Viaplana, Joaquim; Maria-Rosa Lloret, Maria-Pilar Perea i Esteve Clua (2007): COD. Corpus Oral Dialectal. Barcelona: Promociones y Publicaciones Universitarias (PPU). [CD-ROM]
ISBN: 978-84-477-0990-8

Criteria_English

COD QUESTIONNAIRE DATA

I. GENERAL INTRODUCTION

The *Corpus Oral Dialectal* (COD) of contemporary Catalan contains information on the six main dialects of Catalan: Algherese, Balearic, Central, North-Western, Roussillon and Valencian. The data were obtained through a series of interviews carried out in all county (*comarca*) capitals –or equivalent towns– of the Catalan linguistic domain between 1994 and 1996. The interview consisted of two blocks: the production of an oral text on a free topic and answering a questionnaire. This questionnaire was organized into different parts: relevant phonetic aspects, regular verbal morphology, pronominal clitics, articles, possessives, strong personal pronouns, demonstratives and locatives. The results obtained were organized into eight different databases (© Microsoft Access), each with a structure adapted to its contents. We present in this CD-ROM the databases containing all the information transcribed from the questionnaire, which are stored in the local network of the Department of Catalan at the University of Barcelona. For a description of the method of data collection for each field and database structure, see the sections below.[1]

The information collected in the databases corresponds to a total of 303 interviews. As a rule, three autochthonous informants were interviewed at each location, aged between 30 and 45 years old. This age group guaranteed the interviewees had not been schooled in Catalan. The informants also had in common belonging to the middle class, not having spent lengthy periods away from their homes and having no higher education. In order to preserve anonymity, we decided to identify informants by acronyms made up by two or three signs.

Data are transcribed according to the conventions of the International Phonetic Alphabet (IPA). Transcription is wide: diacritics are only used to reflect word accents. In our case, though, the diacritic recommended by the IPA to transcribe the accent ([ˈ]) is not used, and we used instead the one usually employed to transcribe high tone ([´]) on the

[1] The COD was carried out within the framework of projects PB94-0951, PB97-0889, BFF2001-3798 and HUM2004-01504, all with funding from the Spanish Ministry of Education and Science (and from the European Regional Development Fund, ERDF, in the last case). The information on project HUM2004-01504 can be consulted at the following address: <http://www.ub.edu/lincat>.

vowel of the stressed syllable. Unstressed functional words do not have an accent, such as the conjunctions *i*, *ni*, *si* and *que*; the prepositions *de*, *a*, *amb*; articles, etc. The secondary accents caused by weakening of a word sequence are not marked either, nor deaccentuation in the case of an accent clash (therefore, the sequence *és més bo* is transcribed as [éz méz βɔ́] and not [èz mez βɔ́], and the sequence *va beure* is transcribed as [bá βɛ́wɾə] and not [ba βɛ́wɾə]).

The number of records each database contains is specified in the following table:

Database	Number of registers
Verbs	68262
Phonetics	55270
Pronominal clitics	24766
Articles	5049
Possessives	5223
Personal pronouns	1608
Demonstratives	1512
Locatives	803
TOTAL	162.493

Much of the interest of the project lays in providing users with all the data obtained for each answer at each location. Thus, we present three data sets by location, which correspond to the three surveys carried out at each selected location. Therefore, for each question in the survey, the data obtained can match three different situations: total coincidence –the ideal situation– mostly coincident, or total divergence. But the project was also interested in providing users with simplified and organized data in the case of morphological data, which contain the paradigms corresponding to regular verbal inflection, pronominal clitics, articles, possessives, strong personal pronouns, demonstratives and locatives. To achieve this result, the project implemented a filter-program which selected mechanically majority answers by identities. In the case of a total divergence, the most systematic answer was chosen by hand, that is, the answer which suited the most the set of location data. In the case of the data corresponding to the phonetic aspects, we decided to present all data collected without selecting the majority form, because these are not closed paradigms but gather different phonetic realizations of the segments in different contexts. In practice, the databases in the CD-ROM contain the registers obtained from the three informants surveyed at each location, and in the morphological data, the majority form resulting from the filter –automatic or by hand– is stressed among all answers. Data can be consulted from the whole data or filtered data.

