*Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022* , pages 56–60
Marseille, June 25, 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC

# Setting Up Bilingual Comparable Corpora with Non-Contemporary Languages

**Helena Bermúdez Sabel, Francesca Dell'Oro, Cyrielle Montrichard, Corinne Rossari**
University of Neuchâtel
2000 Neuchatel (Switzerland)
{helena.bermudez, francesca.delloro, cyrielle.montrichard, corinne.rossari}@unine.ch

**Abstract**

This paper presents the project *Les corpora latins et français: une fabrique pour l'accès à la représentation des connaissances (Latin and French Corpora: a Factory for Accessing Knowledge Representation*) whose focus is the study of modality in both Latin and French by means of multi-genre, diachronic comparable corpora. The setting up of such corpora involves a number of conceptualisation challenges, in particular with regard to how to compare two asynchronous textual productions corresponding to different cultural frameworks. In this paper we outline the rationale of designing comparable corpora to explore our research questions and then focus on some of the issues that arise when comparing different diachronic spans of Latin and French. We also explain how these issues were dealt with, thus providing some grounds upon which other projects could build their methodology.

**Keywords:** comparability of language stages, genres, modality, annotation

## 1. Introduction

The project *Les corpora latins et français: une fabrique pour l'accès à la représentation des connaissances* (*Latin and French Corpora: a Factory for Accessing Knowledge Representation*), which started in February 2022, aims at comparing the use of modality in two languages which have a filiation relation: Latin and French. By the term 'modality' we refer to the linguistic expression of the stance of the speaker towards the propositional content of an utterance (Nuyts, 2005).

This project stems from the empirical observation of the variety of markers showing the speaker's stance in different languages (Narrog, 2016, Boye, 2016). Choosing two languages that are temporally distant, but genetically connected, enables us to pinpoint the continuity and the discontinuity in the usage of modal forms. We have thus to deal with textual productions belonging to different chronologies, which is not usually the case when speaking of comparable corpora[1] (for an exception, however, cf. van der Auwera and Diewald, 2012). To manage this issue, we needed to elaborate a dedicated methodology and set up corpora that could be compared as being representative samples of selected language stages. To this end, we took into account two different chronologic spans (according to traditional periodisations) for each of the two languages: Classical Latin (1st BCE to 3rd CE) and Early Mediaeval Latin (6th CE to 9th CE), and Classical French (1650-1799) and Modern French (1800-1979), respectively. Then, we set up the four corpora based on the selection of comparable genres in the two languages and in the four spans, though the notion of 'comparable genres' is a challenging one, when dealing with several asynchronous stages. Moreover, as one of our goals is also that of applying statistical methods to study the corpus, a major difficulty lies in comparing linguistic stages that involve each a different grammatical, orthographic and morphosyntactic evolution.

In this paper we outline our methodology to set up comparable corpora in Latin and French considering time period, genres and logistical means such as the availability of texts (especially in Latin). First, we describe the goals of our ongoing project and its aims, specifically considering its stumbling blocks. Then, we outline the choices we made to achieve the setting up of the corpora. Finally, we present how we devised to deal with different annotation tagsets and how those choices allow us to compare the two languages in a tool-based linguistic approach.

## 2. Studying and Comparing Modality Markers in Latin and French

### 2.1 Main Goals of the Project

The goal of the project is to identify the markers of modality, such as morphological or lexical devices and their uses, in the two languages, while taking into account the different genres (informative, ordinary writing, legal, among others) at different historical stages. After having collected the data, the obtained results will be compared in order to measure the similarities and differences in the use of the markers in terms of their presence/absence, their frequency (specificity score) and their association properties (co-occurrence specificity score). In order to do so, we plan to use textometric tools and specifically TXM[2] (Heiden, 2010). This is a platform that provides statistical tools (co-occurrence specificity score, specificity score, factorial correspondence analysis etc.), annotation tools (Heiden, 2018) and an easy access to the full text or to the view of keywords in context.

As mentioned above, the aim of the project is twofold: comparing modality in Latin and French and, at the same time, looking at the differences due to discourse genres in each language and between the two languages. The underlying methodology which mixes chronological spans and discourse genres, is particularly relevant for modality, whose values are instable, but it is efficient for any linguistic enquiry, since it allows one to better evaluate which linguistic elements are genre-dependent and which ones are specific to a particular period. The results of this

[1] See, for example, the catalogue of comparable corpora available at the Virtual Language Observatory (CLARIN 2021).

