The *Corpus of Galician/Spanish Bilingual Speech* of the University of Vigo: Codes tagging and automatic annotation

Xoán Paulo Rodríguez-Yáñez
Universidade de Vigo

Facultade de Filoloxía e Traducción
Universidade de Vigo
Campus As Lagoas-Marcosende, s/n
E-36200 - Vigo, Spain
xoanp@uvigo.es

Håkan Casares-Berg
Seminario de Sociolingüística da RAG

Seminario de Sociolingüística
Real Academia Galega
r/ Hórreo, 31
E-15701 - Santiago de Compostela, Spain
hakancasares@hotmail.com

Abstract

Firstly, we present a brief explanation of this research project, the *Corpus of Galician/Spanish Bilingual Speech* (*Corpus de Fala Bilingüe Galego/Castelán*, abbreviated as *CoFaBil*), currently being complied at the University of Vigo. This ethnographic-conversational based corpus has been recorded in a wide range of informal and spontaneous communicative situations, subsequently transcribed in detail with those conventions normally applied to conversation analysis. Secondly, we explain the manual annotation process of the corpus. The CHAT annotation system, applied in tagging this corpus, requires specifying the linguistic-communicative code to which each word belongs. So, we shall explain the problems to which this word by word tagging leads us. These problems cover phenomena characteristic of both bilingual conversation and languages in contact, but with the specificity that the scarce interlinguistic distance between the varieties of Galician and of Spanish call for adopting certain tagging values (presented in the text) that respond to the complex nature of the different phenomena detected. Thirdly, we present the solutions conceived for the automatic annotation of this corpus. The most important result is the computer application *Anotador 1.0*, which makes it possible to note down a substantial part of the phenomena appearing in the *CoFaBil* more speedily, while doing away with the interpretative biases involved in human annotating. Also, due to the versatility of this tool, it may be used as a corpora annotator of bilingual speech for any pair of languages.

Key words: bilingual corpus, Galician/Spanish, tagging, bilingual conversation, languages in contact.

Resumo

En primeiro lugar presentaremos brevemente o proxecto de investigación en curso *Corpus de fala bilingüe galego/castelán* (abreviadamente, *CoFaBil*) que estamos formando na Universidade de Vigo. Trátase dun corpus de base etnográfico-conversacional, gravado nunha ampla gama de situacións comunicativas informais e espontáneas, e transcrito en detalle aplicando convencións usuais na análise da conversa. En segundo lugar, explicaremos o proceso da súa etiquetaxe manual. O sistema de anotación CHAT, aplicado na etiquetaxe deste corpus, obriga a especificar para cada palabra a súa pertenza a un ou outro código lingüístico-comunicativo. Así, imos expor os problemas ós que nos conduce esta etiquetaxe palabra por palabra. Estes problemas abranguen os fenómenos característicos tanto da conversa bilingüe coma do contacto de linguas, pero coa especificidade de que a escasa distancia interlingüística entre as variedades de galego e castelán demanda a adopción duns valores de etiquetaxe (presentados no texto) que respondan á complexa natureza dos distintos fenómenos detectados. En terceiro lugar, presentaremos as solucións concibidas para a anotación automática do corpus. O resultado máis importante é a aplicación informática *Anotador 1.0*, que permite anotar unha parte importante dos fenómenos que aparecen no *CoFaBil* con maior rapidez, á vez que elimina os sesgos interpretativos da anotación humana. Ademais, dada a súa versatilidade, esta ferramenta podería empregarse como anotador de corpora de fala bilingüe de calquera par de linguas.

Palabras clave: corpus bilingüe, galego/castelán, etiquetaxe, conversa bilingüe, contacto de linguas.

1. Presentation of the *Corpus of Galician/Spanish Bilingual Speech*, (*Corpus de Fala Bilingüe Galego/Castelán*, *CoFaBil*)[1]

1.1. Conceptualization

Despite the fact that in a considerable part of linguistic tradition bilingual communicative competence is identified with the *bilinguisme des hommes cultivés* (Pohl, 1965), i.e., in practice a *double monolingualism* (cf. Rodríguez Yáñez, 1993:

[1] This research Project, headed by Xoán Paulo Rodríguez Yáñez, was subsidized by the Autonomous Government of Galicia (XUGA30202A98) and by the Vice-Rectorship for Research at the University of Vigo (1998, 1999 and 2000 calls). In its current stage, and as an integral part of a broader-based project, headed by Xavier Gómez Guinovart, it is subsidized by the Autonomous Government of Galicia (PGIDT01PXI30203PR) and by the Ministry of Science and Technology (BFF2002-01385). Participating researchers in the same are Virginia Acuña Ferreira, Sonia Álvarez López, Ana Ameal Guerra, Håkan Casares Berg, Eloi Gestido de la Torre, Anxo M. Lorenzo Suárez, Fernando Ramallo, Xoán Paulo Rodríguez Yáñez and Miguel Valverde Juncal.

252-54), with a formal separation of the two languages in discourse practices, and therefore, a quirk of the cultured class, the extensive bibliography published over the past thirty years shows us that real bilingual practice, the bilingual speech of the majority of speakers in most communicative situations, involves important pragmatic-conversational characteristics represented, as a whole, by what we may call *bilingual conversational style* (Rodríguez Yáñez, 1995, 1997). This style is extremely active in the colloquial varieties in Galicia, and therefore, in the informal interactions between its speakers. It is based on code-switching inserted in the processes of conversation construction of the participants' identities (and, where applicable, in the processes where the choice of code is negotiated). Bilingual conversational style is then, a significant part of the communicative repertoire of many speakers with bilingual competence, for whom taking a position at one point or another of the structural continuum has a stylistic and/or identitarian value.

Bilingual communicative competence also entails this verbal repertoire including grammatically mixed (i.e., with code-mixing phenomena) and diachronically interfered varieties (interferences and loans).

It is in this context of theoretical differentiation between at least these three types of phenomena (conversational, grammatical and diachronic-structural) where the conception of this corpus lies (see also Acuña et al., 2001; Rodríguez Yáñez et al., 2001: 2).

1.2. Objectives

The general objective is to build a *Corpus of Galician/Spanish bilingual speech* (*CoFaBil*) for use in research[2].

We understand corpus of *bilingual Galician/Spanish speech* as that comprising representative samples of the continuum of colloquial discursive varieties in Galicia. These range from the strictly monolingual varieties in Galician to other strictly monolingual varieties in Spanish, covering the entire range of conversational and/or structurally intermediate varieties.

