Indicators of lexical growth throughout age, genre and modality for a Catalan L1 corpus

Chapter · November 2014

CITATION
1

READS
90

4 authors, including:

Laia Cutillas i Alberich
University of Barcelona
11 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Liliana Tolchinsky
University of Barcelona
87 PUBLICATIONS **1,228** CITATIONS
SEE PROFILE

Elisa Rosado
University of Barcelona
23 PUBLICATIONS **206** CITATIONS
SEE PROFILE

LAIA CUTILLAS, LILIANA TOLCHINSKY, ELISA ROSADO, JOAN PERERA

Indicators of lexical growth throughout age, genre and modality for a Catalan L1 corpus

Abstract

Lexical development is a key facet of later, school-age language development. The goal of the study is twofold: on the one hand, to describe quantitatively the text-embedded lexicon of a corpus of texts produced by informants from nine years of age to adulthood, and on the other hand, to identify which characteristics of the lexicon can be considered as indicators of age, discourse genre, and modality of production differentiation.

The GRERLI-CAT1 (*Grup de Recerca per a l'Estudi del Repertori Lingüístic, Català L1*) corpus is constituted by a set of narrative and expository spoken and written texts in Catalan that were produced by 79 bilingual Catalan/Spanish speakers that have Catalan as their home language. They were distributed in five groups according to level of education: elementary, secondary, high school and university level, and language teachers. The corpus comprises 316 texts, which include a total of 84,081 tokens, 40,612 types and 31,811 lemmas.

Four measures were applied for characterizing the corpus lexically: (1) lexical diversity, (2) lexical density, (3) word length, and (4) productivity of verbs. Lexical diversity was selected to gauge the range of vocabulary displayed in the texts, lexical density as an indicator of textual richness and informativeness, word length was taken as an indicator of lexical complexity, and productivity of verbs as an indicator of lexical richness within the verbal domain. We discuss the utility of these four measures as indicators of text construction development in different genres and modalities.

1. Introduction

The general framework of this study is Corpus Linguistics, which approaches the study of language empirically. It is based on authentic samples of language use, so data treatment is external to the speaker, directly observable and, therefore, demonstrable. It uses computerized corpora, which represent an important source of quantitative information. The use of these samples enables to determine frequency of occurrences and lack of specific linguistic elements (Civit 2003: 2). Corpus linguistics is a method applicable to every area of linguistic research, from phonology to discourse. Moreover, corpus-based research is applicable to linguistic education both in first (L1) and second (L2) language learning. Native corpora, like the one used in this study, can be very useful for education professionals because they show what speakers say or write in a specific communication situation, as well as their typical difficulties (Nesselhauf 2004: 144-145).

In this study, this methodology is applied to first language acquisition and, more specifically, to the later, school age language development of Catalan. Catalan is a Romance language that typically displays a rich morphological system: nouns and adjectives are inflected for gender and number; pronouns are inflected for gender, number, person and case; and verbs are inflected for person/number, tense/aspect and mood. Moreover, it requires gender and number agreement between noun and adjective and between pronoun and its antecedent, as well as person and number agreement between subject and verb.[1]

1 Catalan is spoken in four Spanish autonomous communities (Aragon, the Balearic Islands, Catalonia and the Valencian Community), in the French region of Rousillon, and in the city of Alghero, on the Italian island of Sardinia. In Catalonia, Catalan and Spanish are equally recognized as official language, but Catalan is the language of education. Therefore, children educated in Catalonia are bilingual (Catalan/Spanish) living in a multilingual environment: "Due to a major surge of immigration over the past decade (3% in 2000 to 13% in 2008), an increasing percentage of children speak a language different from Catalan and Spanish at home" (Llauradó / Tolchinsky in press).

Several Catalan corpora have been compiled during the recent years. The CCCUB (*Corpus del Català Contemporani de la UB*) (Boix et al. 2002), compiled by the *Grup d'Estudi de la Variació* (GEV) at the *Universitat de Barcelona* (UB), contains texts in different dialects and registers, but only in one modality of production, the spoken one. The AnCora-CA (Anotated Corpus-Catalan) (Taulé et al. 2008) includes written journalistic texts, so it could only be assessed one genre and one modality. The CesCa (*Català Escolar Escrit a Catalunya*) (Llauradó et al. 2012) contains different types of written texts: narrations of a film storyline, recommendations of a film, or definitions of words and jokes, but only in one modality of production. These two latter corpora are compiled by the *Centre de Llenguatge i Computació* (CLiC) of the UB. The CICA (*Corpus Informatitzat del Català Antic*) (Torruella 2009), compiled by the *Grup de Lexicografia i Diacronia* (SFI) of the *Universitat Autònoma de Barcelona* (UAB), is a diachronic corpus which includes books written from the 12th to the 15th centuries. The CTILC (*Corpus Textual Informatitzat de la Llengua Catalana*) (Rafel / Solanellas 1986), compiled by the *Institut d'Estudis Catalans* (IEC) with lexicographic purposes, contains both literary and non- literary texts in different genres but, again, only in the written modality. The GRERLI-CAT1 corpus, compiled by the GRERLI (*Grup de Recerca per a l'Estudi del Repertori Lingüístic*) at the UB, includes texts in different genres (narrative and expository) and modalities of production (spoken and written), produced by subjects of different school levels (elementary, secondary and high school, university level and language teachers).

