TOOLS FOR DIALECT SYNTAX: THE CASE OF CORDIAL-SIN (AN ANNOTATED CORPUS OF PORTUGUESE DIALECTS)

Ernestina Carrilho
Universidade de Lisboa

Abstract

This paper addresses methodological issues of concern to the study of morphosyntactic variation. While the empirical basis of dialect syntax is still a matter of elaboration, the focus will be here on the role of dialect corpora as tools for the study of linguistic variation in this particular domain. The case of CORDIAL-SIN, an annotated corpus of Portuguese dialects, will be presented along with some initial advances in Portuguese dialect syntax. Two levels of tools for the study of linguistic variation will thus be addressed here: (i) corpora as general tools for dialect syntax; and (ii) tagging and syntactic annotation within a dialect corpus as tools that ease the way how variation in morphosyntax can be studied.

Section 1 introduces methodological remarks concerning the empirical ground for dialect syntax; the CORDIAL-SIN is presented in section 2; section 3 briefly illustrates how this tool has enhanced the development of Portuguese dialect syntax.

Key words: *Dialect syntax; morphosyntactic variation; dialect corpus; annotated corpus; European Portuguese dialects.*

1. On the empirical basis of dialect syntax

It is well known and often mentioned that the study of syntax has only played a very marginal role in traditional dialectology. Dialectologists have mainly been concerned with the study of phonological and lexical variation, and it was for this purpose that data were systematically collected in dialect surveys and linguistic atlases, which represent the main data source for traditional dialectology.[1] In fact, in the atlas projects all over the world only a scarce part of the published dialect maps involve syntactic data (Cornips and Jongenburger 2001: 1).

This neglect of syntax in dialect studies certainly owes much to the methodological difficulties that syntactically oriented fieldwork raises. The classical method of dialectological interviews, conducted with the help of a questionnaire, hardly combines with the gathering of specific syntactic constructions, for which naming ques-

[1] To this respect, it is worth remembering some notable exceptions such as Remacle's (1952-60) work on the syntax of the Walloon dialect of *la Gleize*.

tions or even completing questions happen to be fairly unhelpful. Also, oral translation from the standard language, a method used for eliciting syntactic properties in most European linguistic atlases, is far from unproblematic. It has been acknowledged that this kind of elicitation raises several problems, among which a high risk of getting an answer influenced by the standard construction (a.o. Bucheli and Glaser 2002: 3). Besides, such a method is only conceivable (despite its imperfections) for those areas where variation meets different linguistic systems (which may be the case of Italy, France or Switzerland, but is not the case for the most part of the Portuguese territory, for instance).

The last decade of the 20th century witnessed however a general and renewed interest in syntactic variation[2] and, concomitantly, new methodological concerns were brought about the empirical ground for this domain of linguistic inquiry. In fact, over the last two decades, several projects dealing with the syntax of dialects have been independently established in different European countries, some of which are still under development: among others, the *Syntactic Atlas of the Dutch Dialects* (SAND); the *Syntactic Atlas of Northern Italy* (ASIS); the *Audible Corpus of Spoken Rural Spanish* (COSER); the project *English Dialect Syntax from a Typological Perspective*; and, more recently, the supranational projects *ScanDiaSyn (Scandinavian Dialect Syntax*), across the Scandinavian dialect continuum, and *Edisyn* (*European Dialect Syntax*), an European project specifically aimed at developing cooperation among dialect syntax projects in Europe through similar or common methodologies (regarding data collection, data storage and annotation, data retrieval, cartography). Among these projects, the data that feed the empirical demands of dialect syntax range from a corpus of independently collected speech (as in the project *English Dialect Syntax from a Typological Perspective*) to written questionnaires requesting translations into the informants' dialect (as in the first phase of ASIS). The advantages and disadvantages of both naturalistic methods behind corpus-data and elicitation techniques have been recurrently discussed. It may easily be acknowledged that none of them is exempt of problems.

Although naturalistic corpus-based data can hardly circumvent problems such as the lack of negative evidence and the weak representation (if any) of sentence types that are rare in spontaneous speech, the experience of fieldwork data collection through elicitation has however proven that this method is not free of difficulties either (see also Labov 1996).

> Every elicitation situation is artificial, because the subject is being asked for a sort of behavior that is entirely different from everyday conversation (cf. Schütze 1996: 3). Sociolinguistic research has clearly shown that the response of subjects on direct judgement tasks ('Is this a good sentence in your dialect?') often tends to reflect the form which they believe to have prestige or obeys the learned norm, rather than the form they actually use (Labov 1972: 213). A reasonable alternative is to use more indirect elicitation tasks (e.g. 'Do you encounter this sentence in your dialect?') Different levels of speech style (informal and formal) yield another complicating factor for syntactic data elicitation. (Barbiers and Cornips 2002: 8-9)

[2] Actually, such an interest was already sketched in the late seventies within the international geolinguistic project ALE, *Atlas Linguarum Europae*, which already stated the convenience for syntactic theory to count on a comparative inquiry into dialect syntax (see Lehman 1980, Kruijsen 1983).

