French through Corpora

French through Corpora: Ecological and Data-Driven Perspectives in French Language Studies

Edited by

Henry Tyne, Virginie André,
Christophe Benzitoun, Alex Boulton
and Yan Greub

CAMBRIDGE
SCHOLARS
PUBLISHING

French through Corpora:
Ecological and Data-Driven Perspectives in French Language Studies,
Edited by Henry Tyne, Virginie André, Christophe Benzitoun,
Alex Boulton and Yan Greub

This book first published 2014

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-5831-5, ISBN (13): 978-1-4438-5831-1

TABLE OF CONTENTS

PART 1. DIACHRONIC STUDIES

PART 2. CORPUS AND SYNTAX

PART 3. SOCIOLINGUISTIC STUDIES

PART 4. LEARNING AND TEACHING FRENCH

FOREWORD

This volume presents a range of ongoing research on French that is informed by corpus linguistics. It is of particular interest because there is a dearth of publications that focus entirely on French while being accessible to non-French speaking linguists. Corpora offer the best way of accessing ‘authentic’ language, given certain methodological caveats which are implicit in the term ‘ecological’, used here to refer to the collection and analysis of ‘natural’, ‘undoctored’ language that is respectful of the context in which the data is gathered.

Throughout this volume interesting light is shone on the status of corpus linguistics itself. While the latter can be seen as a branch of linguistic research with its own range of theoretical questions, it can only be addressed with reference to specific issues in those areas of linguistics for which it acts as a research tool, and as an empowering method of collecting and analysing data on a scale not previously imaginable. As well as providing scale, the use of corpora allows for an analysis that is both more complex (e.g. combining databases or connecting parameters to reveal previously neglected tendencies), and more refined (distinguishing between hitherto indistinguishable sub-systems). Thus certain dichotomies that operated in the past can be questioned: diachronic and synchronic descriptions are less separate than they were held to be post-Saussure, for example. This facilitating, or bridge-building, aspect of corpora is illustrated in exemplary fashion here by the way in which the different contributions, although grouped into Diachronic Studies, Corpus and Syntax, Sociolinguistic Studies, and Learning and Teaching French, often address a similar range of issues.

Some issues nevertheless arise that are specific to a particular area. In the case of the use corpora in diachronic research, for example, there is the difficulty of avoiding categories that may be anachronistic. From the section on learning and teaching French, it emerges that different definitions of ‘authenticity’ of language may be appropriate in applied, as opposed to descriptive, linguistics. A major preoccupation for both descriptive and applied linguists working on (metropolitan) French has been the divide between the spoken language and the written, and the lack of adequate descriptions of the former. The pioneering *Groupe aixois de recherches en syntaxe* collected a (relatively limited) spoken corpus, mainly for the purpose of analysing the grammar of spoken French. While this work has continued to be built on, more recently the focus of attention of many researchers has been on the pragmatic and interactional aspects of French. Indeed, there is perhaps not a section entitled Pragmatics because pragmatic issues permeate many of the contributions.

Despite the fact that several authors point out that there is still no large-scale corpus of metropolitan French, a variety of smaller-scale corpora are used to great effect in the diverse research projects presented here. This volume therefore gives us a substantial body of corpus-informed research that demonstrates the importance of the work being carried out by linguists in France, Britain and elsewhere on the French language.

Carol Sanders
Emeritus Professor
University of Surrey, UK

INTRODUCTION

This book looks at the French language through corpora, and comprises four parts dealing respectively with diachrony, syntax, sociolinguistics and issues arising in the learning and teaching of French. Each part is headed by a chapter that provides an overview of the given field in relation to the themes running through the book. Other than contributing to our general understanding of the French language today, this book specifically addresses the use of *corpora* for the study of language and the links between tools, methods and analyses. How do we use corpora? What are the underlying theoretical and/or methodological considerations? How have these changed our way of formulating linguistic descriptions? What are the implications for descriptive accounts of French? What are the applications of corpus studies?

Each chapter focuses on specific aspects of French and addresses (often indirectly) issues such as the ways in which researchers use existing resources, the reasons for producing new resources, or questions arising from different types of data use. One aim is to challenge or complete existing work, not least in relation to the possibilities that are now made available through corpus use.

Corpora provide *data*, and a common theme throughout the book is one of empiricism. A distinction is sometimes drawn between corpus-based and corpus-driven approaches, and this is apparent here: a corpus-based study looks to gather findings that test a certain idea or model, whatever its scope or ambition; a corpus-driven study is more ecological insofar as it aims to build conclusions on the sole basis of the findings (Tognini-Bonelli 2001). These are perhaps two ends of a continuum, and in many cases there will be a continuous interaction between data consultation and the questions one has, each influencing the other at all stages, from corpus compilation through to final outcomes. But in all cases it is important not to lose sight of the fact that language is not just a neutral collection of data, hence the inclusion of the concept *ecology* in the title.

Ecology has become something of a buzz-word of late in many spheres. The rationale for the ecological leitmotif here derives from a shared interest in language and environment: ecology is to do with understanding language not as an abstract system but as an integral part of human existence (Haugen 1972). Ecology also refers to the authentic nature of the data and issues to do with their collection, transcription and editing, mark-up and tagging, or any other treatment resulting from human intervention whether deliberate or not.