The goals of this paper are (1) to analyse quantitatively the GRERLI-CAT1 corpus text-embedded lexicon and (2) to identify which characteristics of the lexicon can be considered as indicators of lexical growth throughout age, genre, and modality. With this purpose, four measures are evaluated: lexical diversity, lexical density, word length and productivity of verbs.

Lexical diversity (Malvern et al. 2004: 19) is selected because it can gauge the range of vocabulary use. Lexical density (Read 2000: 200) is considered a good indicator of textual richness and informativeness. Word length (Strömqvist et al. 2002: 48), is chosen

as a measure of lexical complexity. Productivity of verbs is selected as an indicator of lexical richness within the verbal domain. This measure embraces both the diversity of verbal lemmas, that is, the number of different verbal lemmas used in the corpus, and the productive use of inflectional morphology of verbs, that is, the number of different tense forms (Serrat et al. 2004: 221).

2. Corpus compilation

The GRERLI-CAT1 corpus was compiled within the framework of a cross-linguistic project designed to analyse the development of text construction abilities in different languages. The corpus was compiled in 1998 as part of the international project "Developing literacy in different contexts and in different languages", Spencer Foundation, Chicago, United States of America (P.I.: R. A. Berman, Tel Aviv University). The languages that took part in this study were Dutch, US English, French, Hebrew, Icelandic, Spanish, and Swedish, with the subsequent addition of Catalan. The main goals of this project were to understand how speakers/writers develop their discursive abilities throughout different educational levels, to analyse how they use the resources of their respective languages to construct discourse in different genres (expository and narrative) and modalities of production (spoken and written), and, finally, to find out common and language-specific patterns of language use when producing spoken and written texts (Berman 2002; Aparici 2010).

Language usage, both spoken and written, is always framed within a specific discourse genre. As Tolchinsky (2004: 235) says:

> There is no such thing as neutral use of language: people constantly attune their speech to specific intentions, purposes, and interlocutors. Thus, development is viewed in terms of the acquisition of different discourse genres, and the way that the cultural conventions of genres constrain the use of linguistic forms.

The two genres selected for the project were expository and narrative. Several reasons lead to the selection of these two genres. Firstly, because the narrative genre has been vastly studied from a developmental psycholinguistic point of view while the expository genre has not, since it is a late developing genre, and secondly, because of the contrast in personal involvement that characterises these two genres. A first person narrative and the discussion of a topic illustrate the extremes of a continuum of personal involvement. Personal involvement of the speaker/writer is demonstrated, among other things, in the election of the main characters. In personal narratives the main character is the speaker/writer and, in expository texts, the topic gains prominence (Aparici 2010).

Regarding modalities of production, spoken and written texts, differ from the point of view of the demands imposed by processing. The spoken modality is constrained by on-line processing and can cause mistakes, breaks, repetitions, reformulations or lack of referents. On the other hand, the written modality is not subject to time limitations. It enables the writer to plan thoroughly what he or she wants to write, to outline or to proofread. For this reason, written texts are prone to be better organized and be more cohesive and coherent than spoken texts.

Informants were selected from nine years of age onwards because at this age they have already attained a command of narrative structure and of the mechanisms of discursive cohesion (Berman / Slobin 1994). The sample also included adults because the development of discursive abilities is a long process that spans from childhood to beyond adolescence (Berman / Verhoeven 2002: 14).

2.1. Informants

The texts were produced by 79 informants from Barcelona, who were bilingual speakers of Catalan and Spanish and whose home language

is Catalan.[2] Participants were distributed in five groups, according to their educational level and age. The elementary school group comprises 20 students from the 4th year elementary school. The secondary school group comprises 19 students of the 1st year secondary school. The high school group comprises 20 students of the 1st year high school. The university group comprises 10 students from different courses and specialities, specifically, four Science students and six Humanities students. Finally, the teachers group comprises 10 high school Catalan language teachers. This group is expected to represent a level of professional use of the language.

Table 1 shows data of all informants' groups:

Elementary School		*Secondary School*		*High School*		*University*		*Teachers*	
n = 20		*n* = 19		*n* = 20		*n* = 10		*n* = 10	
M	Range	*M*	Range	*M*	Range	*M*	Range	*M*	Range
10;3.	9;6. - 10;9.	13;0.	12;4. - 13;4.	17;4.	16;3. - 18;10.	21;9.	19;1. - 24;2.	43;10.	37;8. - 54;11.

Table 1. Number of informants, mean age and age range.

2.2. Tasks

All informants produced four texts. After watching a three-minute video without text, participants were asked to produce a spoken expository text, a written expository text, a spoken narrative text and a written narrative text. The video shows different conflictive situations in schools, such as fights, shunning classmates, cheating in exams, etc. The purpose was that all texts produced by the informants had a common topic to enable the comparison of the text linguistic characteristics.