In fact, past experiences have often shown that the results obtained through elicitation data may differ from those appearing in spontaneous speech (Cornips 2003); also, different elicitation methodologies may often lead to different results (Auckle, Buchstaller, Corrigan and Holmberg 2007). At least, such results appear to suggest that more research is still needed in order to decide on the reliability of the different elicitation techniques.

The pratice developed within the SAND project (Cornips and Poletto 2005, Barbiers et al. 2007), known as a "layered methodology", may be taken as exemplar. The phases of planning the SAND data collection involved, as a first step, a comprehensive literature study. As a second step, a written questionnaire has been prepared on the basis of the syntactic phenomena described in the literature. As Cornips and Jongenburger (2001) report, this questionnaire was carefully prepared to provide insight into (i) the geographic distribution of syntactic variation; (ii) the validity of each type of (written) elicitation; (iii) areas particularly interesting with respect to syntactic variation. As such, the questionnaire served as the input for the next phase, which consisted of oral fieldwork. Preparing the oral fieldwork for the SAND project involved the consideration of an appropriate elicitation task for each syntactic variable to be investigated. The results with respect to the usability of both written and oral elicitation techniques show that not every task is easy to perform and that not every task is adequate to every type of syntactic variable. Fundamental aspects concerning the reliability and the workability of elicitation in dialect syntax may thus be evoked: (i) though useful, elicitation tasks are not without problems; (ii) the negative effects associated to each elicitation task must always be carefully evaluated; (iii) combining different types of data collection methods may help obviating some problems; (iv) dialect syntax analysis requires careful consideration of the relation between the collected data and the effects induced by the method by means of which the data are obtained.

Dialect syntax projects may thus take advantage from a layered methodology that can combine different sources of data collection tasks: besides the appropriate elicitation tests, also interview techniques that can generate more naturalistic speech. Such a practice has recently been adopted in large-scale dialect syntax projects, such as the *ScanDiaSyn*.

As a matter of fact, naturalistic data have also played a non-negligible role in setting up recent advances in dialect syntax. Within different linguistic domains, dialect corpora became also important heuristic tools for the study of non-standard syntax. For instance, we may refer to the *Freiburg English Dialect Corpus* (FRED), a computerized corpus of English dialects, within the project *English Dialect Syntax from a Typological Perspective* (Kortmann 2002), or to the above mentioned COSER, seminal to recent research on several aspects of Spanish dialect syntax (see Fernández-Ordóñez 2009).

Thus, even if we acknowledge the importance of introspective syntactic data that only elicitation tasks may provide, the limits of such a data collecting methodology are also to be remembered when it comes to the study of non-standard syntax. The linguist preparing such data collecting tasks can hardly be familiar with the different varieties of his native language.

> One point which might be made is that this method [the introspective method, EC] cannot be used to study any language or language variety not known to the in-

> vestigator, and since academic linguists are seldom competent speakers of non-standard dialects or uncodified languages, can in practice be used for describing only fully codified languages. This is not of course to deny that those who have grown up as native speakers of a dialect (for example, Peter Trudgill in Norwich […]) may have intuitions about its structure; so also might non-native speakers who have developed an intimate knowledge of the structure of a dialect (see J. Milroy 1981 for an example). But descriptions of non-standard dialects generally use intuition as an aid to focusing the investigation, rather than a basic method; […]. (Milroy 1987: 76)

Above all, introspection alone could hardly be invoked as a source of hypotheses-motivating data central to elicitation tests' design. In this context, thus, dialect corpora can play a different role. Observations sometimes formulated with respect to the empirical basis of linguistic research in general appear especially significant when referring to empirical methodologies applied to the study of dialect syntax:

> The advantage of working with a corpus is, of course, the enhanced objectivity of the data and of all the research that is based on it. In comparison with the other approaches, the possibilities for the researcher to manipulate the data are minimized. Another great advantage is that a corpus the researcher has not produced himself may be varied, heterogeneous, full of surprises and a constant source of inspiration. Exposing oneself to spontaneous data is, in fact, the safest way of discovering those categories of a language [EC: or of a dialect] that are peculiar to it and that the researcher did not expect. (Lehmann 2004: 201)

This is so much so to the extent that comprehensive or specific descriptive dialect syntactic studies are often unavailable for some languages. A dialect corpus, if available, may then be *full of surprises*.

2. CORDIAL-SIN: the syntax-oriented corpus of Portuguese dialects

2.1. Background

By the end of the 20th century, the condition of dialect syntax in Portuguese studies was not significantly different from what happened in other linguistic domains. The major Portuguese dialect surveys had not generally contemplated any kind of syntactic variation, and the questionnaire of the linguistic atlas ALEPG (*Atlas Linguístico-Etnográfico de Portugal e da Galiza*) explicitly stated that "for practical reasons" it did not include syntactic questions (Gottschalk, Barata and Adragão 1974).