II. Description of the COD databases

II. 1 The verb database

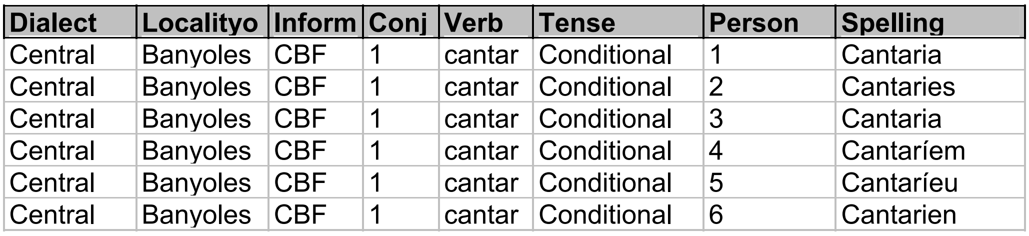
The database consists of the five model verbs: *cantar* (1) 'to sing', *perdre* (2a) 'to lose', *beure* (2b) 'to drink', *sentir* (3a) 'to feel' and *servir* (3b) 'to serve'. This classification, which subdivides verbs from the second and third conjugation according to the absence / presence of the 'extension' segment, is based fundamentally on Viaplana (1984, 1986).[2] Initially, these are the verbs all informants conjugate; nevertheless, if an informant has difficulties in conjugating one, it is replaced by another one of the same model (for instance, *patir* 'to suffer' by *servir*). The method used to obtain the answers has been the completion of sentences read by the interviewer (*Cada dia jo ... aigua* 'Everyday I ... water'). As a rule, the form the informant produces spontaneously is taken as the valid answer, and is, therefore, entered into the database.

In addition to the five model verbs, sometimes additional verbs have been used to confirm answers in case of doubt of the informant (for instance, in doubt between [sentís] and [sentiɣés] 'I felt (subjunctive)', the interviewee is made to conjugate the form for *dormir* 'to sleep') or to reflect characteristic phenomena of the dialect in question (such as the vowel harmony of [mɔ́ɣɔ] for *moga* 's/he move (subjunctive)'. When the informant conjugates all or most of the paradigm of the additional verb a new template is created in the database and the verb is classified according to the

[2] Viaplana, Joaquim (1984) "La flexió verbal regular del valencià", in Casanova, Emili (ed.), *Estudis en memòria del professor Manuel Sanchis Guarner: Estudis de Llengua i Literatura Catalanes*, I, 391-407, University of Valencia, Valencia; Viaplana, Joaquim (1986) "Morfologia flexiva i flexió verbal catalana", *Llengua i Literatura* 1, 385-403. Other works along this line are: Lloret, Maria-Rosa and Joaquim Viaplana (1998) "Variació morfofonològica. Variants morfològiques", *Caplletra* 25, 43-62; Clua, Esteve (1998) *Variació i distància lingüística. Classificació dialectal del balear a partir de la morfologia flexiva*, doctoral thesis, University of Barcelona; Perea, Maria Pilar (2000) *Estructura i variació en el verb català de començaments de segle: La flexió verbal en els dialectes catalans (Alcover-Moll)*, doctoral thesis, University of Barcelona.

conjugation and presence or absence of extension.[3] In Valencian, the additional verbs which can appear conjugated, either completely or partially, are *moldre* ‘to grind’, *coure* ‘to cook’, *dormir* ‘to sleep’, *patir* ‘to suffer’ and *moure* ‘to move’. These three last ones can also be found in North-Western Catalan, together with *veure* ‘to see’ and *plorar* ‘to cry’. In the case of Algherese, the verbs *odiar* ‘to hate’, *canviar* ‘to change’, *conèixer* ‘to know’, *dir* ‘to say’, *eixir* ‘to go out’ (used as a model verb instead of *sentir*) and *dormir* ‘to sleep’ are also conjugated. As in this last dialect the conjugated verbs are more numerous and the extension segment is present in all conjugations, the following classification was adopted: *cantar* (1a), *odiar* (1b), with palatal extension, and *canviar* (1c), with velar extension; *perdre* (2a), *conèixer* (2aa), without extension, and *beure* (2b), *dir* (2bb), with velar extension; *eixir* (3a), *dormir* (3aa), without extension, and *servir* (3b), with palatal extension.

The verb database consists of 19 boxes. The first three ones specify the informant dialect, location and code. The four following boxes show the conjugation, infinitive form, tense and person of the verb form which will be transcribed in the FONÈTICA (Phonetics) box. In addition to this, to make the search easier, the orthographic form corresponding to the Central standard is included.



Dialect	Localityo	Inform	Conj	Verb	Tense	Person	Spelling
Central	Banyoles	CBF	1	cantar	Conditional	1	Cantaria
Central	Banyoles	CBF	1	cantar	Conditional	2	Cantaries
Central	Banyoles	CBF	1	cantar	Conditional	3	Cantaria
Central	Banyoles	CBF	1	cantar	Conditional	4	Cantaríem
Central	Banyoles	CBF	1	cantar	Conditional	5	Cantaríeu
Central	Banyoles	CBF	1	cantar	Conditional	6	Cantarien

As to the FONÈTICA box, we should bear in mind that, although the informant usually pronounces verbal forms within a sentence, they are transcribed in isolation, and do not reflect, for instance the voicing that may have occurred because of the contact between the verb and the following word in the recorded sequence. In the case of the periphrastic present perfect, we do find reflected the assimilation phenomena ([vám bə́wɾə] for *van beure*, and [báŋ kantá] for *van cantar*) and the voicing phenomena ([báz βɛ́wɾə] for *vas beure*) which can be caused by the contact between the auxiliary and the infinitive. In the case of the sibilant fusion usually affecting the second person singular of this same

[3] A template is created provided the informant conjugates, at least, the present indicative and present subjunctive, the imperfect indicative and imperfect subjunctive, and the imperative. If the informant conjugates fewer verb tenses, they will be transcribed in the OBSERVACIONS (Observations) box.

tense of the verbs *sentir* and *servir*, we always chose to preserve whole the form of the main verb: [bá sərβí] for *vas servir*.

