Transcription, Tagging and Coding of Bilingual Corpora via LIDES

M. T. Turell, M. Forcadell

Sèrie Informes, 21

Barcelona
Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
1997
Direcció de les Publicacions de l'IULA: M. Teresa Cabré

Primera edició: 1997

© els autors
© Institut Universitari de Lingüística Aplicada
La Rambla, 30-32
08002 Barcelona

Dipòsit legal: B-34.229-2002
Transcription, Tagging and Coding of Bilingual Corpora via LIDES

M.Teresa Turell  Montserrat Forcadell
turell@upf.es  forcadell_montserrat@trad.upf.es

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

L'objectiu d'aquest informe és presentar l'aplicació d'una sèrie de propostes sobre transcripció, etiquetatge i codificació a dos corpus: el corpus bilingüe LC (La Canonja (Català-Espanyol)) i el corpus trilingüe CSCD (Code-switching as Communicative Design (Català-Espanyol-Anglès)). Aquestes propostes, que constitueixen l'aportació de l'equip IULA-LIPPS (Language Interaction in Plurilingual and Plurilectal Speakers) al manual de codificació del sistema LIDES (Language Interaction Database Exchange System), adoptat pel grup europeu LIPPS, poden ser útils per transcriure, etiquetar i codificar dades provinents de llengües tipològicament properes i distants.

This report presents an overview of the contributions of the IULA-LIPPS (Language Interaction in Plurilingual and Plurilectal Speakers) team to the LIDES (Language Interaction Database Exchange System) coding manual and their application to two corpora: the LC (La Canonja) bilingual corpus (Catalan-Spanish) and the CSCD (Code-switching as Communicative Design) trilingual corpus (Catalan-Spanish-English). These contributions to LIDES can also be exploited for transcribing, tagging and coding language interaction data from both typologically close and distant languages.

1 This research was possible thanks to the CICYT (Comisión Interministerial de Ciencia y Tecnología) grants, SEC93-0725 and SEC96-0627.
Índex

1. LIPPS and LIDES ............................................................................................................. 1
   1.1 Introduction ............................................................................................................... 1
   1.2 LIDES coding manual ............................................................................................ 1
2. Contents of the LIPPS database ........................................................................................ 2
3. The corpora ..................................................................................................................... 2
   3.1 The LC corpus ........................................................................................................ 2
   3.2 The CSCD corpus .................................................................................................... 3
4. Transcription .................................................................................................................. 3
   4.1 Database format ....................................................................................................... 3
   4.2 Tier transcription in the basic LIDES version of the database ................................. 3
   4.3 Transcription unit needed ....................................................................................... 4
      4.3.1 Definition of INTROS ...................................................................................... 5
      4.3.2 INTROS typology ............................................................................................ 5
   4.4 The IULA-LIPPS proposals for the transcription of Catalan-Spanish bilingual data .... 6
      4.4.1 Transcribing words ......................................................................................... 6
      4.4.2 Transcribing Morphemes on the main tier ....................................................... 6
5. Tagging ......................................................................................................................... 7
   5.1 Main tier ................................................................................................................ 7
      5.1.1 General conventions ....................................................................................... 7
      5.1.2 Language tagging ......................................................................................... 8
      5.1.3 Language interaction tagging ......................................................................... 8
   5.2 Dependent tiers ..................................................................................................... 9
      5.2.1 Word-related tiers ....................................................................................... 10
      5.2.2 Utterance-related tiers ................................................................................. 10
6. Coding ....................................................................................................................... 10
   6.1 Word-related tiers .............................................................................................. 10
   6.2 Utterance related tiers ...................................................................................... 11
   6.3 Turn-related tiers ............................................................................................. 13
7. References .................................................................................................................. 15
Appendices ..................................................................................................................... 16
   Appendix 1: Headers and transcription sample for file LC/26.asc (Catalan Matrix Language) ... 16
   Appendix 2: Headers and transcription sample for file LC/02.asc (Spanish Matrix Language) .. 19
   Appendix 3: Sample TURN/INTROS for file CSCD/prgaed6.asa ........................................ 22
   Appendix 4: Coded fragments for file LC/02.asc .......................................................... 24
   Appendix 5: Six-step file (prgaed6.asa) ........................................................................ 28
   Appendix 6: List of abbreviations .............................................................................. 34
1. LIPPS and LIDES

1.1 Introduction

The analysis of bilingual data, that is, corpora involving language interaction phenomena (also known as language contact phenomena) requires a system of transcription, tagging and coding which allows us to isolate, mark and analyze the phenomena at hand in such a way that enables us to understand linguistic structure better and shed some light on linguistic theory. Such system has been proposed as the **Language Interaction Data Exchange System** (LIDES), which derives from the tools included in the CHILDES system (MacWhinney 1991), used to automatize and analyse child language, but involves further adaptation in order to be able to transcribe, tag and code bilingual corpora and develop new computational tools.

The LIDES system has been proposed by the **Language Interaction in Plurilingual and Plurilectal Speakers** (LIPPS) group which was founded with the aim of creating an international language interaction database to encourage research on all aspects (linguistic, sociolinguistic, pragmatic) of language interaction phenomena between typologically close and distant pairs of languages.

The LIPPS project was forged at the group meetings held in Ljouwert/Leeuwarden (Sept. 1994), London (Jan. 1995), Barcelona (Sept. 1995), Nijmegen (Apr. 1996) and Bellaterra/Barcelona (Jan. 1997). LIPPS includes researchers from the following universities: Amsterdam, Autònoma de Barcelona, Lancaster, London, Nijmegen and Pompeu Fabra (Barcelona).

Although the LIPPS group is also exploring other systems, the advantages of adopting LIDES are manifold, since a) it is user-friendly, b) it is specially created to analyse language interaction data and can always be adapted to meet researchers' specific needs, c) it can be used to encourage comparative analysis between different pairs of language, and d) it is compatible with further quantitative analysis and automatic multimedia analysis.