Nevertheless, even if no comprehensive description of syntactic variation phenomena was available, there existed sparse allusions to syntactically relevant variation in European Portuguese. These could in fact be found from the pioneering work in Portuguese dialectology by Leite de Vasconcellos (1901) to many different dialect monographs written near the mid-20th century. However, the place for syntax was usually very marginal when compared to that of lexicon, phonology or even morphology.

It was against this background that CORDIAL-SIN began to be compiled, in 1999. The acronym stands for the Portuguese name "*Cor*pus *Dial*ectal para o Estudo da *Sin*taxe" ("Syntax-oriented Corpus of Portuguese Dialects") and it has mainly been conceived as a major empirical resource for the study of dialect syntax.

As a very important condition for the CORDIAL-SIN genesis, I shall mention the existence of a rich collection of tape-recorded dialect speech, gathered by the Centro de Linguística da Universidade de Lisboa. At this Center, the research group working on Linguistic Variation has been committed to several projects of dialect geography, for which fieldwork interviews have been conducted from the mid-seventies till the beginning of 2000. In the course of such interviews, informants often speak about their story of life, make observations on aspects inquired by the questionnaire, comment on etnographic issues, which amounts to an important extent of spontaneous speech, collected in fairly controlled and homogeneous conditions.

The CORDIAL-SIN project aimed precisely at making available for researchers in general (and especially for those interested in dialect syntax) a significant amount of spontaneous and semi-directed speech drawn from these data. More specifically, this project also aimed at providing fast and systematic access to precise morphological and syntactic information —which motivated the building up of an annotated corpus, marked up with morphological and syntactic information—.

By compiling and making available such an empirical resource for dialect syntax, the CORDIAL-SIN team has also been engaged in the enhancement of research activity on syntactic dialect variation in European Portuguese.

2.2. A dialect corpus

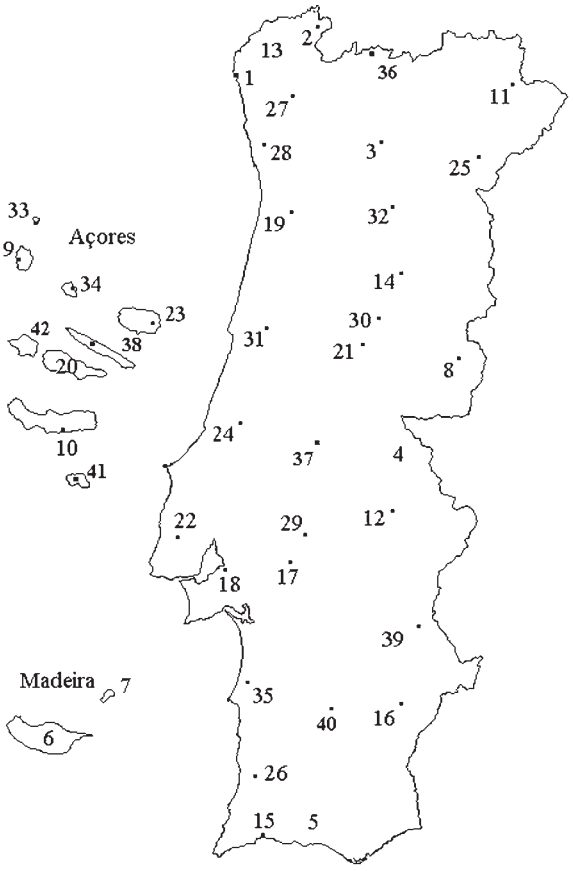
A team coordinated by Ana Maria Martins has been committed to the selection, transcription, annotation and publication of this corpus, compiled from sources such as the ALEPG, the ALLP, the ALEAç and Segura (1987). The corpus amounts to 600,000 words, collected from 42 locations within continental Portugal and the archipels of Madeira and Azores.[3]

The speakers' sociological profile is fairly constant across locations. Given the sources for this corpus, informants correspond to the traditional type of informant in dialect geography: old, non-educated, rural and born and raised in place of interview.

The corpus is freely available through the internet (http://www.clul.ul.pt/english/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php), under three different formats, for the moment: (i) *verbatim* orthographic transcripts; (ii) 'normalized' orthographic transcripts; and (iii) morphologically tagged texts. In the future, CORDIAL-SIN will also be available as a syntactically annotated corpus.

Verbatim orthographic transcripts include the marking up of some syntactically relevant phonetic and morphological variants, and of generalized spoken language phenomena, such as hesitations, filled and empty pauses, repetitions, rephrased segments, false starts, truncated words, speech overlappings, unclear productions (see (3) below). 'Normalized' orthographic transcripts correspond to a simplified version of *verbatim* transcripts, automatically obtained through elimination of the marked up features of spoken language and of phonetic transcriptions. The normal-

[3] CORDIAL-SIN has been funded by FCT (*Fundação para a Ciência e a Tecnologia*), through the following projects: PRAXIS XXI/P/PLP/13046/1998; POSI/1999/PLP/33275; POCTI/LIN/46980/2002; PTDC/LIN/71559/2006.