One of the objectives stemming from building this corpus is to make available a broad bilingual Galician/Spanish database to be computerized, annotated and digitalized in its audio part (and, where possible, in the video part), for research both into diachronic-structural phenomena (e.g., the study of loans from Spanish in current Galician varieties and their distribution according to age, sex, and rural/urban habitat, etc.) as pragmatic-conversational phenomena (e.g., the study of

[2] Currently, the *CoFaBil* is being subjected to a partial research exploitation, as in the following works: Acuña Ferreira (2002) on discursive genres in women and men interaction; and in works on colloquial syntax headed by José María García-Miguel Gallego at this University of Vigo. Also, researchs on bilingual conversation analysis and etnography of communication are being maked.

the stylistic patterns of code-switching in oral narratives) or grammatical phenomena (the hypothetical restrictions of the grammar of Galician-Spanish bilingualism). It will be possible to study the processes of formal restructuring and of linguistic change derived from languages contact, and to research into any of the levels of urban varieties of modern Galician, the speech of rural immigrants living in towns such as Vigo or A Coruña, the varieties of neo-speakers of Galician, etc.

Furthermore, application of the CHAT transcription conventions, contained in the LIDES system (*Language Interaction Data Exchange System*), thought up by the *LIPPS Group* (*Language Interaction in Plurilingual and Plurilectal Speakers*), and grouped in the *LIDES coding manual* (2000), will allow us to integrate our corpus into the international database of this research group where corpus are currently being built on bilingual situations in various European countries, and deal with one of so many questions pending: the comparability of the phenomena found in each of the situations, and the specific features of each of them.

1.3. Methodology

Sound material was obtained from 1992 to 2001, primarily through participant observation with a hidden microphone. The viewpoint adopted is ethnographic-conversational. Following these lines, recordings cover a large number of diverse communicative situations, the immense majority being of an informal and spontaneous nature: a high proportion of the recordings were made at the homes of the participants (chats over coffee, family meals, etc.) besides conversations between neighbours, housewives, students, returned emigrants, interactions between infants and their carers, interactions in groups of friends (male, female and mixed), with strangers in the street, telephone conversations and also in all types of public settings: urban and village markets, groceries, department stores, chemist's shops, cafeterias, bars, hairdresser's, etc. The data of these communicative situations, and those concerning the external variables of the participants (age, sex, place of birth, place of residence, socio-economic class, educational level, job, sociolinguistic history, number of participants, relationship between them, etc.) are incorporated into a specific database in order to solve gradually the shortages observed by incorporating new materials.

There are currently about 250 hours recorded, of which around a fifth have been conversationally transcribed. The transcription conventions used are a version commonly used by conversation analysts, including a highly detailed transcription of the facts linked to the turn-taking system, and the most relevant prosodic phenomena. Basically, we have adopted those conventions proposed by Álvarez Cáccamo, 1990 (generally, see *inter alia*, Hutchby & Wooffitt, 1998). The speakers' codes are identified by three different types of letter: round for Galician, bold for Spanish and italic for the formally ambivalent segments.

2. Bilingual speech and languages in contact: Problems of identification and manual code-tagging

The conversational corpus as it is established is subject to tagging (largely performed manually), using the LIDES-CHAT conventions (2000) referred to above. In this regard, our difficulties focus on the one hand on considering the theoretical problems underlined through the tagging word for word, as we are compelled to do by the CHAT system (in such a manner that each word must be ascribed to a code-language) and, on the other hand, on evaluating the possibility of creating an automatic tagging program for the codes-languages taking into account the solutions given to the preceding problem. At this point, we shall deal with the first of these two problems.

As stated above, the nature of the phenomena characterizing this corpus is widely varied: diachronic-structural (interferences and loans), conversational (code-switching) or grammatical (code-mixing). These three basic types of phenomena may affect, simultaneously or in combination, the same given speech segment: in summary, this entails three types of facts which are relatively independent of each other, or expressed more precisely, that may be dealt with and analyzed independently of each other. Whatever it may be, they all complicate the task of identifying and tagging codes-languages in presence.

In synthesis, we may group the segments into five types to identify the different codes, annotated using the LIDES and CHAT conventions mentioned above, and that we term (in a simple yet effective manner) as follows.

2.1. Tag 1 (*Galician* code)

Segments-words formally identifiable as *Galician* code are included at this category. This tag is written with a @1 placed after each word, e.g.:

(1) eu@1 fixen@1 o@0 axeitado@1
(I did what was right)

As we shall see, an important part of the forms of "Galician code" will, in fact, be under tags @0 and @3, i.e., "Galician" and "Spanish" tags are imperatives of the system itself but not facts which in speech would necessarily be real. As we can see, there is no reason for communicative codes to coincide with the *language* tags understood as monolingual, discrete systems[3]. This is a central theoretical problem that imbues the entire *CoFaBil*.

[3] It lies beyond the objectives of this work to deal with the *a priori* nature of the notion of *language* (or of *code-language*, as we are using the term) applied to the analysis of bilingual speech. Recovery of Roman Jakobson's notion of *code* as *communicative code* (see Álvarez Cáccamo, 1998, 2000) applied to this type of problems opens up a far more flexible theoretical framework. Regarding the inappropriateness of continuing to apply the notion of discreteness of linguistic systems in the study of

2.2. Tag 2 (*Spanish* code)

Under tag 2 (@2) we include segments-words formally identifiable as *Spanish* code.

(2) yo@2 hice@2 lo@2 correcto@0
(I did the right thing)
(3) ella@2 lo@2 hizo@2
(she did it)

which we can compare with example (1), and with (4):

(4) ela@1 fíxoo@1
(she did it)

Nevertheless, and in a similar manner to the case of the *Galician* code, one part of the forms attributable to the Spanish code are to be included under the tags explained below.

The majority of these problems are, in fact, localized under tags 0 (@0), 3 (@3) and 4 ([$4]), under which a substantial proportion of the corpus is to be covered.

2.3. Tag 0 (formally equivalent code)

Formally equivalent segments-words in the two preceding codes are tagged as 0 (@0). These forms are not ascribable to one single code. It must be noted that this category is the result of formally coinciding codes and not, as we shall see, due to loan phenomena or interference, nor to code-switching potential cases (included in tags 3 and/or 4).

(5) dáme@0 un@0 botón@0
(give me a button)
(6) era@0 carlos@0
(it was Carlos)

bilingual speech see for example, Gardner-Chloros (1995) or Gafaranga (2000). We may state, therefore, that in our conversational corpus neither "Spanish" nor "Galician" are languages *per se*, but are rather easily handled labels deriving from the traditional linguistic construct of *language* and that, in discursive practices, are expressed in a non-determinable *a priori* series of communicative codes, which include switching varieties. So, our use of the term *code-language* only obeys an imperative of the tagging system, and the same may be said of our endeavour to identify the addressee segments of each tag proposed: we are not identifying communicative codes (only subsequent, detailed conversational analysis can take on such a task), but rather demarcating some of the external (formal) phenomenon that may be related to their functioning. These problems as a whole have been clearly synthesized by Romaine (1995: 1) where his well-known book *Bilingualism* starts by stating, "It would certainly be odd to encounter a book with the title *Monolingualism*. However, it is precisely a monolingual perspective which modern linguistic theory takes as its starting point in dealing with basic analytical problems such as the construction of grammars and the nature of competence".

Any of these pieces (generally, cases of homophone diamorphs) is morphosyntactically[4] identical in the two codes[5].