The form transcribed in the FONÈTICA box is segmented morphologically; the result of this segmentation is distributed in the following boxes: ARREL (root), EP (epenthesis immediately after the root), EXT (extension), TEMA (theme vowel), MODE (mark of infinitive, gerund and participle), MT (mood/tense/aspect), EP2 (epenthesis after mood and mood/tense/aspect morphemes) and NP (number/person). We include below an excerpt from the database with examples for each of these boxes:

Phonetics	Root	Ep	Ext	Theme	Mode	MT	Ep2	NP
kəntəɾíə	kənt			ə		ɾíə		
pərðərɛ́m	pərð	ə				ɾɛ́		m
bəvént	bə		v	é	nt			
bə́wɾə	bə́w				ɾ		ə	
sénten	sént						e	n
serβîʃko	serβ		îʃk			o		

These boxes, which include the phonetic form and the morphological segmentation, can be left blank in some cases:

a) When the informant does not conjugate a form or whole verb tense, either because he/she forgets or because it escapes him or her at that moment.
b) When the informant gets a tense or person wrong.
c) When the informant speaks too quickly or does not finish conjugating a verb.

In all these situations, the appropriate comment is included in the OBSERVACIONS box. This box can also include additional comments on the verb forms elicited by the informant: some relevant phonetic or phonological aspect, appreciations of the informants as to use (generational, social or geographical factors) and comments on the doubts or difficulties the informant may come across when conjugating a form. Also the forms of several verb tenses the interviewer asks are included as complementary information (see note 2).

We include below some of the most frequent sentences in the OBSERVACIONS box:

- The informant does not conjugate this person.
- The informant makes a mistake: he/she says [x].
- The informant speaks too quickly and does not finish the form.
- The informant uses both forms without distinction.

- The informant says he/she uses form [x] more frequently.
- The informant also accepts the form proposed by the interviewer: [x].
- The informant is also asked to conjugate the verb ‘dormir’: [x].
- The informant explains the autochthonous form is [x].

Finally, mention must be made to the ALTERNATIVES (Alternatives) box, where the alternative forms proposed by the informant are transcribed –should there be more than one, they appear separated by commas. Among the usual alternatives we can mention the following:

a) Forms the informant can use spontaneously with the elicited forms without distinction.
b) Forms the informant uses less frequently.
c) Standard varieties used in a formal context.
d) Forms marked by their use by one generation or another.
e) Autochthonous forms of the informant’s town, but which he/she has stopped using because of self-correction.

As mentioned above, should the informant comment on the use of these forms, it will be included in the OBSERVACIONS box. It is convenient to point out that in the ALTERNATIVES box only the forms the informant has pronounced are transcribed. When the informant accepts alternative forms proposed by the interviewer, but does not pronounce them, the ALTERNATIVES box is left blank, but it is recorded in the OBSERVACIONS box.

II. 2 THE PHONETICS DATABASE

This database includes the phonetic transcription of a series of isolated words and word sequences. From these data, we gathered the different phonetic realizations of one or more segments in a particular context, for each of the six main dialects of Catalan. In the case of word sequences, the specific object of study are the phonological processes which can take place between words, such as voicing and assimilations. Generally speaking, the method employed to collect the data was presenting a series of drawings or photographs the informant had to identify. When this was not possible, work was done through sentence completion, for instance, *El contrari de tou és ...* ('The contrary of soft is ...'). Finally, in word sequences a further procedure was used: the informant had to translate into Catalan a word sequence written in Spanish (in the case of Central, North-Western, Valencian and Balearic Catalan), in French (in the case of Roussillon Catalan) or in Italian (in the case of Algherese Catalan).

Balearic, Algherese and Roussillon have a higher number of records than the other dialects. With this additional information, we intended to gather characteristic phenomena from these dialects.

The phonetics database is headed by the INFORMANT, LOCALITAT and ID boxes – common to all databases. These boxes collect respectively the informant's code and location, and the numeric identifier of each record. The other boxes of the phonetics database are the following:

MOT_ANT: contains the orthographic transcription of the words preceding the word or sequence of words studied.

MOT_ANT_FON: contains the phonetic transcription of the words preceding the word or sequence of words studied.

ARREL: contains the phonetic transcription of the root of the word studied, for example. [**ruð**ɛ́tə]. [4]

VOC_INTER: contains the phonetic transcription of the vowel (if any) preceding the final nasal in some roots, for example [ɔ́m**ə**ns].