1. CORDIAL-SIN locations (see Appendix for identification)

ized transcripts are the input for the tagging and the syntactic annotation. An example of this two-layered transcription is given below:[4]

(1) *verbatim* ortographic transcript
Eu sei que aquilo que{fp} {PH|nũ=não} é por mal, sabe? Mas quem ouve… Vem cá uma pessoa estranha, {PH|nũ=não} é, {PH|nũ=não} conhece e diz: "Ah, [AB|são] são *malcriados, os pescadores" (…).* [Vila Praia de Âncora, VPA15]

(2) 'normalized' ortographic transcript (ASCII version)
Eu sei que aquilo que não é por mal, sabe? Mas quem ouve... Vem cá uma pessoa estranha, não é, não conhece e diz: "Ah, <break> (...) </break> são malcriados, os pescadores" <break> (...) </break>.

[4] On the conventions used in *verbatim* transcripts and their relation to normalized transcripts see *Normas de Transcrição*, http://www.clul.ul.pt/english/sectores/variacao/cordialsin/manual_normas.pdf.

(3) Examples of marked-up spoken language phenomena:
{PH|nũ=não} — phonetic variant for the negation word
{CT|pa=para a} — contraction 'to+the.FEM'
{AB|xxx} — false starts, abandoned sequences
{pp} — empty pauses
{fp} — filled pauses
[underlining] — overlapping
(…) — unclear sequences (also: omitted sequence, e.g. [AB|…], in 'normalized' transcripts)

2.3. Tagging and syntactic annotation in a dialect corpus

Further development of CORDIAL-SIN endowed this dialect corpus with tagging for each word. This has been conceived as a first step towards fast and systematic access to precise morphosyntactic information, which will ultimately be achieved with syntactic annotation.

CORDIAL-SIN tagging and syntactic annotation were both made easier by automatic tools already developed and in use within other related projects. More concretely, tagging and syntactic annotation have been largely inspired by the processes and tools used by the *Penn Parsed Corpora of Historical English* —a set of corpora developed at the University of Pennsylvania by Anthony Kroch and his associates—, and also by the *Tycho Brahe Parsed Corpus of Historical Portuguese*, a corpus of Portuguese authors born from the 16th to the 19th centuries, coordinated by Charlotte Galves at the University of Campinas (Brazil).

Collaborative work with the teams developing these corpora has permitted the tuning of already available tagging and annotation tools in such a way that these could satisfactorily apply to dialectal European Portuguese and serve our purpose. Besides accelerating the tagging and annotation phases, this cooperation also ensures the ease of linguistic information retrieval, since a query tool operating on the annotation system in use is already available.

2.3.1. Tagging

The morphological tagging operation has been to a great extent facilitated by the use of an automated morphological tagger, designed for the *Tycho Brahe Corpus* of Portuguese texts (Finger 1998). After training over a sample of 30,000 hand tagged CORDIAL-SIN words, the rate of accuracy of the tagger proved to be satisfactory enough to encourage the use of its output as the basis for a hand refined and corrected tagged version of the corpus. To ensure the precise format of the tags, an additional automatic tool has been used after manual tag correction and refinement.

Thus, CORDIAL-SIN's morphologically tagged transcripts result from a three steps process involving: (i) automatic tagging by the Tycho Brahe tagger; (ii) manual tag correction and refinement using the CORDIAL-SIN's morphological annotation system; (iii) automatic verification of the corrected tags.

The format of the morphological tags and the basics of the tagset of the CORDIAL-SIN essentially stem from the system designed for the Tycho Brahe automatic tagger (Galves and Britto 1999). Tags have an internal structure consisting of an ever-present main tag, which includes part-of-speech tags (such as D, for determiner), and, in certain cases, sub-tags (for instance, F for feminine, P for plural), diacritics attaching different main tags ("+", "!") or main tags to sub-tags ("–"), and figures indicating clusters, as in the following examples:

Tag	Application	Ex.
/D	singular masculine determiner	*o/D*
/D-P	plural masculine determiner	*os/D-P*
/D-F-P	plural feminine determiner	*as/D-F-P*
/P+D-F	preposition plus singular feminine determiner contraction	*da/P+D-F*
/VB+CL	verb (infinitive) plus enclitic pronoun	*dar-lhe/VB+CL*
/VB-R-1S!CL	verb (future) plus mesoclitic pronoun	*dar-te-ei/VB-R-1S!CL*
/P31	first element of a triple prepositional cluster	*por/P31 mor/P32 de/P33*

2. Examples of CORDIAL-SIN tags

The set of sub-tags codifies inflectional information — tense/mood and person/number for verbs or gender and number for nominal categories. It also specifies in more detail some morpho-syntactic information (such as a -NEG sub-tag that identifies different negative categories, like adverbs, quantifiers, prepositions). The system also allows: (i) main tags attachment for contractions or cliticizations; and (ii) tags and figures combination for multiple words behaving as clusters.

