Building diachronical reference corpora for the French language

Alexei Lavrentiev

А.М. Лаврентьев
A. Lavrentiev

СОЗДАНИЕ РЕФЕРЕНТНЫХ ДИАХРОНИЧЕСКИХ КОРПУСОВ ДЛЯ ФРАНЦУЗСКОГО ЯЗЫКА

BUILDING DIACHRONICAL REFERENCE CORPORA FOR THE FRENCH LANGUAGE

Аннотация. В докладе представлены проблемы создания референтных диахронических корпусов французского языка пути их решения, предложенные несколькими взаимосвязанными проектами: Репрезентативный корпус первых французских текстов (CoRPTeF), Большая историческая грамматика французского языка (GGHF) и Эволюция системы предлогов во французском языке (PRESTO). Многие из этих проблем релевантны для проекта референтного корпуса любого языка, охватывающего широкую диахроническую перспективу.

Abstract. This paper deals with the problems in creating diachronical reference corpora of the French language and with their solutions proposed by several interconnected projects: the Representative Corpus of the First French Texts (CoRPTeF), the Big Historical Grammar of French (GGHF) and the Evolution of the French Prepositional System (PRESTO). Many of these problems are relevant for a project of a reference corpus of any language including a large diachronical dimension.

1. Introduction

Unlike British or American English, or, more recently, Russian, a national reference corpus has never been created for the French language. Although the amount of digitized texts of all kinds available online and the development of methods and tools for using the «web as a corpus» make «old-style» national corpora look ridiculously small and obsolete, reference corpora are still useful, as they are normally compiled from carefully selected and prepared textual data with rich metadata and various kinds of annotation.

Reference corpora are especially important for diachronical studies, as texts are the only source of data for the past states of a language. In fact, the evolution of the frequency of a linguistic phenomenon is often the main indicator of a change in language, and a reliable reference corpus is the best tool to observe such evolution.

As far as the history of the French language is concerned, a number of corpora and text collections were compiled since 1980s for various periods[1], but until recently there has been no corpus covering the entire time span from the 9th till the 21st century. This situation is about to change due, on the one hand, to a number of research projects with a large diachronical outlook (such as GGHF[2] and PRESTO[3]) and, on the other hand, due to the creation of French research infrastructures (such as Corpus Écrits[4] or CAHIER[5]) which promote integration and interoperability of resources created in different contexts. An initiative to discuss methodological and practical issues related to the creation of the French reference corpus was taken in 2012 by the ILF research federation[6] and several sessions of presentations and round tables have already been held.

---

[1] Guillot C., Heiden S., Lavrentiev A., Marchello-Nizia C. Constitution et exploitation des corpus d'ancien et de moyen français // Corpus. 2008. № 7. C. 5 – 23.

[2] La Grande Grammaire Historique du Français (the Big Historical Grammar of French), project coordinated by C. Marchello-Nizia, B. Combettes, S. Prévost, and T. Scheer.

[3] L'évolution du système prépositionnel du français: approche statistique et diachronique (Evolution of the French prepositional system: a statistical and diachronical approach), project coordinated by P. Blumenthal and D. Vigier and co-funded by the French ANR and the German DFG granting agencies (2013-2016).

[4] **http://corpusecrits.corpus-ir.fr**

[5] **http://www.cahier.paris-sorbonne.fr**

[6] Institut de linguistique française (Institute for the French linguistics). **http://www.ilf.cnrs.fr**

In this paper we discuss some of the most common problems (both methodological and practical) that the aforementioned projects and initiatives have to face, as well as the solutions proposed.

The main concern for any corpus is its representativeness. Whereas no corpus can be representative of a language as a whole, a reference corpus should, to quote John Sinclair, «be large enough to represent all the relevant varieties of the language»[1]. Therefore, the first step in designing a reference corpus is to define the relevant variables of text typology.

2. Text typology variables

1.1. Date

The time variable is of primary importance for diachronical studies. Most of the texts created after 1500 can be dated with a precision of at least a year. For earlier periods, the accuracy of the dating of a text can be as rough as a few decades or even a whole century. In addition, the "author's original" is almost never available, and we have to deal with manuscript copies made many ears later and bearing more or less important changes introduced by scribes. Ideally, a text in diachronical corpus should be "equipped" with as much dating information as possible to let the user choose the most appropriate option.

1.2. Region or dialect

Diatopical variation is extensive in medieval texts. However, it may be difficult to assign a regional attribution to a given text, as it may contain forms from different dialects due the long history of

[1] Sinclair J. Preliminary recommendations on corpus typology. Technical report EAGLES (Expert Advisory Group on Language Engineering Standards). 1996.
http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

copying and to a number of other factors. Information on the regional features of the text should be provided in the corpus wherever possible. Otherwise a generic «undefined» value is assigned to a text. After the codification of the literary language in the 16th and 17th centuries, regional features become rare in literary texts, and the value «standard» can be assigned to the regional variable.

1.3. Form, domain and genre

To quote David Lee, «most corpus-based studies rely implicitly or explicitly on the notion of genre or the related concepts register, text type, domain, style, sublanguage, message form, and so forth. There is much confusion surrounding these terms and their usage[1]». The situation has not really changed since 2001, and the terms and taxonomies used in different corpora vary considerably. In addition, the system of text types (or whatever term is used) evolves with the time, and it is complicated to use a single grid over the centuries. A reference corpus should at least adopt a clear system of text description in this area and try to ensure as much compatibility as possible with the systems used in comparable projects. In the projects of French diachronical corpora we are involved in, it has been decided to use the three terms mentioned in the title of this section.

The «form» variable is generally the easiest to determine (verse or prose), although some texts present a mixture of both in various proportions.

