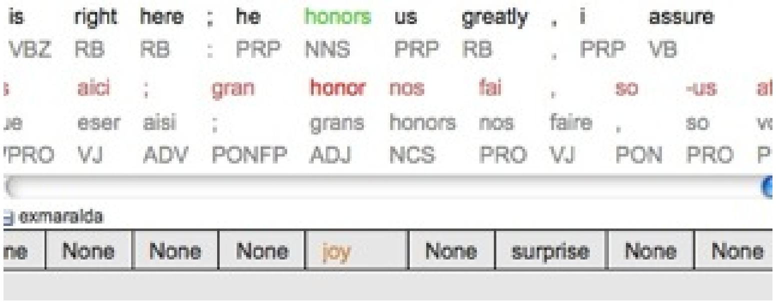
AQL Operator

overlap (_o_)

_o_

speaker = Father

Figure 2: Search for *joy* with the speaker *Father*.



is right here ; he honors us greatly , i assure

VBZ RB RB : PRP NNS PRP RB , PRP VB

s aici ; gran honor nos fai , so -us al

ue eser aisi ; grans honors nos faire , so w

PRO VJ ADV PONFP ADJ NCS PRO VJ PON PRO P

exmaralda

ne None None None joy None surprise None None

Figure 3: Resulting visualization for the query in Figure 2.

charts as a visualization methodology to the literary and linguistic studies. These charts are common in the socio-economic statistic field and are capable of visualizing in motion the development of a phenomenon in question across time. Hilpert (2011, 436) stresses that "the main purpose of producing linguistic motion charts is to give the analyst an intuitive understanding of complex linguistic development". Following Hilpert's methodology, we converted our data into the R data frame format and produced a motion chart, as shown in Figure 5. At present, our emotion analysis can be assessed as a dynamic motion chart, by using GoogleViz[6], internally interfaced via R (R Development Core Team, 2007). The chart allows for displaying emotion by type, color and size across time sequencing. Thus, the user can access this information in an interactive way without having to extract any information.For the future, we plan on adding discourse and word information.

8 Conclusion

We have presented an approach to providing access to an Old Occitan text via parallel word alignment with a modern language and cross-language transfer. We regard this project as a case study, showing that we can provide access to many types of information without the user having to learn cumbersome query languages or programming. We have also shown that the use of methods from computational linguistics, namely cross-language transfer, can pro-

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

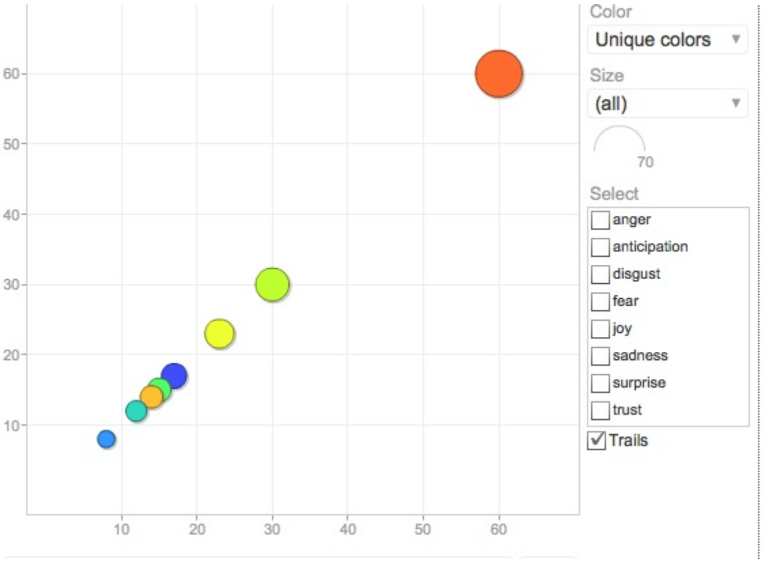


Figure 5: Dynamic emotion analysis with GoogleViz.

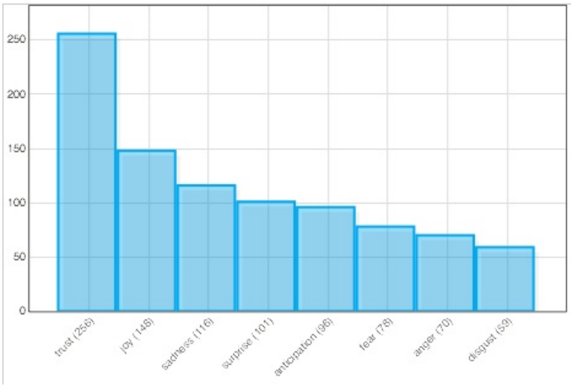


Figure 4: ANNIS frequency analysis.

vide tools for annotating the corpus without having access to an expert in the historical language. Using ANNIS, we showed how a multi-layered corpus can be queried via a user-friendly visual query interface. Finally, we presented a motion chart, which allows the user analyze and trace emotions dynamically without any technical requirements.

This work is part of our on-going project to fully annotate the Romance of Flamenca. Our goal is to provide users with necessary tools allowing for text-mining and visualization of this romance. Given a search query in ANNIS, we plan to develop an R package that will enable users to visualize their individual results exported from ANNIS and processed as motions charts and other statistical plots. Finally, this parallel corpus can be used as a training corpus in machine translation and for parallel dictionary and emotion lexicon building in resource-poor languages, such as Old Occitan.