An important feature of this category is that it serves to tag word for word without being subjected to an interpretative work in terms of the code that the speaker may be using on each occasion (although it is doubtful that such a question should make sense in all cases). In other words, we do not depend of the discursive context on assigning any of the two codes, and this is a considerable advantage from the viewpoint of the operational capacity of tagging and, more particularly, for automatic annotation. In this manner, the form "casa" ("house") will always be casa@0, regardless of the context in which it appears:

(7) mercou@1 unha@1 casa@0
(he bought a house)
compró@2 una@2 casa@0
(he bought a house)

One of the objectives of the code tagging process is to reduce the annotator's margin of interpretability, be it a human or automatic annotator. If we did not adopt this approach, in a fair number of cases the transcriber would take decisions that would introduce interpretative biases and would distort the coded information, making automatic annotation impossible. This would occur in cases, among others, of hesitant interventions where the speaker produces a succession of statements that may (or may not) entail code-switching points (i.e., change of code with a pragmatic value): the analysis of such a possibility should be left for the research stage and may not be solved at the current stage of corpus building and tagging:

(8) 1 A: penso@1 que::@0
(I think that)
2 <2,5>
3 tIña:::s:@1
(you had)
4 ... u::n::@0
(an)
5 <4>
6 un::@0 compromIso@0
(an appointment)

[4] Formal identity at this level should also be confirmed at the phonetic-phonological and prosodic levels. This would call for a phonetic transcription of all these segments, although we have dispensed with this, for the simple reason of the operational capacity of the transcription. We are aware, however, of the fact that, where applicable, the analyst should not neglect this problem.

[5] To the effects of spelling, where Galician and Spanish spellings do not fully coincide, in the transcription we adopt the first of these by default: such is the case of "dáme" (accented) –as opposed to "dame"–, "harmonía" –as opposed to "armonía"–, "gravación" –as opposed to "grabación"–, and so on.

7 algo@0 así@0
(or something like that)

In this manner, tag @0 does not mean that there has been no possible code-switching in any of the intonational sub-units (shown here as differentiated transcription lines) in which this tag appears, but rather that the identification of such a possibility may not be solved by means of a mere lexical and/or morphosyntactic identification.

2.4. Tag 3 (excludable or problematical code)

In tag 3 (@3) we include phenomena both of a diachronical-structural (loans, interferences) and of a pragmatic-conversational nature (i.e., cases where a unit-word potentially has a pragmatic value as switching). Homophone diamorphs are naturally excluded.

We also include here forms that are potentially dialectal varieties of Galician, but which may be interferences from Spanish (and not dialectal forms as such, these being included under tag @1). In any case these would be problems to be solved *a posteriori*. In fact, as a general rule, under this tag many of the cases and phenomena that would be of great interest in research are found. For this reason we do not rule out establishing sub-tags within this category as we streamline the process.

Among these cases we have forms such as:

(9) saiu@1 / saliu@3
(he/she went out)

In the following example, the first three forms are patrimonial variants, whereas the cuchilo@3 form is widespread in the Galician varieties of certain urban areas, plausibly in the more diachronically castillianized areas:

(10) coitelo@1 / cuitelo@1 / cutelo@1 / cuchilo@3 [kuʧilo]
(knife)
(Cf. with Spanish: cuchillo@2 [kuʧiλo])

In the following case we have a voiceless velar fricative segment [x] in “tarjeta” (“card”) instead of the corresponding Galician voiceless pre-palatal fricative [ʃ] (illustrated by the spelling <x>, ‘tarxeta’), a case affecting a large number of words within the varieties of Galician and where, for the moment, we detect an interference.

(11) e@1 largoume@1 unha@1 tarjeta@3
(and he/she showed me a card)

This same word “tarjeta”, however, in the Spanish context would be tagged as Spanish (@2) and not as a type 3 case. In fact the massive incorporation of interferences and loans has operated diachronically particularly in the direction Spanish ➛ Galician, those operating in the opposite direction being comparatively far less abundant, especially in such segments as [x]~[ʃ].

(12) y@2 me@0 largó@2 una@2 tarjeta@2
(and he/she showed me a card)

The case of the *gheada*, a phenomenon particular to certain varieties of Galician (involving making a voiced velar occlusive like an aspirate, normally pharyngeal, and in any position) offers interesting variants:

(13) lechugha@3

(14) Cf.: leituga@1 / leitugha@1 / lechuga@2
(lettuce)

(15) merquei@1 unha@1 lechugha@3
(I bought a lettuce)

(16) no@2 me@0 gusta@0 la@2 lechugha@3
(I don´t like lettuce)

(17) ¿te@0 gusta@0 la@2 leitughiña@1?
(do you like little lettuce?)

In the following case, the word “colo” (lap; cf. Spanish “brazos“ or “regazo“) is almost the only form in Spanish used by many Galicians, from which it may be concluded that, from the conversational point of view, it is an unmarked form within the Galician speech community. But since from a diachronical point of view it is a loan from Galician, this criteria suffices to include it as tag 3 (colo@3). Conversely when this same word appears in a Galician context it is simply marked as colo@1.

(18) y@2 lo@2 cogió@2 en@0 el@2 colo@3

(19) e@1 colleuno@1 no@1 colo@1

(20) y@2 lo@2 cogió@2 en@0 brazos@0

(21) e@1 colleuno@1 en@0 brazos@0
(and he/she took him/her in his/her lap)

The same occurs with words such as “dios” (“God”), diachronically a Castilian form but a conversational unmarked form in the Galician colloquial varieties (cf. with “deus”, the standard form in the Galician formal varieties):

(22) nin@1 pa@0 dios@3

(23) ni@2 pa@0 dios@2
(no talk about it)
(Cf. with: nin@1 pa@0 zeus@0)
(no talk about it)

We see, therefore, that the same form may be marked with two different tags throughout in the corpus. This means a problem when it comes to automatically recovering all the contexts of occurrence of each of the particular forms for study.

As we shall see later, including non-deterministic cases, i.e., cases that can be annotated depending on the context in which they appear, makes it necessary to define this contextual criterion for the automatic annotation. Even so this type of phenomenon is barely susceptible to be given an automated treatment, at least if no lexical information is provided or until the lists of words and of operations carried out by the tagger have been enriched through a complex human feedback process.

This category, like the preceding ones, is only applied word for word, i.e., we are not tagging the problems and phenomena that may go over the level of the word (which, as we shall see, are to be included under tag 4). In this manner, cases such as the following (a word to the effects of spelling) are also to be tagged as 3:

(24) empujoume@3
3rd person sing. past + pronoun object 1st person sing.
(he/she pushed me)

since despite pronominal post-positioning in Galician, we have a segment with a velar fricative [x], and not the alternative forms:

(25) empuxoume@1
(26) empurroume@1
(he/she pushed me)

2.5. Tag 4 (larger fragments than a word)

This tag entails, like the previous one, a set of enormously productive problems for research. Unlike the previous tags, here it is a matter of tagging segments larger than a word. In effect, a merely individualized tagging, word for word, would not take into account a host of problems regarding bilingual grammar (code-mixing) or the pragmatics of bilingual speech (code-switching), phenomena that in most cases go over the spelling frontier of segments-words and that may turn out to be independent of the tagging of each of the words forming the fragment.