Corpora are used nowadays in many areas of linguistics (and beyond), often for purposes not anticipated by the original designers. Burnard (2002: 67), for example, readily admits that the linguists behind the landmark BRITISH NATIONAL CORPUS (BNC) never imagined that it would be bought by individuals, nor that it could be of any direct use for teachers and learners. Nonetheless, where large corpora are now available freely on line, teachers and learners tend to make up the majority of users, as is the case for the BYU or Lextutor interfaces to several large corpora in different languages. In some cases, corpora have been used as simple repositories of linguistic 'facts', almost as archives or databases to dip into when needed and in whatever way required. This approach can be useful in philology, for example, where collections of available texts are often limited and difficult to gather together into a single coherent corpus. Such activity may be considered peripheral to corpus linguistics as an identifiable discipline – or even by some as not true corpus linguistics at all. It does, however, have the virtue of quietly filtering corpus linguistics into 'the rest of linguistics', thereby making it more accessible to a wide research community and extending the influence of its methods and results among the general field. This has also given rise to a debate about whether corpus linguistics represents 'merely' a methodology precisely because it is open to so many uses in so many fields (McEnery *et al.* 2006: 8), or something rather more (Tognini-Bonelli 2001), given its tremendous impact on all that we know about language use – not for nothing do McCarthy (2008) and others talk of a *corpus revolution*.

Turning specifically to French, Gadet (2009: 115) points out that while "for many contemporary linguists, 'doing linguistics' means working with corpora […] the study of French distinguishes itself in this respect from that of other languages". She gives two reasons for the comparative underuse of corpora in French studies: the fact that sizeable corpora for French have been late in coming in comparison to other languages (in Europe, at least); and the non-centrality of many existing French corpora (it is of note that there is no French national corpus). Gadet does concede, though, that the "broadening of the data" over the last few decades has brought about a number of interesting changes, as different research methodologies and theoretical backgrounds find common ground. This is particularly true perhaps for the study of spoken grammar (Blanche-Benveniste 2007: 129), an area where French corpus work has proved particularly innovative through the work of the GARS (*Groupe aixois de* *recherches en syntaxe*) school of linguists. Blanche-Benveniste notes that this is beneficial for grammatical description of the language as a whole, both in terms of suggesting new methods and for discovering new phenomena.

Though we might optimistically imagine the future will hold greater collaboration between sectors of linguistics, positions can become entrenched and exchanges rather heated, as in the "bootcamp debate" (see Worlock Pope 2010) between two camps of corpus linguists – those who are keen to see interaction with cognitive linguistics, and those who are sceptical of any compromise on what may be perceived as an epistemological or even ethical divide between empiricism and intuition/introspection. Corpus linguists may be understandably wary, having been branded as "butterfly collectors" (Chomsky 1979) lacking theory – corpus linguistics as a field simply "doesn't exist" claims Chomsky in an interview reported in Aarts (2001: 5). On the other hand, it is perhaps worth remembering that linguistics has been through a number of pendulum swings over the last century or so (for reasons of technical and methodological possibilities and indeed for cultural or political reasons as much as for theoretical reasons), and that it may be more profitable to seek common ground and collaboration between different approaches (e.g. Fillmore 1992). In other words, should butterfly collectors hunt in a protected area or should they open up to different areas and points of view? Should linguists who do not go in for butterfly collecting become interested in the findings of their collector colleagues, and vice versa? Answers to these questions (any many more) represent a major challenge for the ecology of the linguistics world. As Labov (1971: 119) pointed out several decades ago now, "it is not necessary for everyone to use the same methods – indeed it is far better if we do not […] Data from a variety of distinct sources and methods, properly interpreted, can be used to converge on right answers to hard questions". Labov also draws our attention to the "cumulative principle" of linguistic research, whereby "the more that is known about language, the more we can find out about it" (p. 98). And so it is that this book offers a series of studies that come together in their concern for furthering our understanding of the French language, of its uses, its forms, its variation, its acquisition, etc., as well as in promoting ecological approaches to using corpora for studying these questions.

Bibliography

Aarts, B. 2001, “Corpus linguistics, Chomsky and fuzzy tree fragments”, In C. Mair & M. Hundt (eds.), *Corpus Linguistics and Linguistic Theory*, Amsterdam: Rodopi, 5-13

Blanche-Benveniste, C. 2007, “Corpus de langue parlée et description grammaticale de la langue”, *Langage et société* 121-122(3), 129-141

Burnard, L. 2002, “Where did we go wrong? A retrospective look at the British National Corpus”, In B. Kettemann & G. Marko (eds.), *Teaching and Learning by Doing Corpus Analysis*, Amsterdam: Rodopi, 51-70

Chomsky, N. 1979, *Language and Responsibility* (based on conversations with Mitsou Ronat), New York, Pantheon

Fillmore, C. 1992, “Corpus linguistics or computer-aided armchair linguistics”, In J. Svartik (ed.), *Directions in Corpus Linguistics*, Berlin: Mouton de Gruyter, 35-60

Gadet, F. 2009 “Stylistic and syntactic variation: introduction”, In K. Beeching, N. Armstrong & F. Gadet (eds.), *Sociolinguistic Variation in Contemporary French*, Amsterdam: John Benjamins, 115-120

