Proposal for a morphosyntactic coding of Catalan for CHILDES*

Anna Gavarró
(Universitat Autònoma de Barcelona)

Given that all Catalan files in CHILDES (MacWhinney and Snow 1985, MacWhinney 1991, 1995) are uncoded, and that a growing body of data is available for the acquisition of that language and will presumably be made available through CHILDES, it is important to procure a unified system of morphosyntactic coding, to facilitate the use of our data. The system proposed should be user-friendly both for the person codifying and for the consultant, that is, as simple and transparent as possible, and consistent. Leaving aside the particular markings of each code, the system proposed here for Catalan easily extends to Castillian Spanish, with major changes only being necessary for the pronoun code.

The proposal made here consists of the following. First, the way in which morphosyntactic information is codified in CHILDES is briefly summarised; general constraints and conventions of the system are specified. Second, specialised codes for Catalan and the markedness conventions for this language are identified. Finally, they are exemplified with early language acquisition data. In the appendix, the abbreviations used are indexed alphabetically to facilitate their use.

1. General principles of morphosyntactic codification

As generally proposed in the CHILDES project, the morphosyntactic coding is included in two specific lines, indicated %syn and %mor, for syntax and morphology respectively. The elements in these lines correspond one-to-one with words; we use upper case for the %syn line and lower case for the %mor line, although the system is not case-sensitive. Further information that may be morphosyntactic in nature may appear in the error coding line, %err, if the forms produced by the child are deviant from the target with respect to morphosyntax.

Syntactic structure is thus encoded in the %syn line, while the %mor line includes the remaining morphosyntactic information: the part of speech each constituent belongs to, the presence of clitics and affixes, and the syntactic features that modify each part of

*The author acknowledges the financial suport of the Ministerio de Educación y Cultura through the program of Incorporación de doctores to the project PB-96-1199-C04 at the UAB.

speech. The basic scheme for coding words on the %mor line, as stated in MacWhinney (1991, 1995), is then:

(1) part-of-speech|~clitic#prefix$stem=english+compound&suffix-suffix~clitic

with clitics and suffixes presented in the same position they occupy with respect to the stem. The clitic position is going to be occupied by unstressed personal pronouns and determiners in Catalan; as devised by MacWhinney (1991, 1995), part of speech is repeated after a clitic, as in the Italian example (2). It is possible, although not compulsory, for the transcriber to indicate the English translation of the stem, as also exemplified in (2).

(2) parlagli
v|parl=speak&imp:2s~pro|dat:masc:sg
(from MacWhinney 1995:106)

Since the %mor line is based on the speaker's target, any incorrect forms should have their target clarified in the %err line. Incorrect forms should be followed by the [*] sign both in the data and in the %mor line, and omissions should be signaled 0 in the %mor line only. (That omissions should be signaled represents a problem for the transcriber, as the postulation of empty elements is not always straightforward, nor is it theory-neutral; furthermore, it implies the appeal to adult grammar in ignorance of the principles that child grammar may have. The inclusion of omission marks is nevertheless considered necessary in the CHILDES protocol.)

Language-specific markedness conventions can be set up so that zero morphs need not be rendered in the %mor line.

The markedness conventions followed, as well as the full list of grammatical morphemes used in the codification of a language, must be attached to the corpus files.

2. Codes for the transcription of Catalan

The codes proposed in this section seem sufficient to codify Catalan; they may not be sufficient for the purposes of some trasncribers, in which case new codes may be added; their justification rests with their proponents.

The syntactic primitives (to appear in the %syn line) proposed to codify Catalan are those in (3), and are a subset of those proposed by MacWhinney (1991, 1995) for

English. The codes of MacWhinney for modifier, relativiser/inf and main clause have been left out: appositive phrase and adverbial adjunct cover the cases of modification; the code relativiser/inf seems to be used by MacWhinney for the English of *to* infinitivals and is therefore not necessary; indication of main clause is also dispensable. The code auxiliary ("x") is also left out to have auxiliaries codified syntactically as verbs (as the common morphology of main verbs and auxiliaries in Catalan suggests); this step is also consistent with the existence of a part of speech coding v:aux (MacWhinney 1995: 105), which pressuposes that auxiliaries constitute a class of verbs.