1.2 LIDES coding manual

The goal of the LIPPS group is to make available to language interaction researchers a tool that would enable them to automatize and analyze bilingual data in a systematic way, either to meet their own specific research needs or to be able to contribute to the above-mentioned international database, the LIPPS group is in the process of putting together a coding manual that will help researchers better undertake their tasks and ultimately achieve their research objectives.

This report aims at highlighting the IULA-LIPPS contributions to the LIDES coding manual and their application to two corpora: the LC (La Canonja) bilingual corpus and the CSCD (Code-switching as Communicative Design) trilingual corpus. These contributions involve language-specific proposals that can also be used to transcribe, tag and code data from typologically close and distant languages.
2. Contents of the LIPPS database

The initial LIPPS database contains the following corpora involving the following researchers, speech communities and pairs or triplets of languages (see appendix 6 for spell out of initials):

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Speech communities</th>
<th>Language pairs/triplets</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>Gibraltar</td>
<td>Spanish/English</td>
</tr>
<tr>
<td>PGC</td>
<td>London</td>
<td>Punjabi/English</td>
</tr>
<tr>
<td>PGC</td>
<td>London</td>
<td>Cypriot/English</td>
</tr>
<tr>
<td>PGC</td>
<td>Strasbourg</td>
<td>French/Alsatian</td>
</tr>
<tr>
<td>EE</td>
<td>London</td>
<td>German/English</td>
</tr>
<tr>
<td>MTT,JP, MPB</td>
<td>La Canonja (Tarragona)</td>
<td>Spanish/Catalan, Catalan/Spanish</td>
</tr>
<tr>
<td>MTT, MF</td>
<td>Barcelona</td>
<td>Spanish/Catalan/English</td>
</tr>
<tr>
<td>AB</td>
<td>Netherlands</td>
<td>Turkish/Dutch</td>
</tr>
<tr>
<td>PM</td>
<td>Netherlands</td>
<td>Papiamentu/Dutch</td>
</tr>
<tr>
<td>JT</td>
<td>Brussels</td>
<td>French/Dutch</td>
</tr>
<tr>
<td>MS</td>
<td>Lancashire</td>
<td>Punjabi,Hindi, Urdu, Bengali/English</td>
</tr>
<tr>
<td>MS</td>
<td>London</td>
<td>London English/Jamaican</td>
</tr>
</tbody>
</table>

3. The corpora

3.1 The LC corpus

La Canonja corpus is a Catalan-Spanish/Spanish-Catalan bilingual corpus which consists of one-to-two hour recorded life-stories from thirty informants. La Canonja is a village located at the South-Western limit of Tarragona, the capital of the province. Differently
from what happens with other communities in the outskirts of Tarragona, which attracted immigration during the sixties and have therefore a majority of immigrant population, La Canonja shows an almost balanced distribution between (Catalan-speaking) indigenous people and (Spanish-speaking and Catalan-speaking) immigrant people, since 41,5% of the population (4,402 inhabitants) is of non-local origin.

3.2 The CSCD corpus

The CSCD corpus is a trilingual corpus (Catalan-Spanish-English) which consists of one-to-two hour recorded conversations between trilingual speakers of Catalan, Spanish and English and monolingual English speakers, involved in both formal and informal situations that took place in Barcelona during the summer of 1992.

4. Transcription

4.1 Database format

The files in the basic LIDES version of the database should include the following:

1. All the compulsory headers: @Begin header line, @End header line, @Participant header line, stated in the way indicated in the LIDES coding manual.

2. All the headers which are considered essential to document the nature of any corpus included in the database, in terms of the social variables that refer to the informants, researchers, instruments, size of the corpus and sample, and file matrix language.

3. A main tier transcribed and tagged according to the LIDES coding manual conventions (see sections 4 & 5).

4. Two dependent tiers: the %glo, which is a word-by-word translation to favour potential comparative analysis, and the %eng, which includes an English translation.2

4.2 Tier transcription in the basic LIDES version of the database

At the transcription level, the light version of the database initially involves a main tier, which includes the actual speaker's data and is identified as *INF (for informant), and two

2 See the LIDES coding manual for full details on the exact form of the file, minimal requirements, the form of the rest of the file, documentation file and depfile (where all the CHAT legal characters are stored, together with all those defined by each researcher)
Dependent tiers, which are strongly recommended to allow researchers to explore the potential for comparative analysis between different pairs of languages:

a) the %glo tier: a word related tier that presents literal word-by-word translation into English (or morph-to-morph translation in e.g. Turkish). Each lexeme/morpheme on this tier relates directly to the corresponding lexeme in the same position on the main tier.

b) the %eng tier: an utterance related tier that includes a free translation of the whole utterance into English.

In cases where this correspondence is not straightforward (e.g. Spanish contractions of preposition plus determiner such as al corresponding to to the in English) this is maintained by joining the words on the dependent tier with, for example, an underscore: "to_the".

e.g.:

*INF: pues@2 # me@2 invitaron@2 &a@2 &a@2 <dinar@1> [$c] a@U +...
%glo: well me they_invited to to_to_have_lunch to
%eng: well, I was invited to lunch to...

(LC02/88)

In the cases of data strings that are longer than one line, the %glo tier can be interspersed with the main tier, so linked dependencies between tiers are still easy to follow by the researcher.

e.g.:

*INF: si@2 tiene@2 educacion@2 y@2 te@2 <parla@1> [$c] bien@2
%glo: if has education and you talks well
<tiene@2 que@2> <parlar@1> [$c] la@2 bien@2.
%glo: you_have that to_talk him well
%eng: if he is polite and speaks politely to you, you must
talk to him politely

(LC02/88)

4.3 Transcription unit needed

There are a number of transcription decisions that may seem to be related to syntactic coding but that, in fact, should be taken at the transcription level, and which are set out to

---

3 Researchers should bear in mind, a) (when transcribing on the main tier) that orthography is to be disregarded to prevent any mismatches between ascii signs used in different research contexts, and b) (when coding) that CLAN programs - basic programs in CHILDES - do not make a distinction between upper and lower case.
establish the transcription unit, that is, how the informant’s speech should be presented to make the data more accessible for further analytical purposes.