Haugen, E. 1972, *The Ecology of Language: Essays by Einar Haugen* (ed. A. S. Dil), California: Stanford University Press

Labov, W. 1971, “Some principles of linguistic methodology”, *Language in Society* 1, 97-120

McCarthy, M. 2008, “Accessing and interpreting corpus information in the teacher education context”, *Language Teaching* 41(4), 563-574

McEnery, T., R. Xiao & Y. Tono 2006, *Corpus-Based Language Studies: An Advanced Resource Book*, London: Routledge

Tognini-Bonelli, E. 2001, *Corpus Linguistics at Work*, Amsterdam: John Benjamins

Worlock Pope, C. (ed.) 2010, “The bootcamp discourse and beyond”, *International Journal of Corpus Linguistics* 15(3)

PART 1.

DIACHRONIC STUDIES

CHAPTER ONE

DIACHRONIC LINGUISTICS AND ELECTRONIC CORPORA

BERNARD COMBETTES
UNIVERSITY OF LORRAINE AND ATILF-CNRS RESEARCH UNIT, FRANCE

Introduction

Before looking at the impact of recent developments in corpus enquiry on the field of historical linguistics, it is worth bearing in mind that many of the issues here also concern synchronic studies. The best known and most widely discussed of these is the question of representativity of data (Biber *et al*. 1998; Habert *et al*. 1997). However, certain practical considerations can make the diachronic study of language particularly problematic. Although it is nearly always possible to make general statements from corpus findings, when it comes to looking at earlier periods, obstacles are soon encountered due, quite simply, to the limited range of available texts (cf. chapters 2-4, this volume). For example, Old French from the 11th and 12th centuries is known essentially through literary (mainly poetic) works. Non-literary texts (a small number of chronicles and some administrative documents) comprise a very small part of the total for the researcher to draw on. The paucity of these sources is not helped, of course, by the very status of written documents at the time, together with the use of Latin in the drafting of certain texts, not to mention the absence of spoken records (see Ingham, this volume). And it is difficult, if not impossible, to alter this established fact. The situation is, however, quite different when it comes to studying more recent periods where it is relatively easy to obtain different corpora, not to mention the possibility of taking spoken or 'non-conventional' French into account, with written traces available from the beginning of the Pre-Classical period.

When dealing with the use of corpus data in historical linguistics, it is also useful to distinguish between what is specific to the study of change or development (whatever the particular periods concerned) and what comes under the synchronic analysis of a past language state. The term 'historical linguistics' is somewhat ambiguous since it covers both of these meanings. In this sense, the above remarks concerning representativity do not relate to diachrony as such, but to the observation of past states of the language, whether synchronically or in terms of change.

Another difference is that it is often taken for granted that corpus linguists can be opposed to "armchair linguists", following Fillmore's (1992) caricatured opposition between an empirical approach and a hypothetico-deductive approach using fabricated examples to support a particular argument. This distinction has, in fact, little reason to exist in the study of past language states. The lack of intuition together with the impossibility of tapping into the competence of the native speaker means that the use of an authentic corpus, even a limited one, seems the obvious choice. It should be noted, however, that the study of language change is too frequently assimilated with the study of a (more or less) past language state. Indeed it is quite possible—and much of the work done indicates this—to focus on a short and 'contemporary' diachrony or on on-going developments (cf. Beeching, this volume). In this kind of research, recourse to introspection, to linguistic 'feeling' seems quite plausible: for example, recourse to the considered judgement of speakers from different generations concerning such-and-such an item is not wholly unthinkable. This kind of work on the variation of intuitions and tendencies towards change is not necessarily conducted within the context of corpus linguistics insofar as observation may be based on examples elaborated to suit a particular study's requirements. However, as Benzitoun (this volume) points out, this type of approach may be combined with a corpus study.

Finally, it is worth noting a somewhat paradoxical situation when it comes to using corpora in the study of past language states: our knowledge of old language is acquired through the handling, observation and analysis of texts, but at the same time our interpretations are made possible by the (albeit imperfect) knowledge we already have of the language (cf. Ingham, this volume). From this stems the necessity of increasing the number of corpora we work with, if only for the development of knowledge, for example, of one particular language state. This problem is not new and diachronic linguistics has continued to advance, in particular through the intermediary of philology, thanks to this special attention to the development of corpus use. Change in this field—that has implications for the discipline, as we shall see—has been made possible mainly through the development of electronic corpora and large electronic text databases.

However, despite appearances, this change is not necessarily of an inherently quantitative order. Work carried out well before the arrival of electronic corpora (cf. Imbs 1956 on verb tenses in Old French) was often based on considerable quantities of data of quite respectable size. The true innovation lies, it seems, in the way data is queried and observed.

Positive aspects

There is little need to emphasise the many practical advantages of corpus techniques which, due to their speed and the sheer quantity of occurrences, clearly allow research to be undertaken that would otherwise never have been completed through of lack of time and human resources. This represents a considerable advantage over more 'traditional' studies. The well-known work of Marchello-Nizia (1995, 2004) on demonstratives, for example, shows how traditional approaches often gave a biased account. Working with a large number of occurrences, with the possibility of obtaining reliable word-counts means, for example, that certain isolated cases do not overwhelm the analyses, and other forms may, in fact, appear as a 'normal' part of the system. Computer methods are also indispensable when it comes to connecting different parameters in the analyses, thereby revealing patterns or trends that would otherwise go unnoticed. As mentioned above, some of these comments could just as well apply to a synchronic approach, but are particularly relevant for the field of historical linguistics and have led to a true renewal of the discipline.