NASAL: contains the phonetic transcription of the final nasal consonant (if any) in some roots, for example [ruðó**n**ə].

DIMINUT: contains the phonetic transcription of the diminutive morph (if any), for examples [ʎiβɾ**ɛ́t**].

VOC_FINAL: contains the phonetic transcription of the final unstressed vowel (if any) which is not part of the root, for example [káz**ə**], [rəʎɔ́dʒ**ə**].

PLURAL: contains the phonetic transcription of the plural morph (if any).

[4] The words with an -*n* in the root (*hòmens* 'men' type) receive a different treatment. See the following boxes.

SEQÜÈNCIA: contains the phonetic transcription of the sequence of words in question (see note 1), for example [péjʒ βláw]. This box is only for word sequences, which are transcribed phonetically. In these cases, the previous boxes will be empty.

SENCER: contains the phonetic transcription of the whole word without morphological segmentation, for example [ʎiβɾɛ́t].

FORMES_ALT.: contains the phonetic transcription of the alternative forms proposed by the informant.

OBSERVACIONS: contains various data. It contains the transcription of the answers which do not provide the information corresponding to the question asked or relevant comments which add information or clarify the informant's answer. Also comments on the informant's hesitation between two or more forms, on difficulties in understanding an answer, etc. We present below some examples of the most frequent contents of the OBSERVACIONS box:

- The informant says [x].

  The answer is fully transcribed in this box only; these are cases where the informant's answer is not valid, because the expected sound or the context necessary to observe this sound do not occur in the word the informant pronounces.
- The informant first produces the following form, which he corrects spontaneously: [x].

  In these cases the informant first produces a form, which he corrects spontaneously for another one, which is transcribed in the main boxes. The form first pronounced is transcribed in OBSERVACIONS.
- The answer is not understood.
- The informant is not asked this question.
- The informant says ______ instead of _____. For example, *L'informant diu el femení en lloc del masculí* ('The informant elicitates the feminine form instead of the masculine one').

  This type of remark is used in cases where, although the informant changes the gender or number (in the case of nouns and adjectives) or the tense or person (in the case of verbs), the sound or expected sound contact are maintained. In these cases the form the informant pronounces is fully transcribed in the main boxes.
- The final consonant becomes voiced by contact with the following voiced vowel/consonant.

  In the cases where the expected word ends with a consonant produced by contact with the initial voiced vowel or consonant of the word the informant pronounces next, we decided to transcribe the final consonant as voiceless. Such is the case of *dit*_**s**_ [díts] + *de la mà* (lit. ‘fingers of the hand’), *camp*_**s**_ [káms] + *de futbol* (lit. ‘fields of soccer’), etc., where the words studied are only *dits* and *camps*.[5]

In the case of word sequences where the informant changes the second word for another one which maintains the same phonetic characteristics as the one proposed by the interviewer, the answer is not transcribed in the remarks, but in the main boxes. For example, *serp intel·ligent* (‘intelligent snake’) for *serp hàbil* (‘skilful snake’), where the context under study is *–rp* + *vowel*.

As to the search system, there are a series of boxes which gather the morphological information of the word studied and the phonetic feature we wish to study in each case. These boxes are marked with the values 1 (presence of the feature) or 0 (absence of the feature). We present the list below:

a) Morphological information:

- SUBST_S: singular noun
- SUBST_M: masculine noun
- SUBST_F: feminine noun
- SUBST_PL: plural noun
- DIM: diminutive
- ADJ_S: singular adjective
- ADJ_M: masculine adjective
- ADJ_F: feminine adjective
- ADJ_PL: plural adjective
- VERB
- ALTRES: in the cases where the word does not belong in any of the previous categories.
- CONTACT: in the cases of word sequences.

b) Phonetic feature

- VOC_INICIAL: initial vowel
- VOC_PRE: pretonic/pre-stressed vowel
- E_TON: stressed *e*

[5] This remark is only valid for isolated words, not for word sequences where precisely the phonological processes caused by contact are the focus, for example, [pɛ́z ənórmə] ‘huge weight’.

- O_TON: stressed *o*
- VOC_AL_TON: any stressed vowel, except *e* and *o*
- VOC_POST: post-stressed vowel
- VOC_FIN: final vowel
- GLIDES
- CONS. OCLUSIVA / APROXIMANT: stop / approximant consonant
- CONS. FRICATIVA
- CONS. AFRICADA
- CONS. NASAL
- CONS. LATERAL
- CONS. RÒTICA
- GR_CONS_IN: initial consonant group
- GR_CONS_ME: medial consonant group
- GR_CONS_FI: final consonant group
- GRUP_HETE: heterorganic group
- GRUP_HOMO: homorganic group

II. 3 The pronominal clitic database

The pronominal clitic database contains information from a questionnaire consisting of a series of sentences written in Spanish –in the case of Catalonia, Valencia, Andorra and the Balearic Islands– in French –in the case of Roussillon– and in Italian –for Alghero. From the Catalan translation of these sentences, we expected to obtain the appearance of a pronoun or pronoun sequence in a particular context. Thus, we obtained the forms the weak pronouns *em*, *et*, *es*, *el*, *la*, *els* adopt –accusative third person masculine plural– *les*, *ens*, *us*, *en*, *hi* and *ho* in prevocalic and preconsonantal proclisis and in postvocalic and postconsonantal enclisis. We also obtained information on the combinations of third person datives and accusatives, both in proclitic (without considering whether the verb begins by vowel or consonant) and enclitic position (postvocalic and postconsonantal). The number of sentences presented varies according to the dialect.