4.3.1 Definition of INTROS

We take utterances to be the basic transcription unit. The term INTROS (INformant’s TRanscriptiOn String) is proposed to indicate such transcription unit. We define utterances as units (such as, speech acts, speaker’s turns) which may consist of a single word, single phrases and single and complex sentences. The segments that would characterize and constitute an INTROS would satisfy the following criteria:

4.3.2 INTROS typology

a) single words: (yeah)

  e.g.:
  *BRO: I@3 was@3 +/
  *FER: yeah@3
  *FER: you@3 were@3 there@3
  (CSCD/92)

b) single phrases:

  e.g.:
  *INF: al@1 carrer@1 Bisbe@1 Borras@1.
  %eng: In Bisbe Borras street
  (LC12/88)

c) one-verb simple sentences/clauses:

  e.g.:
  *INF: esta@1 al@1 costat@1 de@1 l’@1 <iglesia@2> [Sb] de@1 dalt@1.
  %eng: it’s right next to the High Street church
  (LC12/88)

d) complex sentences:
- main + subordinate:

  *INF: no@1 te@1 deixen@1 perque@1 passen@1 cotxes@1 i@1 tot@1 aixo@1.
  %eng: they won't let you because there’re lots of cars and all
  (LC12/88)

- utterance [clausal constituent]:

  e.g.:
  *FER: he@3 said@3 # Juan@3 +” <es@2 que@2 tenia@2 mucha@2 mas@2 preparacion@2 teorica@2 la@2 presentacion@2 teorica@2 de@2 ellos@2 que@2 su@2 conferencia@2 > [Sc].
  %eng: He said, Juan: their presentation by the chairperson was much more theoretically grounded than their own lecture
  (LC12/88)
4.4 The IULA-LIPPS proposals for the transcription of Catalan-Spanish bilingual data

4.4.1 Transcribing words

We object to phonetic or phonemic notation in the transcription on the main tier, because this would make the transcript difficult to read. However, since the basic LIDES version should allow researchers to explore the potential for comparative analyses of data sets even at the phonetic and phonemic levels, we propose to include this information together with prosodic features, fundamental to the nature of oral data, within a dependent tier.

4.4.2 Transcribing Morphemes on the main tier

Several questions arise in connection with the transcription of morphemes on the main tier:

1) "Morphemicization on the main tier is intended mostly for initial morphemic analysis..." (MacWhinney 1991), therefore, we propose to transcribe morphemes, clitics, etc. on the main tier in a unified way.

2) Transcription and tagging on the main tier brings up the issue of what the basic structural units should be. This concerns compounds, bound morphemes (suffixes, prefixes) and clitics.
This idiosyncratic consideration is a necessity for analysis since in order to facilitate potential comparative analysis we need to preserve "cells" or "slots" for the word-related tiers.

a) **compounds: one or two words?**

E.g.:

*SMI: I@3 noticed@3 that@3 John@3 code@3 +switched@3 a@3 bit@3.  (CSCD3.asc/92)

b) **bound morphemes:**

It may be worthwhile separating bound morphemes where code-switching could also occur:

E.g.:

*INF: I@3 am@3 not@3 toqu@1 -ing@3 it@3  
%eng: I'm not touching it  (CSCD/92)

Since "toqu-" is Catalan and "-ing" English, we need to separate this inflectional morpheme from the root to be able to account for the code-switch (notice the hyphen to mark suffix dependency).

c) **clitics:**

*INF: dona@1 m'@1 ho@1  
%eng: give it to me

5. **Tagging**

5.1 **Main tier**

Tagging on the main tier involves general conventions, language tagging and language interaction tagging:

5.1.1 **General conventions**

Language strings on the main tier will be transcribed and punctuated according to LIDES transcription conventions.

For example,

*INF: pues # me invitaron &a &a dinar a +... 
%eng: well, I was invited to lunch to  (LC02/88)
presents the repetition of one item due to stuttering, marked by the ampersand (&); there is also a pause after the first introductory word, which is signalled by the hatch mark (#). The symbol +... indicates that the sentence is left unfinished.

5.1.2 Language tagging

Words or morphemes -depending on the type of language transcribed (see also "bound morphemes" in 4.4.2) - will be tagged for language individually; i.e. each word will be followed by the @ symbol plus a number, or a character, denoting the language to which they belong.

E.g.:

@1 = Catalan
@2 = Spanish
@3 = English
@U = undecided or unclassifiable

*INF: pues@2 me@2 invitaron@2 & a@2 & a@2 dinar@1 a@U +...
%eng: well, I was invited to lunch to

(LC02/88)

5.1.3 Language interaction tagging

There is the possibility to mark different types of language interaction (formerly language contact) phenomena on this same tier by applying the following procedure:

1) Strings of one or more words presenting an interaction phenomenon will be set off in angle brackets < > to indicate the scope of the phenomenon.

2) Angle brackets will be followed by square brackets [ ] containing the $ symbol plus another symbol defining the specific type of phenomenon. For instance:

c = code-switch

*INF: pues@2 me@2 invitaron@2 & a@2 & a@2 dinar@1 <diinar@1> [$c] a@U +...
%eng: well, I was invited to lunch to

(LC02/88)

b = borrowing

*INF: i@1 <bueno@2> [$b] no@1 ho@1 se@1.
%eng: and, well, I don't know

(LC02/88)
In the case of corpora involving more than two languages, we propose that for syntactic and semantic calques a number follow the $y$ and $e$ symbols, to give account of the language from which the form is calqued.

For instance, if the utterance

*INF: dema@1 <tinc@1 que@1> [$y2] anar@1 al@1 banc@1.
%eng: I have to go to the bank tomorrow

belongs to a file that contains three different languages such as, Spanish, Catalan and English, without the number 2 after the $y$ symbol, there would be no record of the language from which the form in the scoped string is calqued, since it is not the words themselves that are Spanish, but the syntactic structure they present.

The same applies to semantic calques:

*INF: el@2 <grande@2> [$e1] tambien@2 ha@2 estudiado@2.
%eng: the eldest also has a degree

Gran is used in Catalan to refer to age, and it means "the eldest", whereas in Spanish grande is used for size and applies only to objects.