Modern corpora also allow flexibility in terms of handling that would have been hardly imaginable with an entirely manual analysis. Changes made to more 'traditional' studies were usually considered subsidiary rather than an integral part of the original study. Another 'technical' benefit is the possibility of combining several different databases. A good example of this is the *Dictionnaire du moyen français* (DMF) which provides access to both the dictionary and the texts (cf. Wissner, this volume).

In addition to these practical and quantitative aspects, corpora allow new ways of looking at language, or of reasoning and dealing with problems. These have more or less direct consequences for how the researcher construes the linguistic system and its development. The work of historical linguists before the era of electronic corpora was essentially, as we noted earlier, synchronic in nature, even over relatively extended periods, and did not really concern the question of change. Thus in any given area of study (e.g. the work on verb tenses and forms: Imbs 1956, Moignet 1979, Buridant 2000 on Old French; Martin 1971, Wilmet 1970 on Middle French; Gougenheim 1951 on the 16th century), it is as if the researchers, given the overall mass of data, did not dare take into account several successive language states and were consequently confined to a relatively narrow timeframe corresponding to traditionally recognised periods. This has changed dramatically with the use of large corpora, and it has become common to produce work dealing with change itself (cf. Marchello-Nizia 1999; Carlier & Goyens 1998; Verjans 2009). The implication is that technical innovations have helped give diachrony a new outlook within linguistics in general. It should be noted in passing that the increasing interest shown in grammaticalisation (see Isambert, this volume) goes hand in hand with the availability of fresh opportunities to study change. Most studies referring to this theoretical paradigm would hardly have been possible without the documentation and research tools offered by electronic corpora.

The large number of observable collections of data, together with the possibility of working on broad diachronies, will undoubtedly lead to a fresh take on periodisation and, more broadly, on variation. Historical linguists have become increasingly aware that not all elements of a language system evolve at the same pace, and that there are periods of stasis and periods of change. While this insight is not wholly new, linguists have in the past generally kept to comments on clearly distinct areas, by opposing, for example, lexis and grammar. Use of large corpora can help bring to the fore differences within subsystems (for example, indefinites do not evolve at the same rate as demonstratives or pronouns). This, then, leads us to question the currently accepted periodisation with arguments backed by quantified data. Concepts such as ‘Very Old French’ or ‘Pre-Classical French’ have emerged, in large part due to new means of analysis; and we can assume that the same could apply to the determination of sub-parts within the periods of Classical and Modern French[1].

Another innovation lies in the ‘vertical’ reading of texts via concordances, as opposed to the usual linear reading of a single text. Through the alignment of contexts (left and right), the use of concordances reveals facts that would otherwise have gone unnoticed. It is thus possible to build on a basic automated distributional analysis which brings various contextual parameters directly to the observer’s attention. Of course, the extraction of such information could be performed manually from a detailed reading of the texts, but the use of new tools not only allows a

[1] The *Grande grammaire historique du français* financed by the *Institut de la langue française* (ILF) is considering redefining periods in the history of French.

considerable saving in time, but also the use of a large quantity of data that would otherwise be difficult to handle.

In a completely different vein, certain changes to the way the scientific community functions could be heralded as a positive and promising step forward. Electronic corpora offer interesting perspectives for researchers who are not experts on diachrony or old varieties of language and who would simply not otherwise have ventured into reading texts in Old French or Middle French. Indeed, there is an increasing quantity of work that, while remaining fundamentally oriented towards contemporary French, is supported by data from the history of French and thus throws new light on the modern system. This type of research is well reflected in the successful series of “Diachro” conferences (cf. Fagard *et al.* 2008; Combettes *et al.* 2010), or in the work of the recently founded *Société internationale de diachronie du français* (http://www.sidf.group.cam.ac.uk/).

Finally, the development of large annotated corpora in particular requires collective reflection and interpretation. The management and implementation of a large corpus project requires teamwork between linguists as well as with computer experts (cf. Benzitoun, this volume). In this sense, thinking about new tools and taking time to engage with fundamental analytical questions has indirectly promoted collaborative research and created closer ties between members of the scientific community.

Potential problems

As every coin has its flip side, this section will focus on points that seem to be more negative or, at any rate, that invite a more ‘rational’ use of large corpora. Like the positive aspects of the previous section, these are of various different kinds and are not always specific to diachronic approaches.