[2] Link to the website project: https://txm.gitpages.huma-num.fr/textometrie/en/

analysis could be easily extrapolated for analysing other Romance languages. These corpora will be made freely available to the scientific community under a Creative Commons license. Our corpora will facilitate the exploration of different research questions involving a contrastive perspective, and the semantic annotation can be exploited for studies that are adjacent to modality (e.g. enunciative responsibility).

### 2.2 Some Challenges of Building Asynchronous Comparable Corpora

The definitions given of comparable corpora in the literature and specifically in Corpus Linguistics (Sinclair, 1996; Habert et al., 1997, Talvensaari et al., 2007) are often vague and based on stressing the differences between comparable and parallel corpora.[3] Comparable corpora are thus defined as corpora built with texts in more than one language, with a purpose of comparison and with at least a common point represented by style and/or topics. However, some scholars also point out another required common point: the same time period (Kontonatsios, 2015: 38; McEnery, 2003: 450). Cf. the following list of relevant points for setting up comparable corpora:

> "the parameters that need to be controlled in order to compare languages include:
> – the time when the texts were written;
> – their discursive genre (descriptive, argumentative, etc.);
> – the type of audience targeted and their field (law, science, etc.)." (Zufferey, 2020: 83)

It is important to stress that this view is strictly dependent on a synchronic approach to text corpora. In fact, as shown by van der Auwera and Diewald (2012) comparable corpora can also consist of texts pertaining to distant diachronic spans.

Concerning the other criteria, it seems to us that the ones suggested by Zufferey (2020) are more precise than the notions of 'style' or 'topics' usually used. In fact, the latest may turn out to be problematic when selecting the relevant texts. For instance, a medical topic can be treated very differently according to the type of text and the period (written press, a filmed documentary, academic papers, scientific magazines, etc.). With regard to the audience in the past centuries, we cannot know it with precision. Therefore, this criterion is not applicable in the case of our corpora. Thus, genre and domain become the only suitable criteria for building our comparable corpora.

With reference to the setting up of our corpus, the following issues emerged:

(i) the difference in the time period inherent in our corpus: the two languages are not used simultaneously over time (at least not by native speakers);

(ii) genres are subject to variation over time and this complicates the possibility to compare works from different time spans.

However, we believe that it is possible to work around these two challenges in order to achieve our goal without disregarding them, and thus find a workable solution—maybe an imperfect one, but as Habert et al. put it (1997), working on imperfect data is the only way to contribute to corpus linguistics.[4]

We needed to devise a methodology for the selection of texts in order to master the intrinsic features of the data and the corpora. In the next section we outline such methodology and how we elaborated it.

## 3. Methodology for the Selection of Texts and Related Issues

### 3.1 Building a Corpus to Answer Our Research Questions

It is worth stressing that we adhere to the assertion by Hunston (2002) that a corpus is mainly a tool built in order to explore a research question. Many projects using comparable corpora focus on translation and terminology studies in order to create lexicons and translation resources when parallel corpora are not available (e.g. Delpech et al., 2012; Daille and Harastani, 2013) Our research is slightly different because it does not aim at studying how a modality marker is realised in both languages, but at observing the relations between the use of modality in a language and in one of its descendant languages. In particular, we want to assess which trends with regard to modality are due to diachrony and which ones are due to the genre. Both these questions are very important in the field of linguistics, in particular when analysing semantic change: for instance, it is relevant to take into account the notion of 'post-modality' in order to determine the diachronic evolution of the polysemy of modal markers (such as morphological markers or verbs, e.g. Latin *possum* and French *pouvoir* 'can').

### 3.2 Tackling Temporal Distance

As it is known, French and Latin coexisted during the Middle Ages, though Latin gradually ceased to be the mother tongue of any speaker. Our purpose is to isolate features concerning the use of modality in each language independently of the influence of one language on the other, but drawing on native or native-like speakers. Thus, contact influences between both linguistics systems generate interferences that go against the goals of the project as explained before. This is the reason why we decided to study diachronic spans for each language that do not overlap. In that way, we can take a look at the modal meaning conveyed by a marker in both languages at different time periods. Drawing on this, we will be able to

[3] See McEnery & Xiao (2007: 19 ff) for a discussion of terminological issues concerning parallel and comparable corpora and for a comprehensive definition of the latest term.