5.2 Dependent tiers

All dependent tiers are optional and may be added by the researcher as needed. There are two types of dependent tiers: a) word-related/linked tiers and b) utterance-related tiers.\(^4\)

---

\(^4\) Except for the %glo (word-related) and the %eng (utterance-related) tiers, which are compulsory in the basic LIDES version of the database.
5.2.1 Word-related tiers

In word-related tiers the data is to be inputted in vertically corresponding slots that relate directly to the lexemes on the main tier. This way a linking between the main tier and the relevant dependent tiers is established.

If information on a dependent tier relates to only a part of the main tier, the marking on the dependent tier should be made so that the correspondence between slots is still maintained. That is, if the information inputted on a dependent tier refers to the third and sixth words on the main tier, positions number one, two, four and five should be filled in by a marker such as the semicolon (;), for instance, to preserve the correspondence. This way the maintenance of linked dependence throughout the tiers will still be met.

e.g.:
*INF: el@2 <grande@2> [is1 tambien@2 ha@2 estudiado@2.
%cod: ; symbol ; ; ;

(LC02/88)

5.2.2 Utterance-related tiers

Utterance-related tiers constitute those dependent tiers whose information relates to the utterance as a whole. For these type of tiers there is no one-to-one relationship between items on this tier and the corresponding ones on the main tier to be maintained, therefore they do not require the markers described above.

6. Coding

It should be noted that information coded on dependent tiers should be kept as separate as possible for the sake of clarity and flexibility. This way, different researchers will be able to decide which information they want to merge depending on the nature of their studies.

The IULA-LIPPS team has proposed a list of potentially useful tiers with their possible components. The coding is carried out only for those features (marked in capital letters) that are specific or not further sub-classified.6

6.1 Word-related tiers

a) %mor tier:

This tier refers to the morphological features of a specific language interaction phenomenon under analysis, and some of its components could be:

---

5 Technically, it will not be necessary to retain the (;) marker in order to operate LIDES. However, the option of using the data as input to VARBRUL would make it necessary to keep the symbol so that a concordance program can convert it into the obligatory legal factor (/ = it doesn’t apply) in VARBRUL.

6 Examples of coded dependent tiers are to be found in Appendix 4.
1) **morphemes** (FREE / BOUND)

2) **bound morphemes** (DERIVATIONAL / INFLECTIONAL: each would be coded for different features, e.g. inflectional morphemes would be coded for NUMBER, PERSON, TENSE, ASPECT, etc. in the case of verbs).

### 6.2 Utterance related tiers

b) **%syn tier**:

This tier refers to the syntactic features of a specific language interaction phenomenon under analysis, and its components could be:

1) **string length** (MORPHEME, WORD, PHRASE, CLAUSE, SENTENCE, TURN).

2) **string type** (CONTENT / FUNCTION).

3) **constituent** (NOUN, VERB, NOUN PHRASE, etc.) / non-constituent or **ragged** (GRAMMATICAL, TELEGRAPHIC), (Muysken, 1996).

4) **dependency** (DEPENDENT / NON-DEPENDENT: i.e. modifier or qualifier, or not).

5) **grammatical function** (SUBJECT, VERB, OBJECT, etc.).

6) **type of utterance** (DECLARATIVE, INTERROGATIVE, IMPERATIVE, EXCLAMATIVE, MARKERS).
c) % sit tier

This tier refers to phenomenon site, that is, whether it occurs intersententially (between sentences), intrasententially (between sentence constituents) or extrasententially (tags, confirmation checks, etc.). Its constituents could be:

1) transition (smooth: WITH DUMMY INSERTION, WITHOUT DUMMY INSERTION / flagged), (Muysken, 1996).

2) site (intersentential: TOTALLY INDEPENDENT SENTENCES, COORDINATED SENTENCES, SUBORDINATED SENTENCES / intrasentencial: INTERCONSTITUENT, INTRACONSTITUENT / extrasentential: TAGS, MARKERS)

3) markers: (STARTERS, CONTINUITY).

d) % bor tier

This tier refers to borrowing features, that is, to factors and variables that can constrain the occurrence of loans and all types of borrowing. Its components could be:

1) phonological integration (FULL, PARTIAL, NO INTEGRATION).

2) morphological integration (FULL, PARTIAL, NO INTEGRATION).

3) type of borrowing (UNIQUE, IDIOSYNCRATIC, REPEATED, RECURRENT, EXTENDED).


e) % soc tier

This tier involves all those social and sociolinguistic factors that could be involved in the occurrence of certain language interaction phenomena. Its components could be:

1) sex (MALE, FEMALE).

2) age (YOUNG, MIDDLE-AGED, OLD).

3) education level (PRIMARY, SECONDARY, HIGH).

4) origin (SPECIFIC PLACE(S) TO BE CODED FOR).

5) type of family (INDIGENOUS, IMMIGRANT, MIXED).

---

7 Similar tiers will be needed for other phenomenon types, such as, semantic and syntactic calques.
8 The exact specification of these factors will depend on the sample size and the proportion of participants in each corpus sample.
6) **contact index** (1 - 10).

7) **directionality of participant's language** (NATIVE SPEAKER TO NATIVE SPEAKER, NATIVE SPEAKER TO NON-NATIVE SPEAKER, NON-NATIVE SPEAKER TO NATIVE SPEAKER, NON-NATIVE SPEAKER TO NON-NATIVE SPEAKER).

f) **%prg tier**

This tier refers to those stylistic and pragmatic features intervening in the occurrence and distribution of language interaction phenomena. Its components could be:

1) **pragmatic functions** (TO CHANGE SUBJECT / TO FILL IN SILENCE / TO FULFILL A SPEECH ACT / TO REPRODUCE OBJECTIVELY / TO PERSONALISE / TO OBJECTIVIZE / etc.).

2) **sociolinguistic style** (INTERVIEW, NARRATION, FREE CONVERSATION, etc.).

3) **researcher's identity** (INTERGROUP, INTRAGROUP).