Thus apart from the tag received for each word (in line with the categories explained above) we shall use tag 4 to note that in the relationship between a group of words some type of greater phenomenon occurs. The affected segment is noted by a *main level code type* tag, with the syntax:

<word@1 word@0> [$4]

as occurs in the following examples:

(27) <o@1 origen@3> [$4]
(28) <o@1 orixe@1> [$4]
(Cf. castellano: el@2 origen@2)
(the origin)
(29) <a@1 leite@1> [$4]
(Cf. Spanish: la@2 leche@2)
(the milk)

where we are dealing with a morphosyntactical problem (article and noun concordance) inside the syntagm, as opposed to the feminine concordance for "orixe" (a@1 origen@3) and the masculine concordance for "leite" (o@1 leite@1) in the Galician varieties.

A considerable part of the casuistic for this category, however, is of a more complex nature, covering cases of the pre-verbal position of the pronoun in assertive statements, relatively frequent cases in the urban varieties of Galician and among neo-speakers, although ungrammatical in the rest of Galician varieties:

(30) <me@0 empujou@3> [$4]
(31) <me@0 empuxou@1> [$4]
(32) <me@0 empurrou@1> [$4]
(he/she pushed me)

which we confront with the cases above (25) and (26), and with:

(33) me@0 empujó@2
(he/she pushed me)

In this manner, we come across possibilities such as:

(34) <a@0 mí@2 me@0 gusta@0 máis@1> [$4]
(35) <a@0 mí@2 gústame@1 máis@1> [$4]
(36) <a@0 min@1 me@0 gusta@0 máis@1> [$4]
(I like it better)

which we confront with*:*

(37) a@0 min@1 gústame@1 máis@1
(38) a@0 mí@2 me@0 gusta@0 más@2
(I like it better)

This label is also applicable in the case of discontinuous connectors:

(39) <o@2 seña@1> [$4]
(40) <ou@1 sea@3> [$4]
(that is, thus)

which we confront with*:*

(41) / (42) ou@1 sexa@1 / ou@1 seña@1
(43) o@2 sea@2
(that is, thus)

Finally it is also applicable specially in countless cases of complex segments which pose problems of Galician/Spanish mixing and/or switching:

(44) <mirar@0 a@1 cor@1 de@0 la@2 piel@2 influye@3 moitísimo@1> [$4]
(look, the colour of your skin matters a lot)

which we can contrast with the structures:

(45) mirar@0 a@1 cor@1 da@1 pel@1 inflúe@1 moitísimo@1
(46) mirar@0 el@2 color@0 de@0 la@2 piel@2 influye@2 muchísimo@2
(look, the colour of your skin matters a lot)

or with the case (47):

(47) <ana@0 no@2 te@0 trajo@2 o@1 que@0 lle@1 pediches@1> [$4]
(Ana didn´t bring you what you asked her for)

which we compare with (48) y (49):

(48) ana@0 non@1 che@1 trouxo@1 o@1 que@0 lle@1 pediches@1
(49) ana@0 no@2 te@0 trajo@2 lo@2 que@0 le@2 pediste@2
(Ana didn´t bring you what you asked her for)

or, among so many other possible discursive phenomena, the case of numerical computing sequences, common at fairs, markets and shops, where code alternation are linked to rhythmic patterns (Rodríguez Yáñez, 1995: 204-219):

(50) <un@1 dous@1 tres@0 catro@1 cinco@0 seis@0 siete@2 ocho@2
nueve@2 e@1 dez@1> [$4]
(one two three four five six seven eight nine and ten)

which we confront with:

(51) un@1 dous@1 tres@0 catro@1 cinco@0 seis@0 sete@1 oito@1 nove@1 e@1 dez@1
(52) uno@2 dos@2 tres@0 cuatro@2 cinco@0 seis@0 siete@2 ocho@2
nueve@2 y@2 diez@2
(one two three four five six seven eight nine and ten)

Annotation (and where applicable, sub-annotation) of these complex cases will ease systematic searching at the research exploitation stage of code-mixing and code-switching phenomena. However a systemized treatment of structures larger than the word entails greater difficulty which, despite the fact that computationally it is not specially complex (in the majority of cases), it does call for requirements that can not be solved by the annotator presented in the following section.

3. The creative process of the computer application *Anotador 1.0*

3.1. General comments and basic annotation

Except for tag 4, the categories proposed in the previous section actually correspond to the spelling unit "word". In order to annotate the texts of the corpus automatically, it appears sufficient, therefore, to have a categorical tagger or *part-of-speech tagger* (POS), otherwise known as grammatical annotator, which is in reality the most common form of tagging a corpus, although in this case it will serve for widely differing ends).

The annotation process would involve three prototypical stages. The first would be the pre-editing stage, the second would be assigned to attributing the tagging itself, and a third stage, less desirable although nonetheless essential, where post-editing tasks would be designed.

Pre-editing simply converts the input, i.e., the non-tagged text, into an appropriate format for the annotating programme itself. In our case, the options available are as follows. There is a large amount of conversationally transcribed text in the *CoFaBil* (about 45 hours of recording). If what we are looking for is to annotate the units "word" in reference to the code to which they have to be ascribed (which is the objective of the annotator presented), we should save the document with a *.txt* extension. To do so, the file should be opened with a *.doc* format supporting editor (transcriptions in the *CoFaBil* were made with the popular Microsoft Word text processor) and save as *text only* (despite the fact that the format will be lost and, along with it, encoded information). The files transcribed in the CLAN editor with CHAT tagging may be opened directly from the annotation programme after changing their extension, from *.cha* to *.txt*, for viewing in the writing box.

In the second stage, the programme conducts a word for word analysis and assigns each one the appropriate tag. The simplest version and the easiest to implement is one which distinguishes between only three types of tag. As we saw through section 2, one would annotate forms mistakenly ascribed to the category "Spanish", another the category "Galician", and a third for the "formally equivalent codes", which would comprise the forms excludable from the previous categories for being formally coincident in both codes. The unit word is defined in the programme as "a chain of determined characters". The programme recognizes that it is at the start of a word when it finds a defined character, and recognizes that it is at the end of the same when a non-defined character appears. As noted, the procedure may be somewhat rudimentary but it is nonetheless effective. The pseudo-code would be something on the lines of:

```
        Search (for something starting with
        [ABCDEFGHIJKLMNÑOPQRSTUVWXYZÁÉÍÓÚabcdefghijklmnñopqr
stuvwxyzáéíóúü*@<>&%#/$|¬]
        or that does not start with
        [^ABCDEFGHIJKLMNÑOPQRSTUVWXYZÁÉÍÓÚÜabcdefghijklmnño
pqrstuvwxyzáéíóúü*@<>&%#/$|¬]
        the first is called “word”, and the second “nothing”) end.
```

Once the programme has acknowledged that it is a word, in the ascribing part of the tags the programme uses two main reference word lists. These are obtained from a Galician and a Spanish lexicon. The words appearing in only one of the two lexicon form the word references that the programme uses to assign a mistaken taken corresponding to one of the excluding categories (“Spanish” and “Galician”). The words appearing in both are the words used to recognize those presenting “formal equivalence between codes”.