[4] "Les linguistiques de corpus se révèleront fructueuses comme domaine de recherche si l'on accepte l'imparfait, c'est-à-dire des ressources toujours « impures » […] ". (Habert et. al., 1997: in Chapter X, section 2.3). Our translation: "Corpus linguistics will prove to be a fruitful field of research if we accept the imperfect, that is always 'impure' resources".

create a cartography of modality markers in both languages and see what is relevant in a certain time period and what seems to be subject to variation over time.

This particularity of our research allows us to pinpoint diachronic and cultural differences that go beyond topic or style. Since genres have an impact on the way of saying as shown in Pincemin & Rastier (1999) and Adam (1997), they have more weight in our selection criteria than topic. Such a choice is particularly suitable for our research question, as we are interested in how events are modalised, i.e. how they are presented: the event itself being not relevant.

### 3.3 Dealing with the Audience Criteria and Genre Variation

The parameters of genre should be considered simultaneously to the one of audience target because they are strictly interrelated. In fact, it is really complicated to dissociate, e.g. the genre 'academic paper' from the target audience of the genre.

For our work, we face a double constraint, i.e. (i) finding the 'same' genres attested over the centuries and (ii) finding inside those genres domains that can be compared. For instance, the genre of treaty is attested over time, but the subjects did evolve. Therefore, it is nowadays rare to encounter treaties about mystic topics and conversely to find treaties about communication media in Antiquity.

Moreover, it seems that genres, topics and audience show a great variation which could be related to the digital revolution. This has been documented, among other, by Paveau (2013). She proposes the term 'technogenre' and the following description:

> Ces technogenres sont des aménagements de genres préexistants (en twitterature en particulier) ou des inventions de l'écosystème numérique (Paveau, 2013: 24).
>
> These technogenres are adaptations of pre-existing genres (in twitterature in particular) or inventions of the digital ecosystem (our translation).

Among the variation and the creation of new genres, the 'digital ecosystem' led to the slow mutation of canonical genre such as the genre of correspondence which today could include emails or chats. Moreover, it is by far more difficult to delimit the target audience when the text is intended for the World Wide Web. This was for us the main reason for excluding the 20th and 21st centuries, thus excluding the modern stage of French language.

Second, as we considered it important to take into account the genre variation within a language through centuries, we decided to sample the Latin corpus and the French corpus at different time periods. The result gives us an original comparable corpus with multiple variables.

We propose to summarise what said above in the following schema (see Figure 1). The image shows a timeline in centuries, in which the selection of texts by time period, and genre is represented for each language (coded by different colours).

Figure 1 shows the macro-categories relevant for studying and comparing Latin and French: we separate technical treatises from literary genres and we keep a third category (Other) to include other function-specific genres such as correspondence or legal texts. Each one of these categories is further divided in sub-categories. For instance, technical treatises are grouped by domain: rhetoric and linguistics; philosophy; natural sciences. As an example, in the sub-category 'rhetoric and linguistics' we consider that the Latin works *De verborum significatione fragmentum* by Sextus Pompeius Festus (2nd CE) and *Ars grammatica* by Alcuinus (8th CE) are comparable to the *Grammaire universelle* by Court de Gébelin (1774) and the *Essai de sémantique: science des significations* by Michel Bréal (1887). Each sub-category is between 300'000 and 800'000 words long depending on texts availability (obviously, for Latin we have certain limitations concerning the number of works preserved for certain domains and their availability as free resources).

Figure 1 shows the different variables contained in our comparable corpora that will be exploited to investigate modality: it allows us to compare languages, genres, diachronic spans independently or in combination.

## 4. The Annotation Tagset

### 4.1 Automatic Lemmatisation and Part-of-Speech Tagging

For reasons of feasibility, we decided to carry out an automatic linguistic annotation of the corpora. As figure 1 shows, we retrieved texts in both languages from different chronological stages. One of the issues that arise from this is tied to the graphical representation of data. For instance, in Classical French verbs do not present the same endings. For example, the various forms of *devoir* 'must, have to' in Classical French do not have the same graphical representation as in Modern French, when the verb is conjugated. Similarly, Early Mediaeval Latin can display more recent variants with respect to Classical Latin. In order for us to avoid working based on graphic forms, which are very likely to change over time, we need to annotate our corpora and work with units that are less likely to change, i.e. lemma and morphosyntactic categories.

In order to obtain the best performance with regard to the automatic annotation, we are not only implementing language-specific annotation models, but also period specific models. We selected the following three morphosyntactic taggers:

- Treetagger and the annotation dataset for contemporary French
- Presto, an annotation dataset for Classical French designed during the implementation of the PRESTO project (Blumenthal and Vigier, 2018)
- Treetagger with the model trained by Gabriele Brandolini for Latin.