4) **performance characteristics involved in the context of the phenomenon** (repetition: OF ONE OWN'S UTTERANCE, OF ANOTHER PARTICIPANT'S / HESITATION / pause: LONG, VERY LONG / interruption: SELF-INTERUPTION, BY ANOTHER PARTICIPANT / RETRACING AND CHANGE / FILLER / etc.).

**g) %cva tier** (conversational analysis features)

6.3 Turn-related tiers

Instead of having two types of dependent tiers we propose three different ones: word-related tiers, utterance-related tiers and turn-related tiers.

Turns could be set off by means of the GEM option, the %trn (turn) tier could be included inside the scope of the GEM. This tier would contain information as to turn number, which would facilitate tracing of different participants, and as to whether the phenomenon occurred at turn boundary or across turn boundary.

e.g.:

Bg: 001
*FER: do@3 you@3 know@3 her@3?

*FER: have@3 you@3 met@3 her@3?
*FER:  <es@1 molt@1 maca@1> [$c].
%eng:    she's very nice
Eg:    001
Bg:    002

*BRO:   <of@3 course@3 # in@3 Tarragona@3> [$c].
Eg:    002

(cscd6.asa/92)
7. References


LIDES Coding Manual (forthcoming), LIPPS group.
Appendices

Appendix 1: Headers and transcription sample for file LC/26.asc  
(Catalan Matrix Language)

@Begin  
@Filename: 26.asc  
@Participants: INF Moran Informant, ENT Pujol Entrevistadora  
@Sex of INF: male  
@Age of INF: 36  
@Job of INF: pipe-fitter  
@Mother tongue of INF: Spanish  
@Most usual language of INF: Catalan  
@Matrix language(s) of interview: Catalan/Spanish  
@Comment: INF is a Spanish immigrant married to a Catalan speaker  
@Date of recording: 17-JUL-1992  
@Interviewer: Pujol  
@Date of coding: 11-APR-1996  
@Transcriber and Coder: Turell, Forcadell  
@Warning: FRAGMENT OF FILE

*ENT: pero@1 per@1 exemple@1, es@1 la@1 gent@1, els@1 canongins@1 que@1 et@1 diuen@1 +" o@1 # pero@1 tu@1 no@1 ets@1 d@1 aqui@1.
%trn: 001
%glo: but for example is the people the Canongians who you say oh but you not are from here
%eng: but for example, it is the people, the people from La Canonja that tell you "well, but you are not from here"

*INF: home@1 # directament@1 no@1 t@1 ho@1 diuen@1 mai@1.
%trn: 002
%glo: man directly not you it they_say never
%eng: well, they never say it directly

*ENT: no@1.
%trn: 003
%glo: no
%eng: no
*INF: pero tu per exemple si estas fent +...
%trn: 004
%glo: but you for example if are doing
%eng: but you, for example, if you are doing

*INF: per exemple estas en una taula rodona i tu estas aqui tens gent al radera sentada i un parla en castella.
%trn: 004
%glo: for example, you are in a table round and you are here and have people in the behind sat and one speaks in Spanish
%eng: for example, you are at a round-table discussion and you are here and you have people sitting behind you and one person speaks in Spanish

*INF: doncs i sents el del radera <+ el charnego este con los anos que lleva aqui y mira como habla todavia> [Sc].
%trn: 004
%glo: so and you hear the from the behind the charnego this with the years that takes here and look how speaks still
%eng: so... and you hear someone from behind saying "look at that charnego, he has been living here for ages and see the way he is still speaking"

*INF: coses d'aquelles que alguna manera +...
%trn: 004
%glo: things of these that of some way
%eng: things like that that somehow ...

*ENT: o sigui li diuen a una altra persona.
%trn: 005
%glo: or it be them they say to a other person
%eng: that is they say it to someone else
*INF: si@1 home@1 # pero@1 1 [/-] perque@1 com@1 tu@1 tambe@1 ets@1 nascut@1 a@1 #### doncs@1 d’@1 alguna@1 vegada@1 te@1 una@1 sentir@1 # doncs@1 d’@1 un@1 altre@1 llloc@1.
%trn: 006
%glo: yes man but because since you too
are born in well of some time
you they make to feel so from a other place
%eng: of course but, because since you were also born in ...
therefore sometimes they make you feel as if you were
from somewhere else

*INF: o@1 sigui@1 <que@2 yo@2 me@2 siento@2 de@2 La@2 Canonja@2 # soy@2 del@2 aqui@2 # hace@2 tiempo@2 que@2 vivo@2 aqui@2> [$c].
%trn: 006
%glo: or it be that I myself feel from
La Canonja am from here makes time
that I live here
%eng: so, I feel I belong to La Canonja, I'm from here, it's
been a long time since I've lived here

*INF: <y@2 bueno@2 y@2 ademas@2 participo@2 mucho@2 de@2 la@2 vida@2 del@2 pueblo@2> [$c].
%trn: 006
%glo: and well and besides I participate much of
the life of the town
%eng: and well... and besides I participate a lot in the
village activities

*INF: estoy@2 metido@2 en@2 muchos@2 sitios@2 y@2 yo@2 me@2 imagino@2 pues@2 que@2 trabajo@2 por@2 la@2 canonja@2 tanto@2 o@2 mas@2 que@2 cualquier@2 canongino@2.
%trn: 006
%glo: am put in many places and I myself
imagine so that I work for La Canonja
as much or more than any Canongian
%eng: I am involved in many activities and I imagine that
I work for La Canonja as much or more than any other
person from La Canonja

@End
Appendix 2: Headers and transcription sample for file LC/02.asc
(Spanish Matrix Language)

@Begin
@Filename: 02.asc
@Participants: INF Fernandez Informant, EN1 Pujadas Entrevistador,
EN2 Ferrer Entrevistador, WIF Dona Wife
@Sex of INF: male
@Age of INF: 64;
@Job of INF: pensioner (gardener)
@Mother tongue of INF: Spanish
@Most usual language of INF: Spanish
@Matrix language(s) of interview: Spanish
@Comment: INF is a Spanish immigrant married to a Catalan
speaker.
EN1: Spanish, EN2: Catalan
@Date of recording: 1988
@Interviewer: Pujadas, Ferrer, Ballester
@Date of coding: 15-APR-1996
@Transcriber and Coder: Turell, Forcadell
@Warning: FRAGMENT OF FILE