References

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional sequencing and development in fairy tales. In Jianhua Tao, Tieniu Tan, and Rosalind Picard, editors, *Affective Computing and Intelligent Interaction*, volume 3784 of *Lecture Notes in Computer Science*, pages 668–674. Springer.

Kate Beeching. 2013. A parallel corpus approach to investigating semantic change. In Karin Aijmer and Bengt Altenberg, editors, *Advances in Corpus-Based Contrastive Linguistics: Studies in Honour of Stig Johansson*, pages 103–126. John Benjamins.

E.D. Blodgett. 1995. *The Romance of Flamenca*. Garland, New York.

Monserat Civit, Antònia Martí, and Nuria Bufí. 2006. Cat3LB and Cast3LB: From constituents to dependencies. In *Advances in Natural Language Processing*, pages 141–153. Springer.

Stefanie Dipper. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin, Germany.

Andrés Enrique-Arias. 2012. Parallel texts in diachronic investigations: Insights from a parallel corpus of Spanish medieval Bible translations. In *Exploring Ancient Languages through Corpora (EALC)*, Oslo, Norway.

Andrés Enrique-Arias. 2013. On the usefulness of using parallel texts in diachronic investigations: Insights from a parallel corpus of Spanish medieval Bible translations. In Paul Durrell, Martin Scheible, Silke Whitt, and Richard J. Bennett, editors, *New Methods in Historical Corpora*, pages 105–116. Narr.

Anna Feldman and Jirka Hana. 2010. *A Resource-Light Approach to Morpho-Syntactic Tagging*. Rodopi.

Suzanne Fleischmann. 1995. The non-lyric texts. In F.R.P. Akehurst and Judith M. Davis, editors, *A Handbook of the Troubadours*, pages 176–184. University of California Press.

Virginia Francisco, Raquel Hervás, Federico Peinado, and Pablo Gervás. 2011. Emotales: Creating a corpus of folk tales with emotional annotations. *Language Resources and Evaluation*, 46:341–381.

Martin Hilpert. 2011. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16(4):435–461.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Jungi Kim, Jin-Ji Li, and Jong-Hyeok Lee. 2010. Evaluating multilanguage-comparability of subjectivity analysis systems. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 595–603, Uppsala, Sweden.

Marijn Koolen, Frans Adriaans, Jaap Kamps, and Maarten de Rijke. 2006. A cross-language approach to historic document retrieval. In M. Lalmas and et al., editors, *Advances in Information Retrieval*, pages 407–419. Springer.

Sandra Kübler and Heike Zinsmeister. 2014. *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Academic.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

France Martineau, Paul Hirschbühler, Anthony Kroch, and Yves Charles Morin. 2010. Corpus MCVF (parsed corpus), modéliser le changement: les voies du français. Départment de Français, University of Ottawa.

Anthony McEnery and Zhonghu Xiao. 2007. Parallel and comparable corpora: What is happening? In G. James and G. Anderman, editors, *Incorporating Corpora: Translation and the Linguist*, Translating Europe, pages 18–31. Multilingual Matters.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 976–983, Prague, Czech Republic.

Saif Mohammad and Tony Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 70–79, Portland, OR.

Saif M. Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53:730–741.

Franco Moretti. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 1086–1090, Saarbrücken, Germany.

Daniela Oelke, Dimitrios Kokkinakis, and Mats Malm. 2012. Advanced visual analytics methods for literature analysis. *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 35–44.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Svetlana Petrova and Michael Solf. 2009. On the methods of information-structuralanalysis in historical texts: A case study on Old High German. In Roland Hinterhölzl and Svetlana Petrova, editors, *Information structure and language change. New approaches to word order variation in Germanic*, pages 121–160. Walter de Gruyter.

Shaver Philip, Schwartz Judith, Kirson Donald, and O'Connor Cary. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33.

R Development Core Team. 2007. *A language and environment for statistical computing*. R Foundation for Statistical Computing.

Luz Rello and Iustina Ilisei. 2009. Approach to the identification of Spanish zero pronouns. In *Student Research Workshop, RANLP*, pages 60–65, Borovets, Bulgaria.

Cristina Sánchez López, To Appear. *Sintaxis histórica de la lengua española. Tercera parte*, chapter Preposiciones, conjunciones y adverbios derivados de participios. Fondo de Cultura Económica, México.

Olga Scrivner and Sandra Kübler. 2012. Building an Old Occitan corpus via cross-language transfer. In *Proceedings of the First International Workshop on Lan-*

guage Technology for Historical Text(s), pages 392–400, Vienna, Austria.

Olga Scrivner, Sandra Kübler, Barbara Vance, and Eric Beuerlein. 2013. Le Roman de Flamenca : An annotated corpus of old occitan. In *Proceedings of the Third Workshop on Annotation of Corpora for Research in Humanities*, pages 85–96, Sofia, Bulgaria.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL: HTL*, Columbus, OH.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, Suntec, Singapore. To appear.

Christine Storm and Tom Storm. 1987. A taxonomic study of the vocabulary of emotions. *Journal of Personality and Social Psychology*, 53(4):805–816.

Dan Tufis. 2007. Exploiting aligned parallel corpora in multilingual studies and applications. In *Proceedings of the 1st International Conference on Intercultural Collaboration*, pages 103–117, Kyoto, Japan.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, PA.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*, pages 161–168, San Diego, CA.

Amir Zeldes, J. Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics*, Liverpool, UK.

Amir Zeldes. 2007. Machine translation between language stages: Extracting historical grammar from a parallel diachronic corpus of Polish. In *Proceedings of the Corpus Linguistics Conference (CL)*, Birmingham, UK.