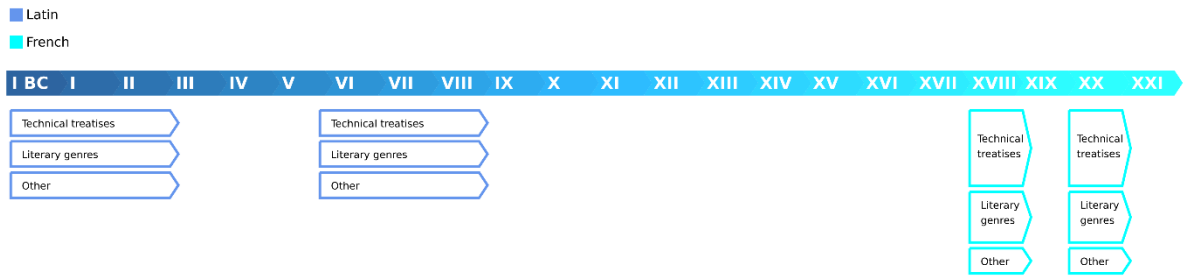


Figure 1. Graphic representation of the linguistic variables present in the project (language, genre, period)

For the various stages of French, there are differences not only at the level of the tagset, but also in the achieved precision. In Classical French we have pieces of information with tags about the subject of the verb according to its conjugation which is not available for Contemporary French. In order to solve this problem, we decided to keep the simplest tagging available, since it would be too time consuming to add a large number of tags.

### 4.2 Semi-automatic Semantic Annotation

As the main goal of our project is the study of modality, we devised an annotation tagset for the manual semantic annotation of modal markers that is appropriate for both languages and for the four linguistic stages of our corpora. We distinguish two major categories[5] (presented in Table 1 as well) of modality—epistemic and non-epistemic—with different sub-categories for each major category:

- epistemic: from weak degree of certainty (*Someone knocks on the door, this may well be the neighbour*) to strong degree of certainty (*Someone knocks on the door, this must be the neighbour)*
- non-epistemic: e.g.
  - capacity (*I can sing very well)*
  - generic possibility[6] (*The tennis court is free, we can go play*)
  - permission/obligation (*You must/may go now*)
  - volition (*I want to go to the movies*)

We devised two possible annotation procedures. As shown in Table 1, a marker which always conveys the same type of modality—e.g. French *peut-être* or Latin *forsitan* 'maybe' which express medium epistemic modality—allows a semi-automatic annotation within the TXM platform (making it possible to annotate at once every occurrence of a lemma). In the case of polyfunctional markers, such as French *pouvoir* and Latin *possum* 'can, to be able' which can express different types of modality—e.g. someone's ability to do something or an epistemic stance— we sample each corpus in order to manually annotate every occurrence of the term according to the type of modality it carries.

| Major modality type | Examples of modal markers that can be semi-automatically annotated | Modal markers that required a manual annotation (meaning is context-dependant) |
|---|---|---|
| epistemic | FR: *certainement / probablement*<br>LA*: forte* | FR: *pouvoir*<br>LA: *possum*<br>Both : morphological markers such as subjunctive/conditional affixes |
| non-epistemic | FR: *vouloir, obligatoirement, nécessairement*<br>*LA: volo* | FR: *pouvoir/ devoir/ falloir*<br>LA: *possum / debeo*<br>FR : / *falloir*<br>LA: *licet*<br>Both: morphological markers such as future affixes |

Table 1. Example of the annotation of some modal markers by type of modality

## 5. Conclusions

In order to achieve our goals and answer our research questions, we had to set up a methodology of selection and processing of texts for both Latin and French to assure the comparability of the corpora.

Our project is still at an early stage of its implementation. The corpora are not set up yet, but a methodology tackling the main challenges and tailored to our research goals has been defined.

This paper shows the different steps in elaborating our methodology concerning the selection and processing of texts. Its interest lays on the lack of documented endeavours working with diachronic comparable corpora.

## 6. Acknowledgements

This work is supported by the Empiris Foundation (fund *Jakob Wüest*).

[5] The definition of the main categories of modality is a debated subject. Our categorization is based on the distinction between epistemic modality and non-epistemic modality which is the most agreed upon.

[6] We distinguish generic possibility from epistemic modality: the latter corresponds to propositional modality and the former to event modality (Palmer, 2001).