As can be speedily deduced, obtaining two rich lexicons is vital for the programme to function. The Anotador 1.0 (Casares Berg, 2002a, 2002b)[6] is distributed with a flexed word list of Galician and another of Spanish.

3.2. Annotating forms with variations

One manner of perfecting the programme would be to introduce a purely orthographic de-ambiguation module. Homophonic words are found in Galician and in Spanish which, with the same meaning, have different spellings, mainly arising from the fact that a same phoneme has two different representations in the written language or that certain letters do not represent any phoneme. Such are the cases of words like /mara’ßiλa/ (“wonder”), written in Galician with a <b> and in Spanish with a <v>, or the case of /armon’ia/ (“harmony”) written with <h> in Galician and without in Spanish. In the conversational transcription conventions used in the *CoFaBil*, it is specified that the ambiguous fragments are transcribed by default with Galician spelling, but their tagging should be that of “formal equivalence between codes”. So that the automatic annotator recognizes these words it is essential to establish a third lexicon resulting from extracting the two prior lexicons from a word list which specifies these cases.

[6] Anotador 1.0 may be downloaded from the website of the *Sociolinguistics and Bilingualism Seminar* of the University of Vigo: http://www.uvigo.es/ssl/hcasares

In the programme, this was carried out by means of a function located before the procedure explained above, in order to detect the Spanish and Galician words, and which simply makes a prior reading of the list *homofonos.txt*. If the word being searched for in the open file, at the time of execution, is there, it directly considers it as "equivalent". The list of homophones may be obtained by means of a search where all the characters of a form are equal in the two lists, and that meet the conditions of having one grapheme or another in the same position (<b> and <v> in the case of marabilla/maravilla) or words with or without an <h> in the initial position (as in the case of harmonía/armonía). This would be a simple procedure to enrich a possible list of homophones. As in the previous case, Anotador 1.0 (Casares Berg, 2002a) is distributed with a short list of homophones, and this is essential despite it not forming a part of the programme itself. As in the case of the other lists (*galego.txt* and *castelán.txt*), *homofonos.txt* can be edited and expanded. In fact, we note that the efficiency of the annotator is linked to the extent and correction of these lists.

A somewhat more complex procedure must be applied, however, to extract the words in which a different spelling from the standard has been used, in order to represent allophones present in certain geographical and social varieties. These are the case of the *gheada* and of *seseo*, which are represented by the digraph <gh> in the first case, and by <s> instead of <c> or <z> in the second. In the first place, the annotator should be able to recognize the form as an alternative of a "standard", i.e., that the chain should be exactly equal, with the exception of a character or chain of characters that, by occupying the same position, appear as one of the possible alternatives. Secondly, the forms obtained may appear in only one of the excluding lists or in both, thus attributing them the relevant "Galician", "Spanish" or "formal equivalent between codes" tag. This problem is solved in a rather heterodox manner, as an internal procedure by a function of the programme and not by an external detection or comparison procedure. In the case of the *gheada*, initially the annotator substitutes the chain <gh> for <g>.

The case of *seseo* is more complex. From a computational point of view, this case is similar to that of *gheada*, with the substantial difference that instead of detecting a chain of relatively infrequent characters, such as <gh>, the annotator should carry out a systematic substitution of all the sequences <se> and <si> for <ce> and <ci>, and <sa>, <so>, <su> for <za>, <zo>, <zu>, respectively, then to run a routine scrutiny in the lists. This process, however, is slow, so that this possibility has been ruled out.

There are also homophone words in Galician and in Spanish with different meanings. This is the case of /polbo/, written with a <b> in Galician meaning "octopus" (*pulpo* in Spanish), whereas in Spanish /polbo/ with <v> means "dust" (*po* in standard Galician). It would appear to be profitable for the same person to separate out these cases in the orthographic de-ambiguation list, to insert them into either of the two lists used as a reference for the programme, in order to grant mistaken tags from one code or another and not have to configure a fourth list of words. It would also appear sufficient to trust in the capacity of the person who drew up the conversational transcription in order to interpret these homophone words contextually, and in view of their meaning to assign them one spelling or another. It is essential to remember that it is not the ambition of the Anotador 1.0 to be more *exacting* than a human transcriber, but rather it is sufficient for it to be *faster* and therefore more profitable. All cases of semantic ambiguity arising due to the inclusion of new variants or for any other reason, are beyond the scope of an annotator with this degree of simplicity, so that de-ambiguating must be relegated to a manual post-editing process.

3.3. Different fragments of the basic unit and exceptions. Tags @3 and [$4]

The nature of the data compiled by the corpus (interactions of real, spontaneous speech) introduce certain difficulties, such as that of an annotator which only considers those possibilities presented to date can not be solved. As we saw through the casuistic presented in section 2, it seems inevitable to foresee a considerable amount of occurrences which do not appear in any of the previous lists, of forms that require a different categorization in terms of the Galician-Spanish linguistic contact processes and/or of their pragmatic-conversational virtuality (which are, in turn, two very different phenomena), seen under @3, and of fragments below or above the unit word presenting specific particularities seen under [$4].

There is a significant occurrence of cases where, from the diachronic viewpoint, we identify as linguistic interference in smaller units of analysis than the basic unit which an annotator of this type is orientated to recognize. Cases of interference at the morphological level, where the components of a word are ascribable to different codes (puertiña@3, “small door”)[7] are an example in point. In view of the fact that in the *CoFaBil* it is not a priority to carry out a morphological level annotation, the need arises to create a fourth category to cover all these phenomena.

The forms susceptible for inclusion in this category would be those from a more heterogeneous background than the previous ones. With all this, the criteria for accepting forms corresponding to the different linguistic and sociolinguistic phenomena should be one of simplicity, seeking to avoid an over-complex annotation. Initially, the most intuitive procedure and the one involving the least effort to give the best results entails ascribing to this category all the forms not appearing in the reference database lists. This would solve the problem of annotating all the forms with morphological particularities, although it would also include those which for any reason are the product of a spelling mistake inserted during the

[7] Which we confront with portiña@1 and with puertecita@2, “small door“.

conversational transcription stage. Computationally, this also appears to be the best option. Following the search routine, when the programme re-writes the words with its own annotation, the words that are not in the lists are given an annotation termed “excludable or problematical” which (as we saw in section 2.4 regarding tag 3) will be of considerable use to the researcher. Post-edition may have a feedback effect on correction of forms with this tag, in the sense that it may reveal forms not included in the reference lists, but which may be included under certain criteria not taken into account when selecting or building the initial lexicons. As we have stated earlier this task does not entail any difficulty whatsoever.

