Syntax and Corpora

Michèle Oliviéri

Référence électronique

Ce document est un fac-similé de l'édition imprimée.

© Tous droits réservés
1. A long-standing controversy

The notion of corpus represents a long-standing and controversial matter in syntax. Before Chomsky, structuralists exploited corpora however these were considered as representative and homogeneous samples of a particular syntactic phenomenon or of a given language. This type of corpus was thus used within a behaviourist approach exclusively with empiricist and taxonomic aims. Then, as soon as Chomsky started to develop generative grammar in the 1950s, it is precisely this type of “corpus” he argued against. Indeed, he shows for instance that language cannot be reduced to a limited number of utterances insofar as the speaker’s competence allows him/her to utter an infinite number of grammatical sentences:

\[
\text{it is obvious that the set of grammatical sentences cannot be identified with any corpus of utterances obtained by the linguist in his field work.}^1
\]

Moreover and quite obviously, it is particularly difficult to guarantee that a sample is representative of a language as well as homogeneous since both characteristics are relatively contradictory. Nevertheless, Chomsky has never totally rejected the necessity of resorting to primary data within a “discovery procedure”. However, he has striven for a theory which looks

\footnote{Chomsky (1957:15).}
beyond the mere observable facts and which accounts for the “general nature of language”. From this point of view,

any grammar of a language will *project* the finite
and somewhat accidental corpus of observed
utterances to a set (presumably infinite) of grammatical utterances.²

The competence defined by Chomsky – nowadays *I-language* – enables the speaker to utter and understand an unlimited number of words and sentences in his/her language but also – and foremost – allows the speaker to carry out judgments of acceptability. Grammaticality judgments have hence always represented a centrepiece in generative research, turning syntactic analysis into introspection. Indeed, as generative grammar aims at describing competence, the investigations focus on the internal knowledge a speaker has of his/her language. Consequently, the analyses aspire at unveiling what happens in his/her brain, which is not necessarily present in a corpus.

Such an approach is then totally justified within this theoretical framework. However, this rationale has also entailed unfortunate consequences and endless debates hinging upon some artificially coined sentences which appear recursively in the literature. Some of these utterances have even become famous despite the fact that they only have little chance of actually emerging spontaneously:

(1) L'homme que je crois que Jean pense qui viendra…
“the man that I think that Jean thinks who will come…”
Qui crois-tu qui/*que téléphonera ?
“who do you think who/*that will telephone?”
Il n'a pas été trouvé de livres.
“it was found no books”

Nevertheless, generativists mainly pursue with this methodology despite the difficulties it entails especially with regard to diachrony since it is not possible to collect judgments from speakers of the previous centuries.

² Chomsky (1957:15).
Quite fairly then, Chomsky’s opponents emphasize these limitations in order to either discredit generativism or advocate other approaches. Indeed, corpus linguistics has considerably developed over the past few years. This expansion is also due to another favourable parameter, i.e. data processing, providing scientists with up to then unavailable possibilities with regard to resources as well as software.

2. On grammaticality

Generative investigations have hence focused on *I-Language* as representing the Faculty of Language, i.e. the “component of human biology that enters into the use and acquisition of language” (Chomsky 2005). This type of research has in turn led generativists to mainly rely on grammaticality judgments.

On the other hand, every linguist who has undertaken such tests knows that grammaticality judgments can easily vary from one speaker to another and that a speaker can sometimes feel uncertain about the grammaticality of an utterance. As a consequence, a sentence can be definitely asterisked by certain authors whereas the same one can be considered as fully acceptable by others.

Moreover, it has often been noticed that speakers’ judgments are not always reliable when it comes to oral and non-normative languages. Indeed, some speakers can categorically reject a structure and nevertheless utter it frequently however unconsciously. Quite obviously, people do not speak as they think they do. How to deal then with such judgments? One of the most striking examples of this discrepancy in French is the sentence presented in (2). Indeed, this example is often asterisked, especially when it is compared with dialectal forms which feature an obligatory subject clitic, despite the fact that (2) actually corresponds to the common structure in spontaneous French.

(2) Marie elle dort.

“Mary she sleeps”

Nevertheless, analyses rely on such judgements, sometimes leaving the reader puzzled. It would then be
M. OLIVIERI

desirable to judge the grammaticality judgments, assess the assessments, but these tasks sound unrealistic. When is an utterance “erroneous”, from the speaker’s point of view, the linguist’s one? When is a sentence the result of variation within a language or even of diglossia?