(3)

s	subject
v	verb
d	direct object
i	indirect object
c	conjunction
p	preposition
a	adverbial adjunct
ap	appositive phrase
rc	relative clause
cc	coordinate clause
cp	complement
pp	prepositional phrase

In the %mor line, the part of speech codes proposed for Catalan are those in (4). They coincide with the word classes selected by Quirk et al. (1985: 67) except for the addition of quantifier, communicator (used for expressions such as *hello* or *aw!*) and wh-word (included by Quirk et al. amongst pronouns). To these, a code for complementiser is added. The category for modal verbs included by MacWhinney (1991, 1995) is not necessary for Catalan, as Catalan modal verbs do not constitute a class contraposed to that of verbs. The remaining codes included by MacWhinney and excluded here are: the code for the infinitive marker *to*, proper noun, number and particle (proper noun and number are included below as codes for the nominal categories only).

(4)

adj	adjective
adv	adverb
n	noun
v	verb
conj	conjunction

det	determiner
prep	preposition
pro	pronoun
quant	quantifier
v:aux	auxiliary verb
wh	wh-word
comp	complementiser
co	communicator

Further features needed to codify Catalan vary depending on syntactic category. For nouns, those in (5) may be needed, for verbs and auxiliaries those in (6)[1], for adjectives those in (7), for pronouns those in (8) and for determiners those in (9).

(5)

prop	proper	n
cmn	common	
sg	singular	
pl	plural	

(6)

1s	first person singular	v
2s	second person singular	
3s	third person singular	
1p	first person plural	
2p	second person plural	
3p	third person plural	
pres	present	
past	past	
fut	future	
part	participle	
ger	gerund	
inf	infinitive	
imp	imperative	
in	indicative	
cond	conditional	
subjv	subjunctive	
perf	perfect	

[1]It may be noted that the proposal of MacWhinney (1991, 1995) includes codes such as "1s" for first person singular side by side with "sg" for the singular of a noun. This means that "singular" is not consistently represented (nor is plural), and that the features of person and number that modify verbs are not presented independently, as would be preferable. We mantain this shortcoming of the system for consistency with the codes for other languages.

	impf	imperfect	
	1c	first conjugation	
	2c	second conjugation	
	3c	third conjugation	
(7)	sg	singular	adj
	pl	plural	
	masc	masculine	
	fem	feminine	
	poss	possessive	
	1	first person	
	2	second person	
	3	third person	
(8)	sg	singular	pro
	pl	plural	
	masc	masculine	
	fem	feminine	
	1	first person	
	2	second person	
	3	third person	
	refl	reflexive	
	nom	nominative	
	acc	accusative	
	dat	dative	
	abl	ablative	
	all	allative	
	imprs	impersonal	
	prs	personal	
	loc	locative	
	partit	partitive	
	indef	indefinite	
(9)	sg	singular	det
	pl	plural	
	masc	masculine	
	fem	feminine	
	def	definite	
	indef	indefinite	

Drawing on the codes for grammatical morphemes of MacWhinney (1991, 1995), we propose the following specialised codes for Catalan:

(10) **Markedness conventions**

nouns	singular, masculine
adjectives	singular, masculine
pronouns	singular, masculine
determiners	singular, masculine
participles	singular, masculine
verbs	3s[2]

(11) **Nominal markings** (for n and a)[3]

-s	pl
-a/-e	fem

Verbal markings

-r	inf
-nt	ger
-t	part:masc:sg
-da	part:fem:sg
-ts	part:masc:pl
-des	part:fem:pl
-ria/-rie	cond
-ré	fut:1s
-rà	fut
-re	fut:2p or fut:1p
-i	subjc:s or subjc:3p
etc.	

[2]Although we might choose e.g. indicative as the unmarked mode, or present as the unmarked tense, we are not going to pursue that line, since they are not morphologically unmarked, at least not for the whole paradigm. Notice that the notion of markedness in the CHILDES protocol has to do with the absence of phonological material, not with the existence of default forms. As a consequence, linguistically relevant formulations of markedness are ignored here. For example, Oltra (1999) analyses the verbal paradigm of Catalan in the framework of distributed morphology and considers the first conjugation as unmarked with respect to the rest; this formulation easily accommodates for the fact that children's errors often involve misclassification of a verb to the first conjugation (*caurat* for *caigut* 'fallen', Joan 2:6); this information is not reflected in the transcription.