In our database we assigned a numeric identifier to each sentence (ID_PF). We have segmented the answers in the following boxes, and presented them in a linear order below –from left to right– as they appear:

MOTS_ANT: contains the orthographic transcription of the word sequence preceding the group formed by the clitic and the verb.

MOTS_ANT_FON: contains the phonetic transcription of the word sequence preceding the group formed by the clitic and the verb.

V_I1: contains the phonetic transcription of the vowel (if any) preceding the consonant of the first pronoun, for example [kwán **ə**d réntu stáz miʎó].

PRO1: contains the phonetic transcription of the consonant of the first pronoun or of the vowel of the pronouns *hi* and *ho*, for example [kwán ə**d** réntu stáz miʎó] i [ə rɛ́wz **j** áz ð əná suβín].[6]

V_FIPRE1: contains the phonetic transcription of the vowel (if any) following the consonant of the first pronoun, for example [l**ə** səlúðə káðə ðíə].

PLUR1: contains the phonetic transcription of the consonant of the plural morph (if any) of the first weak pronoun, for example [lə**z** ɔ́βɾə].

V_FIPOST1: contains the phonetic transcription of the vowel (if any) following the plural morph of the first pronoun, for example [əlz**ə** méɲʒə].

V_I2: contains the phonetic transcription of the vowel (if any) preceding the consonant of the second pronoun, for example [əlz **ə**ŋ kumpɾəɾé wn tɾɔ́s].

PRO2: contains the phonetic transcription of the consonant of the second pronoun or of the vowel in the pronouns *hi* and *ho*, for example [əlz ə**ŋ** kumpɾəɾé wn tɾɔ́s] i [əlz **i** kómpɾəs].

V_FIPRE2: contains the phonetic transcription of the vowel (if any) following the consonant of the second pronoun, for example [li l**e**s kompɾaɾé].

PLUR2: contains the phonetic transcription of the consonant of the plural morph (if any) of the second pronoun, for example [li le**s** kompɾaɾé].

V_FIPOST2: contains the phonetic transcription of the vowel (if any) following the plural morph of the second pronoun, for example [s əlz**ə** méɲʒə kaðə ðíə].

VOC_VERB: contains the phonetic transcription of the verb beginning with vowel (only in proclisis cases), for example [m **əréɣlu**]. It also contains the phonetic transcription of the verb beginning with a vowel which has been ellided, for example [li **splikəɾé**].

[6] In this box the plural morph is also transcribed when it precedes the consonant of the third person pronoun (this happens in some varieties of Minorca). For example [ə**zð** ɔ́bɾə] for *els obre*.

CONS_VERB: contains the phonetic transcription of the verb beginning with consonant (only in proclisis cases), for example [əm **réntu**].

VERB_VOC: contains the phonetic transcription of the verb finishing in vowel (only in enclisis cases), for example [**réntə**m]. It also contains the phonetic transcription of the verb finishing in a vowel which has been elided, for example [**pɔ́z**u].

VERB_CONS: contains the phonetic transcription of the verb finishing in a consonant which is not *–nt*, *–r* or semivowel *–u* (only in enclisis cases), for example [**féz**nə].

VERB_W: contains the phonetic transcription of the verb finishing in semivowel *–u* (only in enclisis cases), for example [**ərəɣlɛ́w**zə].

VERB_R: contains the phonetic transcription of the verb finishing in *–r* (only in enclisis cases), for example [**rəntár**mə]. It also contains the phonetic transcription of a verb finishing in an *r* which has been elided, for example [**rəntá**lzə].

VERB_NT: contains the phonetic transcription of the verb finishing in *–nt* (only in enclisis cases), for example [**pləɲʃan**lə́s].

VO_I1: contains the phonetic transcription of the vowel (if any) preceding the consonant of the first pronoun in enclisis cases, for example [dunɛ́w**ə**lzi].

P_1: contains the phonetic transcription of the consonant of the first pronoun (if any) in enclisis cases or of the vowel of pronouns *hi* and *ho*, for example [dunɛ́wə**l** zi] i [əɾíβə **j** əβiát].

VO_FE1: contains the phonetic transcription of the vowel (if any) following the consonant of the first pronoun in enclisis cases, for example [əkumpáɲəl**ə**].

PL1: contains the phonetic transcription of the consonant of the plural morph (if any) of the first pronoun in enclisis cases, for example [dunɛ́wəl**z**i].