*INF: pues@2 # me@2 invitaron@2 &a@2 &a@2 <dinar@1> [$c] a@U +...
%tm: 001
%glo: well me they_invited to to_have_lunch to
%eng: well I was invited to lunch

*WIF: per@1 Nadal@1.
%tm: 002
%glo: at Christmas
%eng: at Christmas

*INF: <per@1 Nadal@1> [$c]
%tm: 003
%glo: at Christmas
%eng: at Christmas

*EN2: mhm@1.
%tm: 004
%glo: mhm
%eng: mhm
*INF:  y@2 entonces@2 se@2 quedaron@2 todos@2 caray@2 !
%trn:  005
%glo:  and then themselves they_stayed all Goodness_me
%eng:  and then they all stayed for God's sake!

*INF:  y@2 yo@2 # mire@2 # son@2 personas@2 como@2 otras@2 ## son@2 +...
%trn:  005
%glo:  and I look they_are people like others they_are
%eng:  and I, look, they are people like any others, they are...

*INF:  igual@2 son@2 negros@2 que@2 azules@2 todo@2 y@2 todo@2.
%trn:  005
%glo:  same they_are black as blue as everything and everything
%eng:  no matter whether they are black or blue, or whatever

*EN1:  mhm@2.
%tm:  006
%glo:  mhm
%eng:  mhm

*INF:  porque@2 yo@2 le@2 voy@2 a@2 decir@2 la@2 verdad@2.
%trn:  007
%glo:  because I you go to_tell the truth
%eng:  because I'll tell you the truth

*INF:  yo@2 # como@2 muchos@2 dicen@2 ese@2 porque@2 sean@2 negro@2.
%trn:  007
%glo:  I like many say write_down to that because they_be black
%eng:  I, like many people say "write him down because he is black"

*INF:  o@2 sea@2 # eso@2 yo@2 no@2 # porque@2 yo@2 he@2 esto@2 en@2 Lleida@2 con@2 un@2 amigo@2.
%trn:  007
%glo:  or it_be this I not because I have been in this in Lleida with a friend
%eng:  that is, I don't like this because I have been in this in Lleida with a friend

*INF:  dice@2 +" vente@2 unos@2 cuantos@2 dias@2 que@2 vamos@2.
%trn:  007
%glo:  says come there some some days that we_go
%eng:  he says "come along for some days and we'll go"
*WIF: a@1 la@1 poma@1.
%trn: 008
%glo: to the apple
%eng: to pick apples

*INF: <a@1 la@1 poma@1> [$c].
%trn: 009
%glo: to the apple
%eng: to pick apples

*INF: y@2 estuve@2 dos@2 semanas@2 con@2 tres@2 negros@2.
%trn: 009
%glo: and I was two weeks with three blacks
%eng: and I was two weeks with three black men

@End
Appendix 3: Sample TURN/INTROS for file CSCD/prgaed6.asa

TURNS

@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Matrix language(s) of interview: English, Catalan, Spanish
@Date of recording: 30-JUN-1992
@Date of coding: 14-MAY-1995
@Transcriber and Coder: Turell, Forcadell
@Warning: FRAGMENT OF FILE

*SMI: England is very simple.
%trn: 012

*BRO: mechanisms a sort of discourse not at the technological level that's one problem!
%trn: 013

*SMI: like the "storms"; what have you instead of "storm", "storms" and "north wind", or something
%trn: 014
%eng: like the "storms"; what have you instead of "storm", "storms" and "north wind", or something

*BRO: and just call it north wind.
%trn: 015

*SMI: north wind.
%trn: 016

@End
INTROS

@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Matrix language(s) of interview: English, Catalan, Spanish
@Date of recording: 30-JUN-92
@Date of coding: 14-MAY-95
@Transcriber and Coder: Turell, Forcadell
@Warning: FRAGMENT OF FILE

*SMI: England is very simple.
%trn: 012

*BRO: mechanisms # a sort of discourse not at the technological level.
%trn: 013

*BRO: that's one problem!
%trn: 013

*SMI: like the "storms"
%trn: 014
%eng: like the "storms"

*SMI: what have you instead of "storm", "storms" and "north wind", or something
%trn: 014
%eng: what have you instead of "storm", "storms" and "north wind", or something

*BRO: and just call it north wind.
%trn: 015

*SMI: north wind.
%trn: 016

@End
Appendix 4: Coded fragments for file LC/02.asc

@Begin
@Filename: 02.asc
@Participants: INF Fernandez Informant, EN1 Pujadas Interviewer, EN2 Ferrer Interviewer, EN3 Ballester Interviewer, WIF Dona Wife
@Warning: FRAGMENT OF FILE

*EN1: y@2 en@2 el@2 pueblo@2 estan@2 2 un@2 hermano@2 2 y@2 una@2 hermana@2.
%eng: and in the village there is brother and a sister

*WIF: no@1 dos@1 germanes@1.
%eng: no, two sisters

*INF: no@2 dos@2 hermanas@2 si@2 dos@2 hermanas@2 si@2 dos@2 hermanas@2 si@2.
%eng: no two sisters, yes two sisters, yes two sisters

*EN1: y@2 eh@2 # que@2 relaci on@2 tienen@2 2 con@2 con@2 Baza@2 con@2 el@2 pueblo@2 de@2 origen@2?
%eng: and what sort of relationship do they have with Baza, their hometown?

*EN1: tienen@2 alguna@2 relacion@2 ?
%eng: are they still in contact?

*EN1: van@2 y@2 vuelven@2 alli@2 a@2 menudo@2 etcetera@2 ?
%eng: do they often go there, etc?