In order to establish cases requiring a complex contextual interpretation due to going over the limits of the word spelling unit, categorized as tag 4 (see section 2.5), two criterion must be taken into consideration. Firstly, it is essential to demarcate the value of “context” for each of the cases, or rather one which, as a whole, may become a quantifiable phenomenon. So, the context should be defined through categories which either take into account the annotation of previous and subsequent forms, or those for which a statistical data is obtained for all the speaker’s interventions in the transcription or any other measurable procedure. Secondly, it is essential to draw up a casuistic containing all the sequences for which the context conditions its annotation, since initially this would involve ascribable forms to any of the four categories in the first stage of the programme. This casuistic must be the result both of the communicative competence of the sociolinguist him/herself as a member of the Galician/Spanish bilingual speech community, and of the data obtained from the corpus itself in its first tagged version, and in manual corrections.

The problems that this type of annotation present are different depending on each case. On the one hand, there is a theoretical lack of definition as regards the discursive context in which this type sequence is generated. We should take into account that it is in fact impossible to make a distinction, *a priori*, between all the conversational contexts able to be categorized as tag 4. The fact that this is not stated explicitly, and that it is in the end an interpretative problem the solving of which is only possible at the analysis stage of the problems contained therein, means that the phenomenon cannot be reduced to numbers, and therefore, lies beyond the scope of this annotator. Furthermore, there are other types of annotation problems that may only be solved by including grammatical information (concordance, placing of pronouns, etc.). Regrettably, as there is not grammatically annotated list of words available, this option was discarded from the outset.

In the following section we move on to explore the final results of the annotator and its operation.

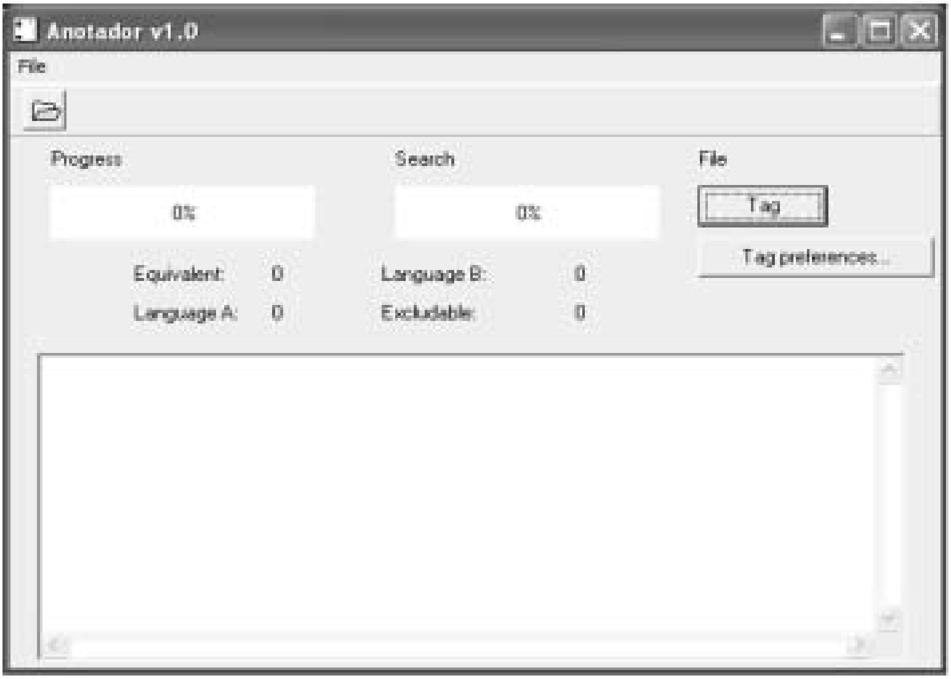
4. Handling the Anotador 1.0

When the Anotador 1.0 was programmed (Casares Berg, 2002a) it was observed that with little effort it was possible to design an annotator for bilingual conversation phenomena. The fact that it uses lists of external reference words makes it possible for them to be written in the languages required. The Anotador 1.0 may turn out to be a particularly useful work tool for researchers who work with bilingual conversation between closely related linguistic varieties, where human annotating (as we saw in section 2) is more laborious.

The Anotador 1.0 was built in C++, using the Borland 3.0 editor and compiler. A Windows application was selected, this being the most widely used operational system in the scientific community for which this application is designed. In the future, it would be interesting to compile a version for Linux and perhaps also for the Mac platform.

4.1. Main functions

Once the application is executed the main window appears (figure 1). Here we can view the file menu. Clicking on this menu, the three basic functions are shown: *abrir* (open), *gardar como* (save as) and *saír* (exit). We also have short cuts to open directly. When we open a file to annotate, we should first ensure that it is a text file. If the file that we wish to annotate is not a text file we proceed to edit it in any text editor programme.



Progress 0%

Search 0%

Equivalent: 0

Language A: 0

Language B: 0

Excludable: 0

Figure 1. Anotador 1.0 main window.

As explained above, the text writing box in the main menu shows both the text to be tagged as the result of the annotation when the process is completed. The files viewed may have a maximum size of 64 KB. In the event that we wish to annotate a large sized file, it may not be viewed on screen, although the annotation process will be carried out, as is logical. In this case, a line of text will appear in the text writing box of the main window, with the following:

****Too large text****

Despite having the option available to save the result with any name we may wish and in the folder that we select, the annotator creates a default text file, *res.txt*, in the same root directory as the file that is being annotated.

If we click on the only function of the annotator, i.e. *anotar* (tag), we can see how the operation develops with the progress bars. The first, a red bar, indicates the progress of the entire annotation process. This is useful for estimating how long the entire process is going to take. The second progress bar, which is blue, indicates the percentage of lists that the annotator requires to run through in order to find the word being searched for.

A small marker appears under the bars, showing the total frequencies of the final tags in a complete annotation process, and therefore the number of words attributable to each code and category.