First, a somewhat perverse effect of corpus-driven research has been to limit the areas of study, the items covered and, more broadly, the different fields observed. The main reasons for such a restriction are quite understandable: the starting point for nearly all work is the study of forms, expressions and specific points of interest made accessible principally through tagged corpora. This has brought about a considerable development in studies on lexis, on grammar-function words and on parts of speech. For example, the diachronic study of noun determiners within the theoretical framework of grammaticalisation and work on modal constructions or connectors within the framework of pragmaticalisation have thus undergone considerable developments (Combettes 2003; Badiou

2005; Féron 2007). However, most syntactic phenomena remain relatively neglected. While governing and ‘support’ verbs have been partly addressed by queries on specific verbal forms (Schøsler 2008; Troberg 2008), and some aspects of the complex sentence (Béguelin & Conti 2010; Combettes 2010) have been studied by way of questions pertaining to subordinating forms, everything contingent on the ordering of components and the structuring of units, appears very difficult to handle. It is perhaps regrettable that work such as Prévost (2001, 2011) on subject postpositioning remains the exception. It appears reasonable to expect that the development of morphological and syntactic annotation will alter this situation in years to come. The same applies to the study of tense and mood, since automated queries are not yet fully operational in work on past language states. However, some research on fixed expressions has been carried out as they are generally automatically identifiable from their surface form. Thus there appears to be a kind of drift, with corpus work entailing a notion of syntax more suited to fixed expressions than to free constructions. This change of perspective, which implies a significant theoretical shift, is rarely made explicit.

One advantage mentioned above is the possibility of taking into account wide-ranging diachronies, even going so far as to cover the entire history of the language. However, the resulting studies give the impression that only specific aspects of a given field can be covered, or a limited subset of a class of expressions, and not the entire language system. While it is true that earlier studies were often confined to relatively narrow time-spans, the fact that they were often conducted against a theoretical background more or less inherited from structuralist thinking meant that they tended to offer wider interpretations of relations within sub-parts of the system. It would be unfortunate if the progress achieved through corpus linguistics led merely to a large number of case studies, and overlooked the fact that change, while only observable in relation to a single form, in fact modifies the overall balance of things within the system. A specific example of this can be observed in the work on noun determiners (cf. Carlier 2001 on articles; Marchello-Nizia 1995, 2004 on demonstratives; Schnedecker 2005, Combettes 2004 on indefinite determiners): this often appears very fragmented, despite the fact that the evolution can be considered as a process of grammaticalisation concerning the category as a whole. The aim, then, should be to strike a better balance between new methodologies, query facilities and corpus handling, on the one hand, and theoretical ambition on the other hand (cf. Isambert, this volume; Ingham, this volume).

There is also a risk in limiting the scope to simple observation of data (cf. the positive side of concordance work mentioned above). Increasing reliance on concordances may lead to reduced 'linear' reading of texts, which may have several consequences. In some areas it may simply influence analysis and hence the results obtained. The types of linguistic environments that are typically queried are often quite limited; moreover, units such as the clause or sentence are not sufficient when it comes to analysing the scope of a connector or the discursive function of a topicalisation marker. On another level, if our knowledge of old language is formed and enriched by knowledge of the texts, we might consider that this enrichment is not fully achieved by consulting concordances, lists or tables. Besides, sustained reading may be the only way to arrive at certain conclusions. For example, certain ecological issues can only be addressed through a certain type of understanding of the data and knowledge (albeit indirect) of the conditions of production and use (cf. André & Tyne, this volume). Moreover, the general flow of language is not easily analysed without a 'natural' reading of the text. And, as pointed out earlier on, when dealing with past language states, there is no possibility of recourse to intuition or native-speaker judgement tests.

Other problems arise for the annotation of texts, which can have quite damaging consequences for the analysis and interpretation of facts. Indexing and coding, along with the development of automated analysers (virtually indispensable for research on morphosyntax) can be quite problematic. In particular, how does one avoid using categories that are likely to be anachronistic? It is well known that concepts are constructed and should always be relativised in order to avoid the pitfalls of naive realism. This issue is particularly complex in the diachronic study of language since linguistic categories are given to change. While some concepts, such as noun, verb or adjective, do not appear to pose any real problems, the same does not hold true for categories such as determiners, pronouns, prepositions or adverbs. In many cases the relatively clean-cut distribution observed in contemporary French only dates from the Pre-Classical period. The same applies to functions: syntactic relations (e.g. between the subject and the verb form, or between a verb and its object) do not always correspond to the same syntactic properties as in modern French, and it is quite plausible that the very notion of transitivity has evolved, for example. Thus we end up with a somewhat paradoxical position: on the one hand, refusing to tag with precision renders automated analysis procedures very difficult; on the other hand, a more accurate coding runs the risk of 'freezing' the analysis in ill-suited and ultimately deceptive categories.

A similar problem arises when trying to identify syntactic units and different divisions in texts. Here again, it seems very difficult to avoid anachronisms. For example, the notion of sentence is not at all clearly established before the end of the 18th century, so how can we define syntactic units in Old French, for example? Also, should we separate texts of the Classical era into specific periods? A further question arises as to the nature of clauses, in particular in relation to the concept of subordinate clauses and embedded sentences. While the situation is relatively clear for content clauses, difficulties quickly arise when it comes to determining the degree of dependence of some adverbial clauses: sentence-initial temporal adjunct clauses in Old French texts appear to be characterised by properties more closely resembling cases of parataxis than hypotactic constructions, and their grouping with causal or purpose clauses seems somewhat simplistic. We are again faced with a paradoxical situation: the corpus should, logically, be tagged according to the specificity of categories of a given time, but this specificity is precisely the object of study and tagging cannot be carried out until the categories are specified. Also, comparison between different periods would be rendered difficult if tags altered. One can only hope that progress will allow the evolution of concepts such as those mentioned here in the not too distant future, and that the problem will somehow be resolved through the continuous improvement of automated analyses. These ‘anachronisms’ may be a necessary but temporary evil that we should perhaps accept on the road towards more satisfactory labels.