## 7. Bibliographical References


Adam, J.-M. (1997). Genres, textes, discours : pour une reconception linguistique du concept de genre. In *Revue belge de philologie et d'histoire*, 75-3. pp. 665-681

Auwera van der, J. and Diewald, G. (2012). Methods for Modalities. In A. Ender et al. (Eds). *Methods in Contemporary Linguistics*. De Gruyter, pp. 121–142.

Blumenthal, P. and Vigier, D. (2018). Présentation. In P. Blumenthal & D. Vigier (Eds.). *Études diachroniques du français et perspectives sociétales*. Peter Lang, pp.7-20.

Boye, K. (2016). The Expression of Epistemic Modality. In J. Nuyts & J. van der Auwera (Eds.). *The Oxford Handbook of Modality and Mood*. Oxford University Press, Oxford, pp. 117–140.

Daille, B., and Harastani, R. (2013). TTC TermSuite - Terminological Alignment from Comparable Corpora (TTC TermSuite Alignement Terminologique à Partir de Corpus Comparables) [in French]. In *Proceedings of TALN 2013 (Volume 3 System Demonstrations)*, pages 812–813. Les Sables d'Olonne, France: ATALA.

Delpech, E., Daille, B., Morin, E. and Lemaire, C. (2012). Identification of Fertile Translations in Comparable Corpora: A Morpho-Compositional Approach. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, [online], San Diego, California, USA: Association for Machine Translation in the Americas. https://aclanthology.org/2012.amta-papers.5.

Habert, B., Nazarenko, A. and Salem, A. (1997). Les linguistiques de corpus. Armand Colin, Paris. [online]. http://lexicometrica.univ-paris3.fr/livre/les_linguistiques_de_corpus_1997/

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation,* [online]. Sendai, Japan. http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24_sheiden.pdf

Heiden, S. (2018). Annotation-based Digital Text Corpora Analysis within the TXM Platform. In S. Bolasco, et al. (dirs).. *Proceedings of the 14th JADT'18,* Roma, UniversItalia, pp. 367-374.

Hunston, S. (2002). Corpora in Applied Linguistic. Cambridge: Cambridge University Press.

Kontonatsios, G. (2015). Automatic Compilation of Bilingual Terminologies from Comparable Corpora. Manchester, UK, The University of Manchester, [Thesis].

McEnery, A. (2003). Corpus Linguistics. In R. Mitkov (Ed.) *Oxford Handbook of Computational Linguistic*s. Oxford: Oxford University Press, pp. 448-63.

McEnery, T. and Xiao, R. (2007). Parallel and Comparable Corpora: What is Happening?. In Gunilla A. & M. Rogers (Eds.), *Incorporating Corpora: The Linguist and the Translator*. Bristol, Blue Ridge, pp. 18-31. https://doi.org/10.21832/9781853599873-005

Narrog, H. (2016). The Expression of Non-epistemic Modal Categories. In J. Nuyts & J. van der Auwera (Eds.), *The Oxford Handbook of Modality and Mood*. Oxford University Press, Oxford, pp. 89–116.

Nuyts, J. (2005). The Modal Confusion: On Terminology and the Concepts behind it. In A. Klinge et al. (Eds), *Modality. Studies in Form and Function*, Equinox Publishing, pp. 5–38.

Nuyts. J. (2016). Analyses of the Modal Meanings. In J. Nuyts & J. van der Auwera (Eds.), *The Oxford Handbook of Modality and Mood*. Oxford University Press, Oxford, pp. 31–49.

Palmer, F. R. (2001). *Mood and Modality* (2nd ed.). Cambridge University Press.

Paveau, M.-A. (2013). Genre de discours et technologie discursive. In *Pratiques* 157-157, pp. 7-30. DOI: https://doi.org/10.4000/pratiques.3533

Pincemin, B. & Rastier, F. (1999). Des genres à l'intertexte. In *Cahiers de praxématique* 33, pp. 83-111. DOI: https://doi.org/10.4000/praxematique.1974

Sinclair, J. (1996). Preliminary Recommendations on Corpus Typology. In *Rap. tech.*, EAGLES (*Expert Advisory Group on Language Engineering Standards*), CEE.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M. and Keskustalo, H. (2007). Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. ACM (Association for Computing Machinery) Trans. Inf. Syst. [online]. https://doi.org/10.1145/1198296.1198300

Zufferey, S. (2020). Introduction to Corpus Linguistics. John Wiley & Sons.


## 8. Language Resource References


CLARIN (2021). Virtual Language Observatory. https://vlo.clarin.eu