This is when the corpus comes to the rescue of introspection. Indeed, only a corpus can actually attest of an occurrence which would otherwise be left out of the analysis. It is also the corpus which can help a researcher to assess the frequency of a structure and contrast forms. And last but not least, it is the corpus which gives access to now extinct languages.

3. Broad corpora and linguistic variation

Dialectologists – to include myself – are used to fieldwork and data compilation. Luckily, gone are the days when “theory” and “empiry” were separate, with subordinate fieldworkers on the one hand and theorists who did not leave their labs on the other hand. Indeed, collecting data – and collecting them personally before analysing them – is a widely acknowledged task nowadays3. Thus, choosing the speakers, devising the questionnaires, applying various techniques, listening to and transcribing data now all represent key steps in corpus compilation and scientific theorization since there is a permanent two-way feedback between the field and the lab. Indeed, the corpus allows the scholar to forward an analysis, but the elaboration of this corpus is predetermined by a framework and a point of view. As a consequence, theory and corpora are tightly bound.

Now, modern technologies offer many different possibilities. Indeed, recent progress in data processing not only provides recordings of better quality and of easier accessibility but it also enables linguists to compile and use very big corpora. A corpus hence no longer stands as a “sample”, whether representative or not, but as a multidimensional and elaborate object, which is constantly modifiable and searchable in many

3 Dalbera (2002), among others.
different ways, and whose size guarantees reliability (according to the famous “more data, better data”). Even though broad and changing corpora still represent only partial visions of the unlimited number of possible utterances and despite the fact that the linguist has to temporarily define their boundaries in order to achieve his/her investigations, it now appears that Chomsky’s initial criticism against corpora is not as strong as it used to be. Why not (re-)conciliate generative grammar with corpus linguistics then? Besides, Chomsky (1965) was already advocating that:

It is not, of course, necessary to assume that empiricist and rationalist views can always be sharply distinguished and that these currents cannot cross. [...] [They] can be presented as explicit hypotheses [...] about the innate structure of a language-acquisition device. [...] The grammars that result from application of the postulated procedures to a sufficiently rich selection of data will be descriptively adequate — in other words, [...] the set of procedures can be regarded as constituting a hypothesis about the innate language-acquisition system.4

This is precisely what several researchers have initiated, especially dialectologists who hold important masses of “actual” data allowing them to test the hypotheses based on one or two utterances from standard languages. In parallel, diachronic syntax, which had only little developed so far within generativism, is currently expanding thanks to broad corpora of written texts as well as dialectal data which reveal successive linguistic stages. For instance, relying on data from the morpho-syntactic module (MMS) of the THESAURUS OCCITAN database (THESOC)5, I forwarded a parameterized account for the

4 Chomsky (1965:52-53).
5 THESOC is a multimedia database displaying the whole range of Occitan dialects and developed within UMR 6039 "Bases, Corpus, Langage" (dir. J-Ph. Dalbera and M. Oliviéri): http://thesaurus.unice.fr. This database includes lexical data as well as annotated and syntactically analysed oral texts. See Brun-Trigaud & Oliviéri (2009), and Georges (2005, to appear).
emergence of subject clitics in Romance languages. Indeed, if one dialect is considered as revealing one state of Saussure’s chess game, then one can “read” the geographical distribution of different dialects on a linguistic map differently and reconstruct diachrony.

The huge advantage of these broad corpora is that they enable researchers to study linguistic variation, whether diatopic or diachronic, without limiting the investigations to a typological point of view.

Indeed, since variation is present at different levels, the facts are often blurred. The first type of variation, the inter-linguistic one, is widely acknowledged and leads to a typological sorting of the languages in the world. In order to achieve such a classification, generativists have been working at devising “macro-parameters” since the GB framework.

The second type of variation, which is gradually becoming a centrepiece in the analyses, is micro-variation. It accounts for more fine-grained differences and is observed between cognate dialects. Nevertheless, it also represents inter-linguistic variation since it occurs between close, yet structurally distinct systems. As a consequence, syntacticians studying dialects have forwarded “micro-parameters” hence accounting for more subtle and more changing behaviours than those found in standard and normed languages.

Finally, a third type of variation exists, namely internal or intra-linguistic variation. Dialectologists know it well but historically syntacticians have not taken it sufficiently into account despite its omnipresence in human languages. Indeed, it has generally been considered as stylistic or sociolinguistic and, as such, has not often been seriously investigated from a syntactic point of view. Actually, speakers of one same language do not use identical internal systems and this practice is sometimes identified as “diglossia”. Consequently, every

---

6 Oliviéri (2009, 2010b), consisting of a progression in terms of features, i.e. [Person], [Speaker], [Gender], and [Number].
7 Kayne (1996).
speaker is plurilingual and masters several linguistic systems, each with its own internal logic which needs describing and disentangling. The difficulty lies in the vagueness of such productions, the subjectivity of grammaticality judgments, and the necessity of establishing precise criteria in order to discriminate these different systems, these different “lects”.