One could argue that the avoidance of categories that are too ‘modern’ and therefore ill-adapted should also extend to notions that do not fall strictly within the field of linguistics but nevertheless quite rightly find themselves among language descriptors, such as those concerning genres and text types. For example, what we call a ‘descriptive’ text is not the same thing (i.e. it does not refer to the same linguistic elements) when dealing with 13th and 15th century texts. The very definition even of an argumentative text is far from stable and the well-known opposition between the narrative and argumentative systems does not seem applicable in the same way across all periods in the history of the language.

Conclusion

Summing up, it does not seem that there is any cause for undue alarm or extensive back-pedalling in the diachronic study of French. The overall view seems to be clearly positive and we can only rejoice in the renewal of diachronic studies which undoubtedly owes much to the emergence and development of new technologies. Most of the problematic issues that have been mentioned here are not insurmountable, and the combined effect of critical engagement, methodological rigour and collective reflection will surely be positive in this sense. Finally, it is worth recalling that the methods and results of previous generations of historical linguists contain lessons that are still relevant today: work on large electronic corpora should be mindful of the necessity of reading naturally as a means of understanding textuality; detailed study of particular cases should not detract from wider reflection on the functioning of the system as a whole.

Bibliography

Badiou, C., 2005, "Psychomécanique et évolution du signifiant: le cas du coordonnant négatif à l'aube du français moderne", *Langue française* 147, 84-97

Biber, D., S. Conrad & R. Reppen 1998, *Corpus linguistics. Investigating Language, Structure and Use*, Cambridge: Cambridge University Press

Buridant, C. 2000, *Grammaire nouvelle de l'ancien français*, Paris: SEDES

Carlier, A. 2001, "La genèse de l'article *un*", *Langue française* 130, 65-88

Carlier, A. & M. Goyens 1998, "De l'ancien français au français moderne: régression du degré zéro de la détermination et restructuration du système des articles", *Cahiers de l'Institut de linguistique de Louvain-la-Neuve* 24, 77-112

Combettes, B. 2003, "*Au contraire, en revanche, par contre*: aspects diachroniques", In P. Péroz (ed.), *CONTRE: identité sémantique et variation catégorielle*, Université de Metz, 269-287

—. 2004, "*Quelque*: aspects diachroniques", *Scolia* 18, 9-40

—. 2010, "Aspects diachroniques de la parataxe: les propositions temporelles en position initiale en ancien français", In M.-J. Béguelin, M. Avanzi & G. Corminbœuf (eds.), *La parataxe. Entre dépendance et intégration*, Bern: Peter Lang, 115-137

Combettes, B., C. Guillot, E. Opperman-Marsaux, S. Prévost & A. Rodríguez Somolinos (eds.) 2010, *Le changement en français. Etudes de linguistique diachronique*, Bern: Peter Lang

Fagard, B., S. Prévost, B. Combettes & O. Bertrand (eds.) 2008, *Evolutions en français. Etudes de linguistique diachronique*, Bern: Peter Lang

Féron, C. 2007, "*Pour vrai*, *pour certain*, *pour sûr*: formation et évolution d'adverbiaux en *pour*", *Langue française* 156, 61-75

Fillmore, C. 1992. "Corpus linguistics or computer-aided armchair linguistics", In J. Svartik (ed.), *Directions in Corpus Linguistics*, Berlin: Mouton de Gruyter, 35-60

Gougenheim, G. 1951, *Grammaire de la langue française du seizième siècle*, Lyon: IAC

Habert, B., A. Nazarenko & A. Salem 1997, *Les linguistiques de corpus*, Paris: Armand Colin

Imbs, P. 1956, *Les propositions temporelles en ancien français. La détermination du moment*, Paris: Les Belles Lettres

Marchello-Nizia, C. 1995, *L'évolution du français: ordre des mots, démonstratifs, accent tonique*, Paris: Armand Colin

—. 1999, *Le français en diachronie: douze siècles d'évolution*, Paris: Ophrys

—. 2004, "La sémantique des démonstratifs en français: une neutralisation en progrès", *Langue française* 141, 69-84

Martin, R. 1971, *Temps et aspect, essai sur l'emploi des temps narratifs en moyen français*, Paris: Klincksieck

Moignet, G. 1979, *Grammaire de l'ancien français*, Paris: Klincksieck

Prévost, S. 2001, *La postposition du sujet en français aux 15^e^ et 16^e^ siècles*, Paris: CNRS Editions

—. 2011, *Expression et position du sujet pronominal du 12^e^ au 14^e^ siècle*, unpublished thesis for Habilitation à diriger des recherches, ENS Lyon

Schnedecker, C. "*Certains* et ses avatars: approche diachronique", *Travaux de linguistique* 50, 131-150

Schøsler, L. 2008, "Etude sur les constructions à verbes supports", B. Fagard, S. Prévost, B. Combettes & O. Bertrand (eds.), *Evolutions en français. Etudes de linguistique diachronique*, Bern: Peter Lang, 345-361