[3]This is done on the basis of orthography, which obscures some generalisations (e.g. -a/-e are in many dialects of Catalan different orthographical renderings of the same unit).

(Given the complexity of the Catalan verbal system, and the existing number of irregular verbs, the reader is referred to Mascaró 1986 for a full paradigm of verbal markings.)

Pronoun markings

hi	loc or all
ho	acc:indef:3
els/ls/los	masc:acc:pl or dat:pl
l/el	masc:acc:sg
la	fem:acc:sg
les	fem:acc:pl
li	dat:sg:3
l'hi	dat:sg:3&acc:sg:3
m/em/me	acc:sg:1 or refl:sg:1
t/et/te	acc:sg:2 or refl:sg:2
ens/nos	acc:pl:1 or dat:pl:1 or refl:pl:1
us/vos	acc:pl:2 or dat:pl:2 or refl:pl:2
n/en/ne	partit or abl
s/es/se	refl:3 or imprs

Determiner markings

l/el	def
un	indef
-s	pl
-a/-e	fem

3. Examples

The use of the codes proposed is briefly exemplified with data from the corpus of the speech of my son Joan:

(12)

*JOA:	pota [*] tancada cau [*]
%syn:	S M PP
%mor:	det|*0 n|pota=door v|*0 adj|tancada=closed-fem p|*0 det|*0 n|cau=key
%err:	pota = porta $pho $los; cau = clau $pho $los
*JOA:	això mament [*]

%syn: S AP
%mor: pro|això=this&dem:3s v|*0 adv|mament=badly
%err: mament = malament $pho $los

*JOA: una senyora té barret
%syn: < S <V D> [RC]>
%mor: det|un=a&indef-fem n|senyora=lady comp|*0 v|te=have&pres n|barret=hat

Appendix: Index of codes

a	adverbial adjunct
abl	ablative
acc	accusative
adj	adjective
adv	adverb
all	allative
ap	appositive phrase
c	conjunction[4]
cc	coordinate clause
co	communicator
comp	complementiser
cmn	common
cond	conditional
conj	conjunction
cp	complement
d	direct object
dat	dative
def	definite
det	determiner
fem	feminine
fut	future
ger	gerund
i	indirect object
imp	imperative
impf	imperfect

[4] As it is to be codified in the %syn line; in the %mor line it appears as “conj”.

imprs	impersonal
in	indicative
indef	indefinite
inf	infinitive
int	interrogative
loc	locative
masc	masculine
n	noun
nom	nominative
p	preposition[5]
part	participle
partit	partitive
past	past
perf	perfect
pl	plural
poss	possessive
pp	prepositional phrase
prep	preposition
pres	present
prs	personal
pro	pronoun
prop	proper
quant	quantifier
refl	reflexive
rc	relative clause
s	subject
sg	singular
subjv	subjunctive
v	verb
wh	wh-word
1	first person
2	second person
3	third person
1c	first conjugation
2c	second conjugation
3c	third conjugation
1p	first person plural

[5] As it is to be codified in the %syn line; in the %mor line it appears as “prep”.

2p	second person plural
3p	third person plural
1s	first person singular
2s	second person singular
3s	third person singular

References

MacWhinney, B. (1991) *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey, Hove and London. Second edition, 1995.

MacWhinney, B. and C. Snow (1990) 'The Child Language Data Exchange System: An update. *Journal of Child Language 17*, 457-472.

Mascaró, J. (1986) *Morfologia*. Enciclopèdia Catalana, Barcelona.

Oltra, I. (1999) 'On the constituent structure of Catalan verbs'. *MIT Working Papers in Linguistics*, 33. Eds. K. Arregi, B. Bruening, C. Krause and V. Lin.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1991) *A Comprehensive Grammar of the English Language*. Longman, London and New York.

8th March, 2000

Anna Gavarró
Departament de Filologia Catalana
Universitat Autònoma de Barcelona
Facultat de Lletres, Edifici B
08193 Bellaterra
agavarro@seneca.uab.es