Thus, stemming from a study on subject clitics in child French, Palasis (2009, to appear) illustrates that the two French systems she describes both have an internal coherence and that the ungrammatical utterances are in complementary distribution in these two systems, as shown in (3):

(3) Normed French:  Jean dort.  "John sleeps"
    Il dort.  "He sleeps"
    Jean, il dort.  "John, he sleeps"

    Spontaneous French:  (a) Jean i-dort.
                        (b) *Jean, il dort.
                        i-dort.  *Jean dort.
                        *Il dort.

This point of view then leads to reconsider the facts which are to be analysed and compared with attention in order to determine the universal principles and the parameters pertaining to each system. Can two utterances stemming from two different systems really be put on the same level without taking the existence of the two systems into account? 10 Quite obviously then, oral and spontaneous languages must be compared with other oral systems, i.e. dialects can be collated with oral French, oral Italian or oral Portuguese but not with normed and standard systems. Otherwise, conclusions could well be erroneous due to flawed grammaticality judgments.

---

9 Oliviéri (2010a:58).
10 As stated in Scheer (2004:58): “l’ambition générative à découvrir les universaux doit s’accompagner d’un contrôle structuraliste” (the generative ambition of unveiling universals must be backed up with a structuralist control).
4. Presenting the contributions

This special issue on corpus syntax displays an overview of approaches in this field. The contributions in this volume thus pertain to different theoretical frameworks and address a range of matters in diverse languages from various points of view. Nevertheless they all demonstrate how important it is for a researcher to work from a (broad) corpus, compiled and elaborated by the researcher him/herself in order to forward syntactic analyses.

Katérina Palasis’ paper focuses on child data and L1 acquisition. Within the generative framework and thanks to personally recorded data, she illustrates how fieldwork, corpus compilation, and theorization are closely interwoven and how the type of data has impacted on some past conclusions. In order to study the pro-drop phenomenon in child French, K. Palasis has hence compiled a new corpus which will eventually be shared on CHILDES. She forwards a fine-grained analysis of several issues on data compilation and lays principles in order to obtain a relevant corpus which can then be shared with the scientific community. Her analysis of these data, which takes the entire child verbal system into account, leads her to reconsider some previous studies on the null subject parameter and to forward new theoretical hypotheses.

Maria-José Ezeizabarrena and Sandrine Aeby also focus on acquisition but in a bilingual context. Their article, which also has a generative background, deals with code-switching in adult-child and child conversations in a Basque-Spanish bilingual context. M.-J. Ezeizabarrena et S. Aeby methodologically explore different types of Basque-Spanish code-switching and put previous analyses to the test of facts. They detail their two longitudinal and cross-sectional corpora from a methodological as well as a scientific point of view. The analysis of their data then unveils the fragility of previous approaches and shows that code-switching is amenable to factors which belong to the syntax-pragmatics interface.

Inés Fernández-Ordóñez’s contribution is dedicated to presenting an important corpus, the Audible Corpus of Spoken Rural Spanish (COSER, after its Spanish abbreviation), which is
applied to the syntactic and morpho-syntactic study of Spanish dialects. I. Fernández-Ordóñez details the methodology and functionalities of this corpus and illustrates how several syntactic phenomena are accounted for thanks to this tool. Indeed, COSER quantitatively as well as qualitatively complements facts and observations from preceding linguistic atlases. The analyses based on COSER hence bring a new point of view as well as further prospects on these dialects. This contribution also highlights that facts which had not, or only partially, been investigated beforehand are not inconsiderable and are actually part of a coherent system.

Catarina Magro works within the Distributed Morphology framework. Her contribution deals with the well-known phenomenon of interpolation in Portuguese dialects. She relies on another broad corpus, the Syntax-oriented Corpus of Portuguese Dialects (CORDIAL-SIN) and questions common belief on interpolation as an archaic structure, showing that this characteristic is actually productive in modern European Portuguese. The enormous advantage of C. Magro’s approach lies in the possibility of collating actual data to previous hypotheses and predictions which had been devised without any data. Consequently, this corpus allows the author to forward a new point of view on a much-described issue and to conclude with further proposals.