4.2. Configurability

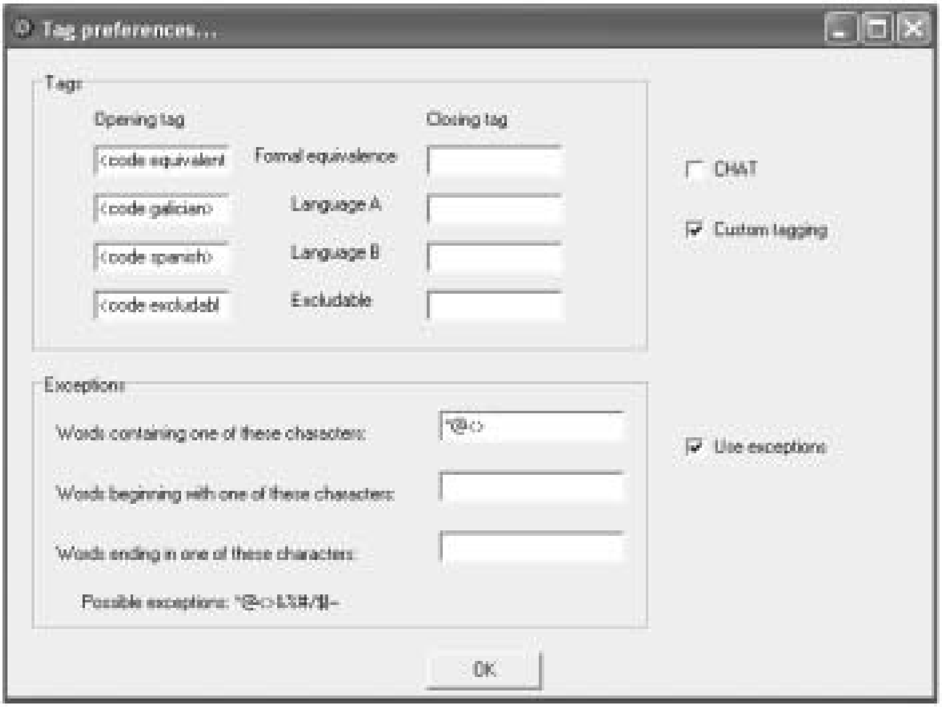
The final option available to us on the main window is to click the configuration key. This displays the programme configuration on the screen (figure 2).

As we can see here the default annotation convention for the programme is CHAT, as is logical, since it is designed for use by the research personnel of the *CoFaBil*. As we have already seen, the annotations to define a code-language in CHAT are made with a very simple annotation, involving adding a small chain of text to the word, namely an @ symbol and a number:

esto@0 é@1 un@0 exemplo@1
(this is an example)

For easy conversions or quite simply, to achieve the universality required, the user has the option to use the textual annotation type (on the condition that it comprises chains of characters placed before, after or on both sides at the same time, of the word in question) that he or she may wish. In this manner it is possible to annotate a text, for instance, with COCOA:

<equivalent code>this <Galician code>is <equivalent code>an <Galician code>example



Formal equivalence: <code equivalent
Language A: <code galician>
Language B: <code spanish>
Excludable: <code excludabl

□ CHAT
☑ Custom tagging
☑ Use exceptions

Words containing one of these characters: *@<>
Words beginning with one of these characters:
Words ending in one of these characters:
Possible exceptions: *@<>&%#/¶~

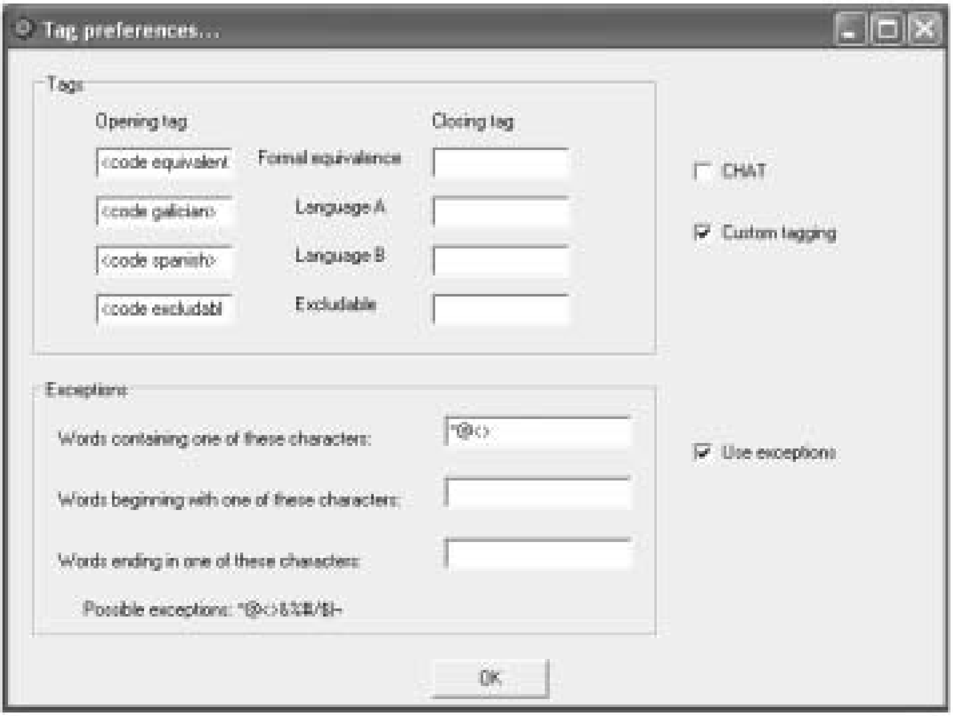
Figure 2. Configuration window.

As we are aware, in the COCOA system annotations function as delimiting agents at the start of a value act on a tag, so that a COCOA annotation keeps its value until an annotation with the same tag but a different value appears (Pérez Guerra, 1998). We will then have a way of defining the form of annotating the same reality in COCOA in the table “Tags”. This would be the appearance of a screen for annotating in COCOA (using the tag names and values from the previous example) a text transcript in bilingual speech (see figure 3).

As we can see, we place the annotations in COCOA in the “Opening tag” column that we used in the previous example for each code-language that we wish to mark. The column “Closin tag” will remain empty as a result of the very characteristics of this annotating system.

In order to annotate a text in line with the conventions of the *sgml* system, after adopting any type of standard or using own marking elements, we have to fill in the first column “Opening tag” with indications as follows:

<equivalent>
<galician>
<spanish>
<excludable>



Tag preferences...

Tags

Opening tag — Closing tag

<code equivalent — Formal equivalence —

<code galician> — Language A —

<code spanish> — Language B —

<code excludabl — Excludable —

☐ CHAT

☑ Custom tagging

Exceptions

Words containing one of these characters: *@<>

Words beginning with one of these characters:

Words ending in one of these characters:

Possible exceptions: *@<>&%#/$|~

☑ Use exceptions

Figure 3. Configuration window for annotating in COCOA.

And in the second column "Closing tag" with the same structures, including the closing final tag for the marking elements in this system:

</equivalent>
</galician>
</spanish>
</excludable>

These examples illustrate the programme's versatility for annotating in line with different conventions. The user may of course use any system of his or her own, on the condition that it must be coherent.

When a text is annotated it sometimes occurs that it is not raw, but that it already has some kind of annotation. The transcriptions which are meant to be annotated with this programme usually incorporate some type of annotation, e.g.:

```
@begin
*MOD:   non te quero.
        (I don´t love you)
*CHI:   pareces parva.
        (you seem stupid)
@end
```

If there is no type of information for the tagger to avoid annotating these chains of characters we would obtain results as the following:

```
@begin@3
*MOD@3: non@1 te@0 quero@1.
*CHI@3:   pareces@0 parva@1.
@end@3
```

Here we can apreciate that together with the transcription itself, some other parts have been tagged, e.g. the speech turn indicator (*MOD: and *CHI:) or the transcription opening and closing tags (@begin and @end) have also been "annotated", i.e. provided with a @3, as the tagger interpret them as being "excludable code".