The «domain» of the text is related to its primary function (e.g. «entertain» for literary texts or «regulate the social life» for juridical texts). For medieval French, six major text domains have been

[1] *Lee D.* Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle // Language Learning & Technology. 2001. Vol. 5. P. 37-72.

identified and described by the CoRPTeF project[1]: literary, didactical-scientific; religious, historical, juridical and practical[2]. New domains, such as political texts, news or private correspondence appear later. A given text can belong to multiple domains (e.g. literary and religious).

The «genre» is defined internal by structural properties of a text. It is of course related to the literary genre but also applies to non literary texts. A «genre-like» term is often present in the traditional title of a text (e.g. «Chanson de Roland» or «Roman de la rose»), but the same term can change its meaning with the time (e.g. «chanson de geste» corresponds to «epic» in the modern system and has nothing to do with lyrical songs in later periods). The system of genres is probably the least stable over the large diachrony, it is therefore impossible to define a unique list of values for all the periods, however, it is important to ensure that the same term is not used for different genres in different parts of the corpus and that similar genres are identified by the same term across the corpus.

3. Text selection constraints

An «ideal» reference corpus should be perfectly balanced with respect to the selected variables, however a number of constraints make this task extremely complex.

First of all, the definition of text unit needs to be clarified. It can be generally agreed that a text should have a single author (or group of authors) and be composed at a particular moment or relatively short period of time. The question arises whether a collection of short poems by the same author may be considered the same text. This

[1] Corpus Représentatif des Premiers Textes Français (Reprensentative corpus of the first French texts), project coordinated by Céline Guillot and funded by the French ANR agency (2008-2011). **http://corptef.ens-lyon.fr**

[2] *Guillot C., Lavrentiev A.* Présentation des descripteurs du projet CORPTEF. Lyon, 2009. **http://corptef.ens-lyon.fr/IMG/pdf/descripteurs-corptef.pdf**

option was chosen in the projects of French diachronical corpora but it should be noted that the corpus query tool allows the user to isolate each poem and work on them individually, if necessary.

The size of texts is extremely variable (from a few hundreds to hundreds of thousands of words), which raises the question of sampling. Sampling may be inevitable if the corpus project implies manual annotation and exhaustive qualitative analysis. An interesting solution was used in the GGHF project which has built two corpora: the «core» where the text size is limited to 40 000 words and the «integral» corpus composed of entire texts of the core corpus and of a number of additional texts.

For the early stages of the history of the French language, the construction of a balanced corpus is impossible, as very few texts are available, and some text types are over-represented (like the anglo-norman dialect, the verse form and the religious domain in the 12th century).

Finally, some texts are subject to copyright restrictions. This regards not only the texts of the 20th century the authors of which are alive or deceased since less than 70 years[1], but also the critical editions of texts from the public domain. The copyright status of the critical editions is actually a complex juridical problem[2], and it is possible that many of the restrictions currently applicable to French editions may be removed in the future. In any case, texts without copyright restrictions or with open licenses (e.g. Creative Commons[3]) should be preferred in the reference corpora.

---

[1] This is the legal term of copyright protection according to French laws. It may be extended in some cases.

[2] *Margoni T., Perry M.* Scientific and critical editions of public domain works: an example of European copyright law (dis)harmonization. *Canadian Intellectual Property Review*. 2011. Vol. 27. P. 157–170. Available at SSRN: **http://ssrn.com/abstract=1961535**

[3] **http://creativecommons.org**

4. Corpus projects

4.1. CoRPTeF

This project aimed at the creation of a representative corpus of texts of the earliest period in the history of the French language (from 9th till the end of the 12th century). Although its diachronical dimension is relatively narrow, as there are very few extant texts composed before the 12th century, the CoRPTeF project provided an opportunity to refine the methodology of text description and selection. The experience of the project showed that it is impossible to consider the early stages of the French language independently of its complex relations to the Medieval Latin. In the continuation of CoRPTeF, a project of bilingual Latin-French medieval corpus was initiated in 2012.

4.2. GGHF

This project aims at the publication of a new corpus-based descriptive historical grammar of the French language (from the 9th till the 21st century). Different parts of the grammar will be written by experts in the corresponding periods and linguistics disciplines, and the use of the corpus is a general requirement to all the contributors. As already mentioned, this project introduced the concept of the «core» and the «additional» corpora. It had to face the problem of changes in text types over the centuries, to elaborate a sampling strategy and to deal various formats of source texts and with copyright issues. The corpus is currently compiled but its usage is limited to the members of the project due to the copyright problems that will be hopefully resolved in the nearest future.

4.3. PRESTO

This joint French and German project has started in April 2013. Its research program focuses on the development of the French prepositional system all over the history of this language. An

important part of the project is devoted to the development of methods and tools of automated distributional analysis of the prepositions, and relies on the construction of a representative lemmatized corpus. PRESTO extends the «core» GGHF corpus to make it more balanced and representative and develops tools for automatic lemmatization of medieval and early modern texts.

5. Conclusion

The list of the methodological problems and of the corpora projects mentioned above is not exhaustive, some others (including MCVF[1], ELICO[2], SRCMF[3], etc.) should have been mentioned if we had more time and place. Nevertheless we hope that the overview we make gives an adequate idea of the corpus-related initiatives and research problems in the field of the French diachronical linguistics. Thanks to the contribution of different projects and to the development of corpus analysis tools and infrastructures, it is possible that in some years the scholarly community will have at its disposal a reliable and freely accessible reference corpus for the history of the French language.

[1] **http://www.voies.uottawa.ca**
[2] **http://elico.linguist.univ-paris-diderot.fr**
[3] **http://srcmf.org**