The tags thus obtained have a structured format that straightforwardly allows for very detailed morphological information, a very appealing solution when tagging a morphologically rich language, such as European Portuguese. As a welcome result a number of possible structured tags higher than 1.000 can be obtained from the CORDIAL-SIN tagset, which reduces to c. 40 main tags plus a smaller set of 25 subtags. (For a detailed description of the entire tagset and its application, see *Manual of the CORDIAL-SIN Morphosyntactic tagging*, http://www.clul.ul.pt/sectores/variacao/cordialsin/manual_anotacao_morfologica.pdf.)

2.3.2. Syntactic annotation

CORDIAL-SIN syntactic annotation is currently under development (2008-2010, within the project DUPLEX).[5] The annotation processes and tools in use have been developed for the *Penn Parsed Corpora of Historical English* (and the same or very similar annotation system is equally used on the *Tycho Brahe* corpus).

[5] Project PTDC/LIN/71559/2006, funded by FCT (http://www.clul.ul.pt/english/sectores/variacao/projecto_duplex.php).

Adopting such a rich annotation system (developed for English corpora) for a Portuguese corpus required a cautious adaptation of the existing system to data that differ from English in many respects. The initial phase of the CORDIAL-SIN syntactic annotation process has thus been devoted to the tuning of the basic annotation system, a task which was carried out in strict collaboration with the *PennCorpora* and the *Tycho Brahe* teams. Hand annotation of a 10,000 words sample of the corpus has served to define and consolidate the main guidelines of the system so as it could apply to Portuguese spoken texts. These general guidelines resulted in the first version of the annotator's manual.

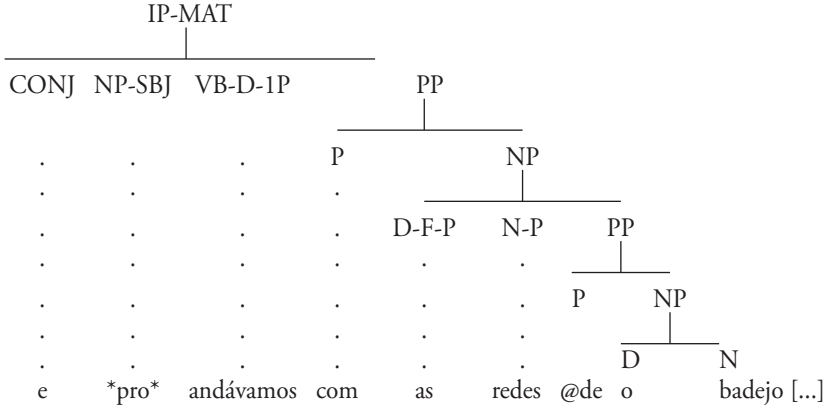
Data annotation itself is usually a very complex task. In the present case, additional complexity was expected, given the spoken and dialectal nature of the corpus. Sentences that call for detailed consideration are frequent, even though the basic lines of the system are already defined. Difficult annotations are decided upon after discussion, and each new difficult example is added to the annotator's manual, in order to assure consistency. Thus, the syntactic annotation guidelines have been progressively enriched during the whole course of the annotation phase, as more data are analysed and as new difficult sentences arise. (The current version of the *Syntactic Annotation Manual* is available at: http://www.clul.ul.pt/sectores/variacao/cordialsin/Syntactic%20annotation%20manual.html.)

The CORDIAL-SIN syntactically annotated transcripts are built on previously tagged texts, such as:

(4) e/CONJ andávamos/VB-D-1P com/P as/D-F-P redes/N-P @de/P @o/D badejo/N ,/, que/WPRO são/SR-P-3P mais/ADV-R baixas/ADJ-F-P .../. [VPA07]

The syntactic annotation produces a tree representation in the form of labeled brackets, where depth of indenting corresponds to depth of structural embedding. As for the *Penn Parsed Corpora*, the annotation represents quite flat trees, allowing for multiple branching nodes and, for some words, projecting only a word-level node (e.g. inflected verbs, negation).

(5) a.



```
b.
(IP-MAT(CONJ e)
        (NP-SBJ *pro*)
        (VB-D-1P andávamos)
        (PP     (P com)
                (NP     (D-F-P as)
                        (N-P redes)
                        (PP     (P @de)
                                (NP     (D @o)
                                        (N badejo))
                        (, ,)
                        (CP-REL         (WNP-1          (WPRO que))
                                        (IP-SUB         (NP-SBJ *T*-1)
                                                        (SR-P-3P são)
                                                        (ADJP (ADV-R mais)
                                                                (ADJ-F-P baixas))))))
        (. ...)) [VPA07]
```

In addition to constituent boundaries and phrase and clause dependencies, the annotation marks up grammatical relations, clause and sentence type, some empty categories (such as null subject and null object), among others. At the word level, morphological labels are preserved. Phrase and clause labels indicate category (NP, PP…), often specified by an extended label indicating syntactic function (e.g. subject, direct object), clause type (e.g. relative, adverbial, interrogative), or other relevant information (e.g. left dislocation, pragmatic marker).