Franziska Hack’s paper investigates interrogation in several Ladin dialects of Northern Italy. The author examines data from pre-existing corpora, such as atlases and databases, and discusses the advantages and disadvantages of these sources. This study leads her to compile an additional corpus thanks to further personal fieldwork whose methodology and choices are accounted for in the contribution. For instance, the main objective of this new corpus is to obtain data which are absent from atlases, e.g. negative data and grammaticality judgments, in order to comply with her generative framework. Here again the theoretical rationale has an impact on the principles of the fieldwork and further drives the compilation of the corpus which, in turn, leads to the analysis.
The article by Leonardo M. Savoia and M. Rita Manzini, also dedicated to some Northern Italian dialects known as “Occitan” or “Franco-provençal” dialects, deals with subject clitics. The authors rely on a broad corpus stemming from their numerous dialectal surveys and account for various mechanisms involving subject clitics, such as doubling of the lexical subject, interrogation, and interaction with object clitics. On the one hand, the fine-grained analysis of these data within the generative framework allows the authors to confirm and precise already-existing generative hypotheses. On the other hand, L. M. Savoia and M. R. Manzini forward new insights into dialectal typology, hence contradicting common belief in this domain.

Sascha Gaglia also examines subject clitics in some Italian dialects, i.e. the Friulian varieties. Within the framework of Optimality Theory, the analysis relies on an oral survey based on grammaticality judgments compared with a written corpus from the Bible. Moreover, S. Gaglia takes previously published data and outcome into account (Renzi & Vanelli (1983) and Heap (2000, 2002), among others). Examining all the subject clitic omissions, he then demonstrates that this phenomenon pertains to the presence of negation, object clitics, and reflexive pronouns. He forwards an analysis in terms of constraints aiming at avoiding syncretism.

Sandra Augendre’s contribution deals with dislocation in Italian and bears on the syntax-pragmatics interface. Her corpus is composed of a particular type of data since these are written however informal texts, i.e. e-mails, chat, and text messages. Within her corpus, the author examines different aspects of dislocation, i.e. subject and object, right and left, from two points of view, syntactic and communicative. Further to the distributional classification S. Augendre forwards of the different types of dislocation and thanks to her corpus and analysis, the author also proposes several hypotheses with regard to the pragmatic functions of the syntactic structures. Indeed, she demonstrates that there are different types of dislocation which each fulfil a particular function.
Christophe Benzitoun, Solène Bresson, Laure Budzinski, Jeanne-Marie Debaisieux, and Klara Holzheimer forward a distributional analysis for the adjective prochain ‘next’ in French. Their investigations stem from an important diversified corpus since half their data is oral whereas the other half comes from written texts. Such a compilation then allows the researchers to compare both systems qualitatively as well as quantitatively. Moreover, collating their results with those from analyses undertaken thanks to specially coined utterances, the authors demonstrate that corpora forward indications that introspection does not unveil. Their work then confirms that oral data are important for linguistic studies and that elaborate writing or dictionary entries only give a partial vision of linguistic reality.

Michael Zimmermann and Georg Kaiser’s paper also illustrates how theorization and data are intertwined. Indeed, their topic – Medieval French – leads these authors to compile an annotated text corpus. It is then shown that coding raises theoretical issues which, in turn, can be solved by going back to data. Thus, while assigning a categorial status to et ‘and’ and ne ‘not’ in Medieval French, M. Zimmermann and G. Kaiser show how different theoretical backgrounds can impact on the annotation and analysing processes of a corpus. As a consequence and further to fine-grained analyses of the facts against previous work, the authors forward a new hypothesis on the status of these elements.

A single issue of Corpus cannot do justice to the diversity of approaches in corpus syntax. However, we have tried to forward if not homogeneous at least representative samples of the kind of research that can be carried out in this domain. Indeed, all these contributions share one same need, i.e. a huge amount of data. Consequently and fortunately for the scientific community, many databases have emerged over the last years and these are more and more elaborate. Moreover, in this issue, a corpus always appears as an asset, i.e. a tool which allows researchers to shed further light on linguistic phenomena. Indeed, all these data-based studies lead
researchers to forward new analyses, to elaborate original hypotheses often against what is generally stated in the literature. Furthermore, taking performance, this E-language which surfaces in corpora, into account also leads to approach the interfaces of grammar and to investigate the connections between syntax and pragmatics or phonology. In turn, all these studies then also appear as very promising in order to understand the underlying mechanisms of human language and to grasp the architecture of its grammar.

References