VO_FST1: contains the phonetic transcription of the vowel (if any) following the plural morph in enclisis cases, for example [dunɛ́wəlz**ə**].

VO_I2: contains the phonetic transcription of the vowel (if any) preceding the consonant of the second pronoun in enclisis cases, for example [əmbíən**ə**lzi].

P_2: contains the phonetic transcription of the consonant of the second pronoun (if any) in enclisis cases or of the vowel of pronouns *hi* and *ho*, for example [dónali**l** es] i [kompɾálz**i**].

VO_FE2: contains the phonetic transcription of the vowel (if any) following the consonant of the second pronoun in enclisis cases, for examples [dónalil**e**s].

PL2: contains the phonetic transcription of the consonant of the plural morph (if any) of the second pronoun in enclisis cases, for example [dónalile**s**].

VO_ST2: contains the phonetic transcription of the vowel (if any) following the plural morph in enclisis cases, for example [əmbíənəlz**i**].

MOTS_POST: contains the orthographic transcription of the word sequence following the group formed by the clitic and the verb.

MOTS_POST_FON.: contains the phonetic transcription of the word sequence following the group formed by the clitic and the verb.

SENCER: contains the phonetic transcription of the weak pronoun (only in proclisis cases).

OBSERVACIONS: contains various data. We transcribed: from answers which do not provide information on the question asked (see below) to alternative forms proposed by the informant. In addition to this, some comments are included, such as the hesitation of the informant between two or more forms, difficulties in understanding an answer, etc. We include below a selection of the most frequent contents of this box:

- The informant also produces the form [x].
- First the informant produces the following form, which he/she corrects spontaneously: [x].
- The informant hesitates between this form and form [x].
- The informant uses this form and form [x] without distinction.
- The informant explains elderly people use the form [x].
- The answer is not quite understood.

Every time the informant's answer was not the one expected, we decided not to segment the sequence and to fully transcribe it in the OBSERVACIONS box. The specific cases are the following:

a) When, according to the question, the pronouns are expected to appear in enclitic position but, instead, they appear in proclitic position, or the other way round (for example, the answer ***Les*** *va haver d'acompanyar al col·legi* 'S/he had to accompany them (fem.) to school') to a sentence such as *Tuvo que acompañar**las** al colegio*).

b) When, according to the question, we observed the informant changed the gender or number of the phrase operating as referent of the pronoun (this is the case of an answer such as *Aquesta nina, no t'oblidis de comprar-**la** a les nenes* ('This doll (fem.), don't forget to by it (fem.) for the girls') to the sentence *Este muñeco, no te olvides de comprárse**lo** a los niños* ('This doll (masc.), ...') –which shows the informant expected information referring to the third person accusative pronoun of the masculine singular and not of the feminine).

c) When, according to the question, the informant expects that the proclitic pronouns appear in prevocalic position, but instead they appear in preconsonantal position, or the other way round (for example, *Sempre **em** vesteixo quan surto de casa* 'I always dress up when I leave home' as an answer to *Siempre **me** arreglo al salir de casa*).

d) When in enclisis cases the pronouns appear postponed to a different verb form to the one expected according to the question (for example, *Renteu-**me** el cotxe*, lit. 'Wash the car for me', as answer to the question *Láva**me** el coche*).

II.3.1 CRITERIA OF ORTHOGRAPHIC TRANSCRIPTION

In the orthographic transcription, as far as morphology is concerned, the dialectal realizations of verb and noun inflection are maintained (for example, *cridos* for *cridis* 'you yell (subjunctive)', *faigues* for *facis* 'you do (subjunctive)', etc.). We also maintained the dialectal forms of the masculine definite article *lo*, *los* and the different forms of the *salat* article (i.e. the one derived from Latin IPSE, IPSA, etc.). The clitic pronouns which precede a verb beginning with consonant are transcribed orthographically with the reinforced form. We did not collect some pronunciations of verb and clitic sequences such as *gità's* or *vendre-lo*, which are transcribed orthographically as *gitar-se* 'to lie down' and *vendre'l* 'to sell it (masc.)'.

As to the lexical aspects, we transcribed dialectalisms (for example, *sàller* and *sallir* from the Roussillon for *sortir* 'to go out') and interferences with other languages. In this last case, we decided to transcribe the forms from French or Italian in agreement with the spelling of these languages (for example, *voiture* for *cotxe* 'car' or *assai* for *molt* 'very'); however, the forms of Spanish origin are transcribed with the spelling adapted to the conventions of Catalan (for example, *unya* for *ungla* 'nail'). The formal variants (such as *cotxo*, *aigo-aiga-auia* 'water', *lego* for *luego* 'afterwards', *aülla* 'needle', *llavò-llavòs-llavons-llavòrens-llavonses* 'then') are transcribed by only one form, the one which appears on the DIEC, to make the search easier.