So that the annotator can overlook the character sequences used to open and close a CHAT file or any indicating a speech turn, the table configuration window includes an "Exceptions" button. Here, the chains containing CHAT annotations are already included by default, so that sequences such as *MOD: or *CHI: indicating speech turns, or @begin and @end indicating the beginning and end of file *.cha*, can be overlooked by the annotator. With the same objective of configurability proposed for annotation, the use is able to define the chains containing any peculiar character in the table of exceptions, or that start or end with a given character.

5. Conclusions

We have presented a proposal for manually tagging *CoFaBil*, incorporating the different phenomena found regarding the codes in Galician/Spanish bilingual speech (pragmatic, grammatical and diachronic-structural phenomena) in five tags (@1, @2, @0, @3, and [$4]), in such a manner that segments-words and also sequences larger than this unit are tagged.

The advantages, however, of having an annotator able to carry out this task automatically are undeniable. The proposed Anotador 1.0 solves the phenomena detected for the tags @1, @2, @0 and partially @3. However the problems categorized under [$4] and those under @3 linked to an interpretation depending on the discursive context, go beyond the scope of an annotator of these characteristics due to their complexity. Later versions of this annotator perhaps may propose at least partial solutions to these problems pending.

As we have seen, manual tagging is directly based on a reflection on the theoretical problems of codes and their role in Galician/Spanish bilingual speech community, and not on the degree of difficulty that its implementation would involve for an automatic annotator.

The distinction between very different levels of analysis (or in other words, between substantially independent types of problems), the arguable status of codes-languages in bilingual speech, and the difficulties posed by their identification in situations of scarce interlinguistic distance (as occurs between Galician and Spanish languages), are factors that the researcher must take into account in the manual tagging of the corpus and when considering how the research itself is to be carried out. An automatic annotator able to tag the greatest possible part of the phenomena detected would be of a great help, although realistically phenomena that can only be tagged manually would lie beyond its scope.

Bibliographical references

Acuña, V., S. Alvarez, A. Ameal, H. Casares, A. Lorenzo, F. Ramallo, X.P. Rodríguez & M. Valverde (2001). "Galician/Spanish bilingual corpus: Some transcription and tagging difficulties". Paper presented at the *Third International Symposium on Bilingualism*, 18-20 April 2001, University of West England, Bristol. [Unpublished].

Acuña Ferreira, V. (2002). *Géneros discursivos en la interacción femenina y masculina: las historias de queja*. Unpublished MA Dissertation, Universidade de Vigo.

Alfonzetti, G. (1992). *Il discorso bilingüe. Italiano e dialetto a Catania*. Milan: Francoangeli.

Álvarez Cáccamo, C. (1990). *The Institutionalization of Galician: Linguistic Practices, Power, and Ideology in Public Discourse*. Unpublished PhD dissertation, University of California at Berkeley.

Alvarez Cáccamo, C. (1998). "From 'switching code' to 'codeswitching': Towards a reconceptualisation of communicative codes". In P. Auer (ed.), *Code-Switching in Conversation. Language, Interaction and Identity*. London: Routledge, 29-48.

Alvarez Cáccamo, C. (2000). "Para um modelo do 'code-switching' e a alternância de variedades como fenómenos distintos: dados do discurso galego-português/espanhol na Galiza". *Estudios de Sociolingüística* 1(1), 111-128.

Auer, P. (1984). *Bilingual Conversation*. Amsterdam: John Benjamins.

Auer, P. (ed.), (1998). *Code-Switching in Conversation. Language, Interaction and Identity*. London: Routledge.

Casares Berg, H. (2002a). "Anotador 1.0". *Seminario de Sociolingüística e Bilingüismo* web site: http://www.uvigo.es/webs/ssl/hcasares

Casares Berg, H. (2002b). *Un etiquetador automático para o* Corpus informatizado de fala bilingüe galego/castelán *da Universidade de Vigo*. *Seminario de Sociolingüística e Bilingüismo* web site: http://www.uvigo.es/webs/ssl/hcasares

Gardner-Chloros, P. (1995). "Code-switching in community, regional and national repertoires: The myth of the discreteness of linguistic systems". In L. Milroy & P. Muysken (eds.), *One speaker, two languages. Cross-disciplinary perspectives on code-switching*. Cambridge: Cambridge University Press, 68-89.

Gafaranga, J. (2000). "Language separateness: A normative framework in studies of language alternation". *Estudios de Sociolingüística* 1(2), 65-84.

Hutchby, I. & R. Wooffitt (1998). *Conversation Analysis. Principles, practices and applications*. Cambridge: Polity Press.

LIDES coding manual (2000). (= *The International Journal of Bilingualism*. Special Issue. *The LIDES coding manual: A document for preparing and analyzing language interaction data* 4,2).

MacWhinney, B. (1991). *The CHILDES Project: Computational tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Muysken, P. (2000). *Bilingual Speech. A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Payrató, L., E. Boix, M.R. Lloret & M. Lorente (coords.), (1996). *Corpus, corpora. Actes del 1r i 2n Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)*. Barcelona: PPU-Secció de Lingüística Catalana, Universitat de Barcelona.

Pérez-Guerra, J. (1998). *Análisis computerizado de textos. Una introducción a TACT*. Vigo: Universidade de Vigo. Servicio de Publicaciones.

Pohl, J. (1965). "Bilinguismes". *Revue Roumaine de Linguistique* X(4), 343-49.

Rodríguez Yáñez, X.P. (1993). "Quelques réflexions à propos de la sociolinguistique galicienne". *Plurilinguismes* 6, 225-58.

Rodríguez Yáñez, X.P. (1995). *Estratexias de comunicación nas interaccións cliente-vendedor no mercado da cidade de Lugo: as alternancias de lingua galego/castelán e a negociación da escolla de lingua*. Unpublished PhD Dissertation. Universidade da Coruña.

Rodríguez-Yáñez, X.P. (1997). "Aléas théoriques et méthodologiques dans l'étude du bilinguisme. Le cas de la Galice". In H. Boyer (ed.), *Plurilinguisme: "contact" ou "conflit" de langues?* Paris: L'Harmattan, 191-254.

Rodríguez Yáñez, X.P., A. Lorenzo Suárez, F. Ramallo, V. Acuña Ferreira, S. Alvarez López, A. Ameal Guerra, H. Casares Berg & M. Valverde Juncal (2001). "El *Corpus informatizado de fala bilingüe galego/castelán* de la Universidad de Vigo: presentación y problemas de identificación y etiquetado de los códigos gallego y castellano". In A.I. Moreno & V. Colwell (eds.), *Perspectivas Recientes sobre el Discurso. Recent Perspectives on Discourse*. AESLA (Asociación Española de Lingüística Aplicada) and Universidad de León: Secretariado de Publicaciones y Medios Audiovisuales. [CDRom edition, 13 pages]

Romaine, S. (1995). *Bilingualism*. Oxford& Cambridge: Blackwell. [2nd edition].