It is worth noting that such an annotation scheme is to be seen as a fairly theory-neutral representation of constituent structure, to which category and function labels are added. The main goal of the syntactic annotation is thus to facilitate automated searches for various constructions, not to associate every sentence with an adequate structural description. Controversial decisions on annotation are avoided (for instance, by omitting undecidable information —such as the attachment of a PP as complement or as adjunt—; or by using default rules), so that the annotation scheme is completely predictable and so exploitable for automatic searches.

Turning now to the annotation process, three different stages must be mentioned: (i) a stage of automatic parsing of the data, in which the *Penn Corpora* version of a statistical parser (Collins 1999, Bikel 2004) runs over the tagged texts (at Univ. Pennsylvania); (ii) a stage of human editing of the parsed output, a time-consuming task carried out with the help of *CorpusDraw*, an editing annotation tool; (iii) finally, the result is a parsed version of the corpus in such a format that allows data retrieval through syntactic configurations (automated searches become possible with the aid of *CorpusSearch*, a search engine for parsed corpora). Both *Corpus Search* and *Corpus Draw* are components of *CorpusSearch*2 —a Java program that supports research in corpus linguistics, developed by Beth Randall at University of Pennsylvania (Randall 2005-2007)—. This program, which is freely available from http://corpussearch.sourceforge.net, is thus useful both for the construction of syntactically parsed corpora and for searching them.

CorpusDraw gives support to the human editing of the parser output, which may involve: changing syntactic tags, adding subcategory information, changing attachment level, adding empty categories, for instance. *CorpusSearch,* in turn, is a dedicated engine for parsed corpora permitting basic search functions that are linguistically intuitive: for instance, *(immediately)precedes, (immediately)dominates, exists, hassister,* among others.

CORDIAL-SIN is now in the intermediate stage of the annotation process: the output of the automated parser is under manual correction with the aid of *CorpusDraw*. As mentioned above, this task is also that of defining the details of CORDIAL-SIN's annotation system within the standards already operative in other corpora and readable by the automatic parser. Adapting the already defined system mainly involves finding solutions required by those grammatical aspects where Portuguese and English differ (a task shared with the Tycho Brahe team) or by other aspects that are characteristic of CORDIAL-SIN's dialectal and spoken data. Also at the level of the label set, a very small number of extended labels have been added, all of them relating to aspects particular to spoken language.

3. CORDIAL-SIN as a tool for Portuguese dialect syntax

The annotated dialect corpus CORDIAL-SIN has been —and still is— the main empirical source for a number of studies on different aspects of Portuguese dialect syntax (see http://www.clul.ul.pt/english/sectores/variacao/cordialsin/projecto_cordialsin_publicacoes.php). Some of these studies have already counted on the available tagging of this corpus. If we refer to a very simple example, investigating inflected gerunds as a dialectal feature in European Portuguese (as it has been achieved in Lobo 2008) could begin by obtaining a list of inflected gerunds in CORDIAL-SIN. Through the tagged corpus, this can easily be achieved just in a couple of seconds with any concordancing program. Concordances can thus easily operate over the relevant tags and sub-tags (here the sub-tags G for 'gerund' and -F for 'inflected'), providing very precise results in a very short span of time.[6] Given the distribution of all CORDIAL-SIN locations, this corpus may also provide insight into the geographic distribution of syntactic variants (see Carrilho and Pereira 2009). Finally, and perhaps more surprisingly, CORDIAL-SIN has revealed the type of till-then-unknown data without which a researcher could hardly prepare relevant and adequate elicitation tests. This was in fact the case of all the wide spectrum of (mostly unknown) constructions featuring expletive *ele* found in CORDIAL-SIN, which motivated a proposal for the re-evaluation of the received view about the grammatical status of this expletive in European Portuguese (Carrilho 2005).

The importance of making accessible to other researchers detailed syntactic information about dialectal data is twofold: firstly, it eases the way to have a closer look at dialectal data relevant for the study of syntax in general and it permits to know their

[6] Within the European project *Edisyn,* a Search Engine is currently under development. An experimental version of this Search Engine, which allows searches for part-of-speech tags across different corpora and databases (among others, the *SAND*, the *ASIS* and a corpus of Estonian Dialects), can already be operated over the CORDIAL-SIN tagged data.

geographical distribution; and secondly, it provides researchers with a wider range of syntactic phenomena, some of which pervasively appear in the dialects, while being almost unknown in the standard variety. In this respect, a general improvement of the empirical foundations for the study of syntax can be achieved. To the extent that the sentence-based syntactic annotation is compatible with already available tools permitting detailed searches, CORDIAL-SIN syntactic annotation will ensure fast and efficient access to massive dialectal data, capable of responding to the different demands of specific research purposes within the domain of Portuguese dialect syntax.