II.3.2 Criteria for phonetic transcription

The phonic sequence is divided into a series of independent boxes. In the case of elisions or fusion of elements which belong to different boxes, we had to choose to assign a single element to one of the two boxes. We detail below the criteria we applied in these cases.

We always gave priority to the sequence formed by the verb and the clitic. Thus, in a sequence such as [káðə ðí **ə**m réntu lə káɾə] (*Cada dia em rento la cara* lit. 'Every day I myself wash the face'), the second neutral vowel, although it can belong both to the noun *dia* and to the first person weak pronoun, was assigned to the latter. The same principle is applied when there is a coincidence of the final vowel of the enclitic pronoun and the initial vowel of the sequence of following words. In a sequence such as *La bicicleta* [kómpɾalil**a** le tʃikétes] (*compra-li-la a les xiquetes*) 'The bicycle (fem.), buy it (fem.) to the girls', the second *a* is transcribed in the box which corresponds to the final vowel of the pronoun and not in the one of the following words.[7]

As a result of the segmentation of the phonic chain in different boxes, it can happen to have a cell where the initial word begins with a sequence of sounds which does not form a syllable structure accepted in Catalan. This is the case of the sequence [kwán ed rénto **st**áz miʎó] (*Quan et rento estàs millor* 'When I wash you, you feel better'). It is also possible to have the opposite case, that is, that the sequence in a box ends with a syllable structure which is not accepted in Catalan, for example [sém**pɾ** əz réntə ləz ðéns̩] (*Sempre es renta les dents* 'S/he always brushes the teeth').

In the boxes for preceding words and following words, the phonic chain is sequenced in lexical worlds, each of which is transcribed with a primary accent. In some cases, certain orthographic criteria were adopted; thus, the proclitics have been transcribed separate from the verb they accompany (for example, [əm réntu] for *em rento* 'I wash myself'). The enclitic pronouns, though, have been transcribed together with the verb they accompany (for example, [réntəm] for *renta'm* 'wash me').

[7] This criterion was followed in all databases which contain the phonetic transcription of the words preceding or following the word studied.

Moreover, within one box we can find contacts of identical vowels which are solved by elision of one of the vowels. When this happens, if one of the vowels is lexical and the other epenthetic, the lexical vowel is transcribed. When the two vowels are lexical, we can have several possibilities. If one is stressed and the other unstressed, the stressed vowel is transcribed, for example, [akét ɔ́m **é**z el métʃe] (*Aquest home és el metge?* 'Is this man the doctor?'). When the two vowels are unstressed, priority is given to the integrity of the word to the right (for example, [m aɣɾáð **a**ɾiβáɾ] for *m'agrada arribar* 'I like to come'), except in the case when only the word on the left belongs to one of the major lexical categories (for example, [kómpɾ**ə** m tərʒɛ́tə] for *compra amb targeta* 's/he buys with (credit) card').

In the case of consonant contact, we always transcribed the consonant in the position of open syllable (for example, [mé **s**uβín] for *més sovint* 'more often'; [reunjón **s**émpɾe] for *reunions sempre* 'meetings always').

II. 4 THE ARTICLES DATABASE

This database consists of 24 records which correspond to the forms of the definite article, the personal article and the masculine definite article preceded by the prepositions *per* 'for/by', *amb* 'with', *a* 'to' and *de* 'of'. The first 16 records, which include the forms of the definite article and the personal article, are common to all informants; the remaining 8 records, which correspond to the forms of the masculine definite article preceded by the prepositions *per*, *amb*, *a* and *de*, are only included in the case of Algherese and Balearic.

As to the contexts of occurrence, both the definite article and the personal article were included followed by word starting by vowel and by consonant. In the case of the singular feminine definite article we also considered, specifically, the contexts where it precedes a word starting by the unstressed *i* or *u* vowels and stressed *i* or *u* vowels.

The answers were obtained through the translation into Catalan of sequences which contained the articles in Spanish –for Balearic, Central, North-Western and Valencian Catalan– in French –for Roussillon– and in Italian –for Algherese.

Apart from the boxes common to all databases (INFORM, LOCAL and ID), this database is organized into the following boxes:

MOT_ANT: contains the phonetic transcription of the sequence of words (if any) preceding the article of the preposition + article contraction.

CONS_ANT: contains the phonetic transcription of the initial consonant of the prepositions *per* and *de*, when they make a contraction with the singular or plural masculine definite article.

VOC_ANT: contains the phonetic transcription of the vowel (if any) preceding the consonant of the article.

ART: contains the phonetic transcription of the consonant of the article.[8]

VOC_FINAL: contains the phonetic transcription of the vowel (if any) following the consonant of the article.

PLURAL: contains the phonetic transcription of the consonant of the plural morph (if any) of the article. In the case of the plural masculine article *es*, we decided to transcribe the final fricative in the ART box and not in the PLURAL box.

MOT_POST: contains the phonetic transcription of the sequence of words (if any) following the article.