*INF: si@2.
%eng: yes

*WIF: no@2 a@2 menudo@2 a@2 menudo@2.
%eng: no, quite often, quite often

*INF: alli@2 pa@2 pasa@2 una@2 2 +...
%eng: there there's a
*INF: allí hacen desto # hacen <mercado> [$c].
%eng: there they have this... there's a market
%syn: $yW $yN $yO
%sit: $iA $i; $iI $i;
%prg: $pO $p0
%soc: $oH $oI $o2
%direc: $d1

*INF: y van.
%eng: and they go
*WIF: no vol dir si anem naltrus.
%eng: no, she means whether we go
*INF: ah.
%eng: I see
*EN1: si.
%eng: yes
*INF: nosotros no.
%eng: no, we don't
*EN1: no.
%eng: you don't.
*WIF: si hi hem quatre vegades.
%eng: yes, we've been there four or five times
*INF: cuatro o cinco vegades [$c].
%eng: four or five times
%syn: $yW $yN $yA
%sit: $iA $i; $iR $i;
%prg: $pH $p1
%soc: $oH $oI $o2
%direc: $d1

*WIF: ja ho he vist tres vegades.
%eng: I've already seen it tree times
*EN1: de vacaciones digamos.
%eng: on holidays, let's say
*WIF: si si.
%eng: yes, yes

*INF: no no de...
%eng: no, no, on...
*EN1: bien.
%eng: ok

@end
**KEY**

Dependent tiers

% **syn:** syntax
% **sit:** site
% **prg:** pragmatics
% **soc:** sociolinguistics
% **dir:** (language) direction

Components

$y$: syntax code
$i$: site code
$p$: pragmatics code
$o$: sociolinguistics code
$d$: (language) direction code

%syn factors

<table>
<thead>
<tr>
<th>Level</th>
<th>Grammatical form</th>
<th>Syntactic function</th>
</tr>
</thead>
<tbody>
<tr>
<td>M Morpheme</td>
<td>N Noun</td>
<td>O Object</td>
</tr>
<tr>
<td>W Word</td>
<td>J Adj</td>
<td>A Adjunct</td>
</tr>
<tr>
<td>P Phrase</td>
<td>R Pro-form</td>
<td>Q Complement</td>
</tr>
<tr>
<td>C Clause</td>
<td>D Adv 3 PrepP U Conjunct T Det etc.</td>
<td>V Verb M Marker etc.</td>
</tr>
</tbody>
</table>
### %sit factors

<table>
<thead>
<tr>
<th>Site</th>
<th>Intersent.</th>
<th>Intrasent.</th>
<th>Extrasent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Intersent.</td>
<td>T Total.indep.</td>
<td>I Interconstituent</td>
<td>G Tag</td>
</tr>
<tr>
<td>A Intrasent.</td>
<td>C Coord.</td>
<td>R Intraconstituent</td>
<td>K Marker</td>
</tr>
<tr>
<td>X Extrasent.</td>
<td>M Main</td>
<td>H (both I &amp; R)</td>
<td>D Disjunct</td>
</tr>
<tr>
<td>B Subord.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### %prg factors

<table>
<thead>
<tr>
<th>Speech acts</th>
<th>Situational</th>
</tr>
</thead>
<tbody>
<tr>
<td>O Orientate</td>
<td>1 Yes</td>
</tr>
<tr>
<td>H Echo</td>
<td>0 No</td>
</tr>
<tr>
<td>E Evaluate</td>
<td></td>
</tr>
<tr>
<td>M Monitor</td>
<td>2 Agree</td>
</tr>
<tr>
<td>Q Quote</td>
<td>etc.</td>
</tr>
</tbody>
</table>

### %soc factors

<table>
<thead>
<tr>
<th>Contact index</th>
<th>Origin</th>
<th>Most usual language</th>
</tr>
</thead>
<tbody>
<tr>
<td>L Low</td>
<td>I Immigrant</td>
<td>1 Catalan</td>
</tr>
<tr>
<td>A Average</td>
<td>C Canonjan</td>
<td>2 Spanish</td>
</tr>
<tr>
<td>H High</td>
<td>S Second generation</td>
<td></td>
</tr>
</tbody>
</table>

### %dir factors

<table>
<thead>
<tr>
<th>Language direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Spanish-Catalan</td>
</tr>
<tr>
<td>2 Catalan-Spanish</td>
</tr>
</tbody>
</table>

Note: when factors do not apply we make use of the semicolon (;).
Appendix 5: Six-step file (prgaed6.asa)

1st step

@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Warning: SAMPLE FILE. ADAPTED (transcription conventions only)

*BRO: I was +/.  
*FER: you were there !  
*SMI: were you ?  
*BRO: +, I was surprised about the topic .  
*BRO: he was being ironical .  
*SMI: not that ironical maybe .  
*FER: vols mes ?  
*SMI: no # gracies .  
*FER: no I do 'nt think so .  
*BRO: no .  
@End
2nd step

@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Warning: SAMPLE FILE. ADAPTED (transcription conventions + turn specification)

*BRO: I was +/.  
%tn: 001

*FER: you were there!  
%tn: 002

*SMI: were you?  
%tn: 003

*BRO: +, I was surprised about the topic.  
%tn: 004

*BRO: he was being ironical.  
%tn: 004

*SMI: not that ironical maybe.  
%tn: 005

*FER: vols mes?  
%tn: 006

*SMI: no # gracies.  
%tn: 007

*FER: no I do 'nt think so.  
%tn: 008

*BRO: no.  
%tn: 009

@End
3rd step

@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Warning: SAMPLE FILE. ADAPTED (transcription conventions + turn specification + language tagging + comments)

*BRO: I was +/
%trn:  001

*FER: you were there !
%trn:  002

*SMI: were you ?
%trn:  003

*BRO: +, I was surprised about the topic.
%trn:  004

*BRO: he was being ironical.
%trn:  004

*SMI: not that ironical maybe.
%trn:  005

*FER: vols mes [% offering] ?
%trn:  006

*SMI: no # gracies .
%trn:  007

*FER: no I do 'nt think so .
%trn:  008

*BRO: no.
%trn:  009

@End
4th step

@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Warning: SAMPLE FILE. ADAPTED (transcription conventions + turn specification + language tagging + comments + phenomenon scope marking and language interaction tagging)