References

ALE: *Atlas Linguarum Europae*, Van Gorcum, Assen Maastricht.
ALEAç: *Atlas Linguístico e Etnográfico dos Açores* (J. Saramago, coord.) (*http://www.clul.ul.pt/english/sectores/variacao/projecto_aleac.php*)
ALLP: *Atlas Linguístico do Litoral Português* (G. Vitorino, coord.) (*http://www.clul.ul.pt/english/sectores/variacao/projecto_allp.php*)
ALEPG: *Atlas Linguístico-Etnográfico de Portugal e da Galiza* (J. Saramago, coord.) (*http://www.clul.ul.pt/english/sectores/variacao/projecto_alepg.php*)
ASIS: *Syntactic Atlas of Northern Italy* (*http://asis-cnr.unipd.it)*.
COSER: *Audible Corpus of Spoken Rural Spanish* (*http://www.uam.es/coser)*.
EDISYN*: European Dialect Syntax* (*http://www.meertens.knaw.nl/edisyn)*.
FRED: *Freiburg English Dialect Corpus* (*http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED)*.
Penn Parsed Corpora of Historical English (*http://www.ling.upenn.edu/hist-corpora)*.
ScanDiaSyn: Scandinavian Dialect Syntax (http://uit.no/scandiasyn).
Syntactic Atlas of the Dutch Dialects (SAND, see Barbiers et al. 2006).
Tycho Brahe Parsed Corpus of Historical Portuguese (*http://www.ime.usp.br/~tycho/corpus*).

Auckle, T., Buchstaller, I., Corrigan, K. & A. Holmberg, 2007, «Speakers can "talk the talk", but can they "walk the walk" too?: Measuring syntactic variability using different instruments.» *Sixth meeting of the UK Language Variation and Change Conference* (UKLVC6), Lancaster University, September 2007.

Barbiers, S. et al., 2006, *Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND)*, Meertens Institute, Amsterdam. *(http://www.meertens.knae.nl/sand).*

— & L. Cornips, 2002, «Introduction to Syntactic Microvariation», in S. Barbiers, L. Cornips & S. van der Kleij (eds.), *Syntactic Microvariation*, Merteens Institute Electronic Publications in Linguistics. 2. (Available at: *http://www.merteens.knaw.nl/book/synmic)*.

—, — & J. P. Kunst, 2007, «The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas», in J. C. Beal, K. C. Corrigan & H. Moisl (eds.), *Creating and digitizing language corpora. Volume 1: Synchronic databases*, Palgrave Macmillan, Hampshire, 54-90.

Bikel, D., 2004, *On the Parameter Space of Generative Lexicalized Statistical Parsing Models,* PhD diss. Univ. Pennsylvania, Philadelphia, PA. (Available at: *http://www.cis.upenn.edu/~dbikel/software.html* «Multilingual Statistical Parsing Engine»).

Bucheli, C. & E. Glaser, 2002, «The Syntactic Atlas of Swiss German Dialects: Empirical and Methodological Problems», in S. Barbiers, L. Cornips & S. van der Kleij» (eds.), *Syntactic Microvariation*, Merteens Institute Electronic Publications in Linguistics. (Available at: *http://www.merteens.knaw.nl/books/synmic/*).

Carrilho, E., 2005, *Expletive* ele *in European Portuguese Dialects*, PhD dissertation. University of Lisbon (Available at: *http://www.clul.ul.pt/equipa/ecarrilho/Carrilho2005.pdf).*