Troberg, M. 2008, "Une étude diachronique des verbes datifs en français", in B. Fagard, S. Prévost, B. Combettes & O. Bertrand (eds.), 2008, *Evolutions en français. Etudes de linguistique diachronique*, Bern: Peter Lang, 385-403

Verjans, T. 2009, *Essai de systématique diachronique: genèse des conjonctions dans l'histoire du français (11^e^ – 17^e^ siècles)*, unpublished PhD thesis, University of Paris 4

Wilmet, M. 1970, *Le système de l'indicatif en moyen français, études des tiroirs de l'indicatif dans les farces, sotties et moralités françaises des 15^e^ et 16^e^ siècles*, Geneva: Droz

CHAPTER TWO

ACCOMMODATING THEORETICAL EXPECTATIONS WITH CONFLICTING CORPORA

PAUL ISAMBERT
FRANÇOIS-RABELAIS UNIVERSITY, TOURS AND LLL-CNRS RESEARCH UNIT, FRANCE

Introduction

Corpora are collections of stubborn data, accidental to some degree, that cannot be trusted at face value; not that they should be discarded without justification, of course, but they cannot be taken blindly for faithful images of reality, especially when they conflict with theoretical expectations. The question, then, is the following: when should we abandon hypotheses belied by available data, and when should we consider that known facts give an incomplete picture of reality? The latter solution of course implies that one should *show* that data are only partial; a simple stipulation to save a theory has zero scientific value.

This chapter investigates such a situation with the French adverb *autrement*. Today, three main uses can be identified: adverb of manner meaning 'another way' as in example 1; a connective denoting a negative hypothesis (example 2); or a discourse marker (not addressed here).

1. Il ne mangeait que du chocolat, j'arrivais pas à l'alimenter **autrement**
   'He ate only chocolate, I couldn't feed him in any other way'
   (PFC CORPUS)
2. Un traitement au laser à ce stade est des plus bénéfiques et efficaces. **Autrement**, la rétine malade va appeler d'autres néo-vaisseaux qui vont proliférer
   'Laser therapy at this point is most beneficial and effective. Otherwise, the diseased retina will call other neovessels which will proliferate'
   (L'EXPRESS)

These three uses seem to fit perfectly the steps of a path to grammaticalisation: the adverb becomes a connective and then a discourse marker, along the path that goes from referential to expressive and textual meaning (Traugott 1982, 1995). However, the connective occurs along with the adverb since the first Old French texts; actually, the very first occurrences of *autrement* (in the *Song of Roland*) are connectives. In other words, the data seem to belie the theory.

In this chapter, I will show how careful analysis of the data can be shown to display patterns that hint at grammaticalisation, despite the apparent contradiction. To do so, I will draw clues from contemporary data on the adverb and show that the connective in the first attestation was in fact highly restricted, and that those restrictions can be tied into a coherent whole through grammaticalisation theory.

Data

The first occurrence of *autrement* in French is a semantically unambiguous connective, as in example 3:

3. En la bataille deit estre forz e fiers,
U **altrement** ne valt .iiii. deners
'In the battle, he must be strong and wild, otherwise he isn't worth four deniers'
(*Song of Roland*, c. 1090)

Here, if we follow the grammaticalisation hypothesis, we could have expected an adverb of manner; ideally, this would be attested for quite some time before the connective gradually appears. However, this is not the case. As noted, the connective is unambiguous and cannot be interpreted as an adverb of manner in any way; the verb phrase *valoir quelque chose* ('to be worth something') cannot take a modifier (the second occurrence of *autrement* in the *Song of Roland* could be seen as an adverb of manner – see example 7).

It is important to note here that the data are sparse: there is no overwhelmingly clear picture of the use of *autrement* in Old French, let alone in one dialect or type of text only. Thus, even if the data did not seem to contradict the theory, a good deal of careful interpretation would be required as it would be rather easy to impose a hypothesis on (and claim it is confirmed by) insufficient data.

Three approaches can be envisaged in relation to this situation:

- Data are too sparse and do not refute the theory. In other words, the adverb of manner grammaticalised into a connective, even though the evolution is not attested in historical records. There is of course no good reason to pursue this line of thought: it is mere theoretical blindness which disregards facts. However, it could be entertained at worst if there were absolutely no relevant data, and we would have to rely on the theoretically most probable scenario.
- The evolution of *autrement* is a counterexample to grammaticalisation: according to records, a connective has turned into an adverb of manner. This approach implies that data, however incomplete, have the final say. It would also rely on the fact that unidirectionality in grammaticalisation has already been proven false (Campbell 2001a), so that the evolution towards increased grammatical function is only one possibility when considering the history of *autrement*. In other words, there is no good reason to be puzzled by data in the first place, unless one is uncritically committed to grammaticalisation theory (which, according to some, is not a theory anyway – Campbell 2001b). However, this idea has two flaws: first, with respect to unidirectionality, no counterexample has been found in the area of connectives (admittedly a ‘special’ kind of grammaticalisation in itself – Traugott & Dasher 2002); second, and more importantly, how can we explain such an evolution? It is not enough to simply state that it has happened, especially with such sparse data: one must show *how* it may have happened, otherwise the hypothesis is unwarranted. In this respect, such an approach fares no better than the previous one, because the data remain unexplained.
- The third option is simply to give up: nothing can be said about the history of *autrement*, end of story. While admittedly safe, this ‘approach’ prevents any testing of a theory with difficult material; it eschews undue generalisations and speculations, but at the same time it does not acknowledge what theories are made for, i.e. to shed light on otherwise unexplained data.