SENCER: contains the phonetic transcription of the article or preposition + article contraction.

OBSERVACIONS: contains various data. We transcribed answers which do not provide the information corresponding to the question asked, alternatives proposed by the informant, comments referring to the hesitation between two forms, mistakes made by the informant, etc.

In order to make the search easier, the database contains a box with the orthographic form corresponding to the Central standard of the definite article, personal article or masculine definite article preceded by the prepositions *per*, *amb*, *a* and *de*, and the word following them. It also has a series of boxes with morphological information on gender and number (MASC, FEM, SING and PL).

II. 5 THE POSSESSIVES DATABASE

[8] In this box, in the case of Pollença, it is possible to find vowel [u], characteristic of the article in this variety.

This database gathers the different forms the possessives take when used preceding the noun they determine. We gathered both the forms of the stressed possessives –of one possessor and several possessors– and the forms of unstressed possessives, although the latter are not systematically asked from all informants. Moreover, third person possessives which indicate more than one possessor *llur*, *llura*, *llurs* and *llures* were only asked in the Roussillon dialect. In all cases, the method used to obtain the data was the translation of a series of short sequences which contained the corresponding forms of the possessives.

Apart from the boxes common to all databases (INFORM, LOCALITAT and ID), this database is organized into the following boxes:

VOC_ANT: contains the phonetic transcription of the vowel (if any) preceding the consonant of the article which appears before the possessive.

ART: contains the phonetic transcription of the consonant of the article.

VOC_FIN1: contains the phonetic transcription of the vowel (if any) following the consonant of the article.

PLURAL1: contains the phonetic transcription of the consonant of the plural morph (if any) of the article.

POSS: contains the phonetic transcription of the root of the possessive.

VOC_FIN2: contains the phonetic transcription of the vowel corresponding to the epentheses or inflection morphs (if any) of the possessive.

PLURAL2: contains the phonetic transcription of the consonant of the plural morph (if any) of the possessive.

MOT_POST: contains the phonetic transcription of the noun pronounced after the possessive.

SENCER: contains the phonetic transcription of the possessive.

OBSERVACIONS: contains different information, such as alternative forms proposed by the informants, unexpected answers or comments on the use of a form.

To make the search easier, the database contains a box with the orthographic form corresponding to the Central standard of the possessive and of the article preceding it. It also has a series of boxes which indicate the gender and number of the possessive (MASC, FEM, SING and PLUR).

II.6 THE PERSONAL PRONOUN DATABASE

This database contains eight records for each informant, which correspond to the singular and plural first, second and third person personal pronouns (masculine and feminine). The data were obtained from the translation into Catalan of a series of short sequences which contained the forms corresponding to the personal pronouns.

The personal pronoun database is headed by the boxes common to all the databases (INFORM, LOCAL and ID). The remaining boxes are presented below:
T_FON: contains the phonetic transcription of the personal pronoun.
MOT_POST: contains the phonetic transcription of the sequence of words following the personal pronoun.
OBSERVACIONS: includes different information, such as, among others, alternative forms proposed by the informant or forms which do not correspond to the expected answer.

To make the search easier, this database contains a box with the orthographic form of the personal pronoun corresponding to the Central standard. It also has a series of boxes with morphological information on gender, number and person (SING, MASC, FEM, PL, PERS).

II. 7 THE NEUTRAL DEMONSTRATIVES DATABASE

This database contains three records for each informant, which correspond to the three possible degrees of proximity of the neutral demonstratives, represented by the forms *açò* (1st degree), *això* (2nd degree) and *allò* (3rd degree). Data were obtained from the translation into Catalan of a series of short sequences which contained the forms corresponding to the neutral demonstratives.

Apart from the boxes common to all the databases (INFORM, LOCAL and ID), this database has the following boxes:
MOT_ANTERIOR: contains the phonetic transcription of the word preceding the demonstrative.
FONÈTICA: contains the phonetic transcription of the demonstrative.

OBSERVACIONS: includes, among others, alternative forms or forms which do not correspond to the expected answer.

To make the search easier, this database has a box which indicates the degree of proximity of the neutral demonstrative (1, 2 or 3).

II. 8 THE LOCATIVE DATABASE

This database contains three records for each informant, which correspond to the three possible degrees of proximity of locatives 'here/there', represented by the forms *ací* (1st degree), *aquí* (2nd degree) and *allí* / *allà* (3rd degree). The data were obtained from the translation into Catalan of a series of short sequences which contained the forms corresponding to the locatives.

Apart from the boxes common to all the databases (INFORM, LOCAL and ID), this database has the following boxes:

MOT_ANTERIOR: contains the phonetic transcription of the word preceding the locative.

TRANS_FON: contains the phonetic transcription of the locative.

OBSERVACIONS: contains different kinds of information, such as, among others, alternative forms proposed by the informant or forms which do not correspond to the expected answer.

To make the search easier, this database has a box which indicates the degree of proximity of the locative (1, 2 or 3).