*BRO: I@3 was@3 +/.  
%tm: 001

*FER: you@3 were@3 there@3 !  
%tm: 002

*SMI: were@3 you@3 ?  
%tm: 003

*BRO: +, I@3 was@3 surprised@3 about@3 the@3 topic@3.  
%tm: 004

*BRO: he@3 was@3 being@3 ironical@3.  
%tm: 004

*SMI: not@3 that@3 ironical@3 maybe@3.  
%tm: 005

*FER: <vols@1 mes@1> [$c] [% offering] ?  
%tm: 006

*SMI: <no@1 # gracies@1> [$c].  
%tm: 007

*FER: no@3 I@3 do@3 'nt@3 think@3 so@3.  
%tm: 008

*BRO: no@3.  
%tm: 009

@End
5th step

@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Warning: SAMPLE FILE. ADAPTED (transcription conventions + turn specification + language tagging + comments + phenomenon scope marking and language interaction tagging + %glo and %eng)

*BRO: I@3 was@3 +/. %trn: 001

*FER: you@3 were@3 there@3 ! %trn: 002

*SMI: were@3 you@3 ? %trn: 003

*BRO: +, I@3 was@3 surprised@3 about@3 the@3 topic@3. %trn: 004

**BRO: he@3 was@3 being@3 ironical@3. %trn: 004

*SMI: not@3 that@3 ironical@3 maybe@3. %trn: 005

*FER: <vols@1 mes@1> [c] [% offering] ? %trn: 006 %glo: you_want more %eng: some more?

*SMI: <no@1 # gracies@1> [C]. %trn: 007 %glo: no thanks %eng: no thanks

*FER: no@3 I@3 do@3 'nt@3 think@3 so@3 . %trn: 008

*BRO: no@3. %trn: 009

@End
6th step
@Begin
@Filename: prgaed6.asa
@Participants: BRO Brown Adult, SMI Smith Adult, FER Ferrer Adult
@Warning: SAMPLE FILE. ADAPTED (transcription conventions + turn specification + language tagging + comments + phenomenon scope marking and language interaction tagging + %glo and %eng + coding)

*BRO: I@3 was@3 +/.  
%trn: 001

*FER: you@3 were@3 there@3 !  
%trn: 002

*SMI: were@3 you@3 ?  
%trn: 003

*BRO: +, I@3 was@3 surprised@3 about@3 the@3 topic@3.  
%trn: 004  
*FER: he@3 was@3 being@3 ironical@3.  
%trn: 004

*SMI: not@3 that@3 ironical@3 maybe@3.  
%trn: 005

*FER: <vols@1 mes@1> [$c] [% offering] ?  
%trn: 006  
%glo: you_want more  
%eng: some more?  
%sit: $s41  
%prg: $p2 $p48 $p72 $p81

*SMI: <no@1 # gracies@1> [$c].  
%trn: 007  
%glo: no thanks  
%eng: no thanks  
%sit: $s41  
%prg: $p1 ; $p72 $p81

*FER: no@3 I@3 do@3 'nt@3 think@3 so@3 .  
%trn: 008

*BRO: no@3.  
%trn: 009
@End
### Appendix 6: List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>Ad Bachus</td>
</tr>
<tr>
<td>b</td>
<td>borrowing</td>
</tr>
<tr>
<td>%bor tier</td>
<td>borrowing coding line</td>
</tr>
<tr>
<td>c</td>
<td>code-switch</td>
</tr>
<tr>
<td>%cva tier</td>
<td>conversational analysis line</td>
</tr>
<tr>
<td>CHILDES</td>
<td>Child Language Data Exchange System</td>
</tr>
<tr>
<td>CSCD</td>
<td>Code Switching as Communicative Design</td>
</tr>
<tr>
<td>DM</td>
<td>Discourse Marker</td>
</tr>
<tr>
<td>e</td>
<td>semantic calque</td>
</tr>
<tr>
<td>EE</td>
<td>Eva Eppler</td>
</tr>
<tr>
<td>%eng</td>
<td>English free translation</td>
</tr>
<tr>
<td>%glo</td>
<td>gloss</td>
</tr>
<tr>
<td>*INF</td>
<td>Informant (*BRO, *FER, *SMI = other informants' names)</td>
</tr>
<tr>
<td>INTROS</td>
<td>INformant's TRanscriptiOn String</td>
</tr>
<tr>
<td>IULA</td>
<td>Institut Universitari de Lingüística Aplicada</td>
</tr>
<tr>
<td>JT</td>
<td>Jeanine Treffers</td>
</tr>
<tr>
<td>JP</td>
<td>Juan Jose Pujadas</td>
</tr>
<tr>
<td>LC</td>
<td>La Canonja</td>
</tr>
<tr>
<td>LIPPS</td>
<td>Language Interaction in Plurilingual and Plurilectal Speakers</td>
</tr>
<tr>
<td>LIDES</td>
<td>Language Interaction Database Exchange System</td>
</tr>
<tr>
<td>MF</td>
<td>Montserrat Forcadell</td>
</tr>
<tr>
<td>MM</td>
<td>Melissa Moyer</td>
</tr>
<tr>
<td>%mor tier</td>
<td>morphological coding line</td>
</tr>
<tr>
<td>MPB</td>
<td>Merce Pujol Berche</td>
</tr>
<tr>
<td>MS</td>
<td>Mark Sebba</td>
</tr>
<tr>
<td>MTT</td>
<td>M. Teresa Turell</td>
</tr>
<tr>
<td>PGC</td>
<td>Penelope Gardner-Chloros</td>
</tr>
<tr>
<td>PM</td>
<td>Peter Muysken</td>
</tr>
<tr>
<td>%prg tier</td>
<td>stylistic and pragmatic line</td>
</tr>
<tr>
<td>%sit tier</td>
<td>phenomenon site coding line</td>
</tr>
<tr>
<td>%soc tier</td>
<td>sociolinguistic coding line</td>
</tr>
<tr>
<td>%syn tier</td>
<td>syntactic coding line</td>
</tr>
<tr>
<td>%trn</td>
<td>turn specification line</td>
</tr>
<tr>
<td>y</td>
<td>syntactic calque</td>
</tr>
</tbody>
</table>