— & S. Pereira, 2006, «On the areal distribution of non-standard syntactic constructions in European Portuguese», paper presented at the *VIth Congress of Dialectology and Geolinguistics*, Univ. Maribor, Slovenia, September.
Collins, M., 1999, *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, PA.
Cornips, L., 2003, «Contact-induced Varieties, Syntactic Variation and Methodology», presented at *European Dialect Syntax ESF/SCH Explanatory Workshop*, Padova, September.
— & W. Jongenburger, 2001, «Elicitation techniques in a Dutch syntactic dialect atlas project», in: H. Broekhuis & T. van der Wouden (eds.), *Linguistics in The Netherlands* 2001, 18. John Benjamins, Amsterdam/Philadelphia.
— & C. Poletto, 2005, «On Standardizing Syntactic Elicitation Techniques (part 1)», *Lingua* 115, 939-957.
— & —, 2007, «Field linguistics meets formal research: How a microparametric view can deepen our theoretical investigation (sentential negation)», unpublished paper presented at *ICLaVE 4,* University of Cyprus, June.
Fernández-Ordoñez, I., 2009, «Dialect grammar of Spanish from the perspective of the *Audible Corpus of Spoken Rural Spanish* (or *Corpus Oral y Sonoro del Español Rural, COSER*)», *Dialectologia* 3, 23-51. (Available at: *http://www.publicacions.ub.es/revistes/dialectologia3).*
Finger, M., 1998, «Tagging a Morphologically Rich Language», in *Proceedings of the First Workshop on Text, Speech and Dialogue (TSD'98)*, Brno, Czech Republic, 39-44.
Galves, C. & H. Britto, 1999, «A construção do *Corpus Anotado do Português Histórico Tycho Brahe*: o sistema de anotação morfológica», in I. Rodrigues & P. Quaresma (eds.), *Proceedings of the IV PROPOR*, Universidade de Évora, Évora, 55-67.
Gottschalk, M. F., M. da G. Themudo Barata & J. V. Adragão, 1974, «Introdução», *Questionário Linguístico,* Instituto de Linguística, Lisboa.
Kortmann, B., 2002, «New Prospects for the Study of English Dialect Syntax: Impetus from Syntactic Theory and Language Typology», in S. Barbiers, L. Cornips & S. van der Kleij (eds.), *Syntactic Microvariation*, Merteens Institute Electronic Publications in Linguistics. (Available at: *http://www.merteens.knaw.nl/books/synmic/*.)
Kruijsen, J., 1983, «La Syntaxe dans l'*Atlas Linguarum Europae*», in C. Angelet, L. Melis, F. J. Mertens & F. Musarra (eds.) *Langue, Dialecte, Littérature. Études Romanes à la Mémoire de Hugo Plompteux,* Leuven U. P., Leuven. 213-223.
Labov, W., 1972, *Sociolinguistic Patterns*, University of Pennsylvania Press, Philadelphia.
—, 1996, «When Intuitions Fail», *Chicago Linguistics Society* 32. Parasession on Theory and Data in Linguistics, 76-106.
Lehmann, C., 2004, «Data in Linguistics», *The Linguistic Review* 21, 175-210.
Lehmann, W. P., 1980, «Dialect Investigations as Basis for Syntactic Study», in J. Kruijsen (ed.), *Liber Amicorum Weijnen*, AFA, Assen. 379-384.
Leite de Vasconcellos, J., 1901, *Esquisse d'une Dialectologie Portugaise*, Centro de Linguística da Universidade de Lisboa/Instituto Nacional de Investigaçã Científica, 3rd edition, 1987.
Lobo, M., 2008, «Variação morfo-sintáctica em dialectos do Português europeu: o gerúndio flexionado», *Diacrítica, Ciências da Linguagem*, Revista da Universidade do Minho, Braga, 22.1, 25-55.
Milroy, J., 1981, *Regional Accents of English: Belfast*, Blackstaff.
Milroy, L., 1987, *Observing and Analysing Natural Language: A Critical Account of Sociolinguistic Method*, Basil Blackwell, Oxford.
Randall, B., 2005-2007, CorpusSearch2 *(http://corpussearch.sourceforge.net).*
Remacle, L., 1952-1960, *Syntaxe du Parler Wallon de la Gleize*, Société d'Édition «Les Belles Lettres», Paris. vol. 1 (1952), vol. 2 (1956), vol. 3 (1960).

Schütze, C. T., 1996, *The Empirical Base of Linguistics: Grammatical Judgments and Linguistic Methodology*, University of Chicago Press, Chicago.
Segura, M. L., 1987, *A Fronteira Dialectal do Barlavento do Algarve*, Doctoral diss. CLUL.

Appendix: List of CORDIAL-SIN locations

01 VPA Vila Praia de Âncora (Viana do Castelo)
02 CTL Castro Laboreiro (Viana do Castelo)
03 PFT Perafita (Vila Real)
04 AAL Castelo de Vide, Porto da Espada, S. Salvador de Aramenha, Sapeira, Alpalhão, Nisa (Portalegre)
05 PAL Porches, Alte (Faro)
06 CLC Câmara de Lobos, Caniçal (Funchal)
07 PST Camacha, Tanque (Funchal)
08 MST Monsanto (Castelo Branco)
09 FLF Fajãzinha (Horta)
10 MIG Ponta Garça (Ponta Delgada)
11 OUT Outeiro (Bragança)
12 CVB Cabeço de Vide (Portalegre)
13 MIN Arcos de Valdevez, Bade, São Lourenço da Montaria (Viana do Castelo)
14 FIG Figueiró da Serra (Guarda)
15 ALV Alvor (Faro)
16 SRP Serpa (Beja)
17 LVR Lavre (Évora)
18 ALC Alcochete (Setúbal)
19 COV Covo (Aveiro)
20 PIC Bandeiras, Cais do Pico (Horta)
21 PVC Porto de Vacas (Coimbra)
22 EXB Enxara do Bispo (Lisboa)
23 TRC Fontinhas (Angra do Heroísmo)
24 MTM Moita do Martinho (Leiria)
25 LAR Larinho (Bragança)
26 LUZ Luzianes (Beja)
27 FIS Fiscal (Braga)
28 GIA Gião (Porto)
29 STJ Santa Justa (Santarém)
30 UNS Unhais da Serra (Castelo Branco)
31 VPC Vila Pouca do Campo (Coimbra)
32 GRJ Granjal (Viseu)
33 CRV Corvo (Horta)
34 GRC Graciosa (Angra do Heroísmo)
35 MLD Melides (Setúbal)
36 STA Santo André (Vila Real)
37 MTV Montalvo (Santarém)
38 CLH Calheta (Angra do Heroísmo)
39 CPT Carrapatelo (Évora)
40 ALJ Aljustrel (Beja)
41 STE Santo Espírito (Ponta Delgada)
42 CDR Cedros (Horta)