In this chapter, I will advocate the idea that carefully interpreted data, building on what we know of *autrement* in contemporary French, can be accounted for in terms of grammaticalisation, which serves as a heuristic. In this way, the history can be (hypothetically) reconstructed if one pays sufficient attention to the fine-grained properties of the attested utterances instead of merely classifying the occurrences of *autrement* according to broad categories (cf. Combettes, this volume). In other words, simply pegging *autrement* in example 3 above as a ‘connective’ hides subtler characteristics which must be teased out.

Interpreting data

What should we look at? How should we go about determining what the 'subtler characteristics' mentioned above are, and whether they exist in the first place? Here, I will rely on data in contemporary French (see Isambert 2010 for a detailed account). This does not necessarily imply that the adverb has not changed in a thousand years, and that we can make up for insufficient historical data with contemporary sources. And yet the modern use displays trends that may be reflexes of Old French, or that we can at least take as clues to be investigated. For example, Torres Cacoullos and Walker (2011) show that well-advanced constructions like the *go*-future nonetheless show striking differences when compared across French, Portuguese and Spanish, and that these differences are relics of the past. Thus the differences between what we call the connective use of *autrement* in Old French and the same function in contemporary French may be telling.

On a different line, important insights from the most recent grammaticalisation studies can be put to good use in our case: most importantly, it is now widely acknowledged that *words* do not grammaticalise, but *constructions* do (Bybee 2003; Traugott 2003; Joseph 2004). Thus *autrement* should not be looked at alone, but in the whole context, and the full expression, where it appears, should be addressed.

Finally, grammaticalisation is a gradual process, so the new meaning of an expression can be found in just a few contexts at first, before it extends to a less determined environment (what Himmelmann 2004 calls "host-class expansion"). This means that if *autrement* has grammaticalised from an adverb of manner into a connective, and provided that this evolution is not too old when *autrement* first occurs in historical records, then we should be able to observe restrictions in its use as a connective.

If we compare the adverb of manner in contemporary use with its counterpart in Old French, no difference can be observed. In particular, no restriction (when compared to contemporary usage) seems to have applied in Old French, as far as can be judged from historical records (perhaps the adverb underwent semantic change before records began). This observation excludes expressions occurring today that did not exist, or are not attested, in the past, for instance *autrement dit* ('in other words') or *voir les choses autrement* (typically found in journalistic texts, meaning 'to have a different opinion or to disagree', and generally introducing direct speech). But the evolution of these and other forms can be tracked, the former to an expression appearing in the 16th century (meaning literally 'said otherwise' at that time), and the latter to modern journalistic style. In other words, they are later entrenchments of the adverb of manner into collocations, themselves perhaps open to new developments, but quite restricted for the moment.

However, there are significant differences in the use of the connective when compared across time. Nowadays, simplifying greatly (ignoring in particular the many cases where it is ambiguous with its third value, the discourse marker), it occurs in three main types of context (see invented examples 4 to 6): deontic (example 4; also example 2 above), epistemic (example 5), and in conditional structures (example 6):

4. Fais ceci, **autrement** il y aura telle conséquence (négative)
   'Do this, otherwise there will be such and such (negative) consequence'
5. X est vrai, **autrement** Y le serait (et ça ne l'est pas)
   'X is true, otherwise Y would be true (and it isn't)'
6. Si P alors Q, **autrement** (c'est-à-dire si non-P alors) R
   'If P then Q, otherwise (i.e. if non-P then) R'

The connective in example 4 is used to convey the idea that something (described as an injunction in the first clause) should not happen and considers the potential outcome negative (the second clause), thus lending support to the injunction. In example 5, it denotes the hypothetical falsity of the first clause and introduces a clause that contradicts common knowledge, thus proving that the first clause is true; in other words, 5 is similar to 4 except that it operates on truth values, not expectations. In example 6, it reverses the polarity of an already hypothetical clause (typically, a subordinate clause headed by *if* or *when*) and introduces a statement true under that hypothesis. Interestingly, of these three main patterns, what we see in example 4 is the most frequent; also, there is a well-known cline from deontic modality to epistemic modality, i.e. from 4 to 5, so we may expect 4 to be older than 5. Example 6, on the other hand, is more concerned with textual organisation, and actually paves the way for the discourse marker alluded to in the introduction.

Only the first, deontic use occurs in the oldest historical records. It can be noted that in example 7, *autrement* does not target the initial subordinate clause; thus the utterance does not follow the same pattern as example 6. In both examples 7 and 8, the speaker asserts that something must be done or else something undesirable will happen, as described in what follows *autrement*. Non-deontic contexts, and especially conditional structures, appear only in the 16th century. Thus the connective is restricted in Old French, and appears unsurprisingly in contexts which may engender other uses (deontic use then allows epistemic use, not the other way