2 It is unlikely to find monolingual speakers of Catalan, since Catalan and Spanish are both official languages in Catalonia. All children use Catalan at school and Spanish is massively present both in the media and in social settings.

2.3. Procedure

Informants produced the four tasks individually. Data were elicited in two sessions and four different orders of text production were followed: A (first session: spoken narrative/written narrative; second session: spoken expository/written expository); B (first session: written narrative/spoken narrative; second session: written expository/spoken expository); C (first session: spoken expository/written expository; second session: spoken narrative/written narrative); and D (first session: written expository/spoken expository; second session: written narrative/spoken narrative).

2.4. Corpus storage

We have used two different versions of the corpus: clean (*net*) and morphologically tagged (*morfo*). In the following sub-sections, the characteristics of these two corpus versions will be described.

2.4.1. Clean version (net)

This version consists of text files without format which contain the production with no additional symbols or indications of spelling mistakes. All marking of ancillary material, such as repetitions, reformulations, pauses and comments, is omitted. The transcription unit is the clause. That is, each text line in this version corresponds to a clause. We follow Berman and Slobin's (1994: 660) definition of this unit:

> We define a clause as any unit that contains a unified predicate. By unified, we mean a predicate that expresses a single situation (activity, event, state). Predicates include finite and non-finite verbs, as well as predicate adjectives.

This version aims at producing an input file for the morphological analyser HSMorfo, which performs the morphological tagging. The counts of tokens and types were computed using the *net* version.

2.4.2. Morphologically-tagged version (morfo)

This is a text version that results from the *net* version of the corpus after having been morphologically tagged using the HSMorfo Linux software.[3] Files of the morphologically analysed texts contain three columns: the first one represents the type, the second one the lemma and the third one includes the EAGLES (Expert Advisory Group on Language Engineering Standards) tag (EAGLES 1996), which shows the grammatical category of the type. The measures for the four analysis dimensions of the current paper were calculated using the *morfo* version.

2.4.3. Other versions

There are three more versions of the corpus: mirror (*rep*), normalized (*nor*) and CHAT (*cha*). The mirror version (only for writing texts) consists in the transcription of the written productions into a MS Word file, with the original disposition of paragraphs and lines, punctuation and spelling. In the normalized version, texts (both spoken and written) are transcribed in CHAT (Codes for the Human Analysis and Transcripts) format (MacWhinney 2012). Spoken productions are transcribed orthographically (not phonetically) including processing information (pauses, repetitions, reformulations, etc.). In written texts, following CHAT conventions, spelling mistakes are followed by the correct word (e.g. *vastant* [: *bastant*] 'enough'). Finally, the CHAT version only differs from the *nor* version in that the transcription unit is the clause.

[3] It is important to remark that some files have been analysed using the FreeLing software, because they included a character which provoked a malfunctioning on HSMorfo and it finished the analysis before arriving to the end. However, the working method of both programs is the same, because they use the same analysis process, as well as EAGLES tags on the tagging process. They only differ in the treatment of the apostrophized words (FreeLing attaches them to the next word and HSMorfo separates them) and in the contractions (FreeLing separates them into two words and HSMorfo treats them as one word). These treatments have been subsequently fixed using a software specially designed to correct these errors.

2.5. Corpus processing

The *net* version provides the input files for the HSMorfo Linux software, which performs the morphological analysis and tagging, the *morfo* version of the corpus. The HSMorfo software uses the tags developed by the EAGLES group for morphosyntactic annotation of European languages, which are adapted to Catalan in Civit (2003). Each position of the tag corresponds to a feature (whose number varies for each category), the value of this feature and a code representing them. Zero shows that a given feature has no value.

Below are some examples of the morphological analysis of three specific cases of segmentation, which could be a problem for morphological analysers:

- Multi word expressions, linked by an underscore (_), such as *punt de vista* 'point of view':
  punt de vista ⇨
  1. *punt_de_vista punt_de_vista* ('point of view') NCMS000 (Noun Common Masculine Singular)
- Words containing a dash (-) and/or an apostrophe ('):
  a) Words with clitic pronouns, which are attached to a verb using a dash and/or an apostrophe, such as *controla'ls-ho* 'control it to them':
  controla'ls-ho ⇨
  1. *controla controlar* ('to control') VMM02S0 (Verb Main Imperative 2nd Singular)
  2. *'ls ell* ('he') PP3CP000 (Personal Pronoun 3rd Common Plural)
  3. *-ho ho* ('it') PP3NN000 (Personal Pronoun 3rd Neuter Invariable)

  b) Apostrophized words, such as *l'únic* 'the unique':
  l'únic ⇨
  1. *l' el* ('the') DA0CS0 (Determinate Article Common Singular)
  2. *únic únic* ('unique') AQ0MS0 (Adjective Qualifying Masculine Singular)

- Contractions (prepositions attached to an article) are treated by the analyser as compound prepositions, and they include grammatical information about gender and number. In other words, they present the preposition features (three first values of the tag) and the article features (the two last values of the tag), such as *dels* 'of the':

  dels ⇨

  1. *dels del* ('of the') SPCMP (Adposition Preposition Complex Masculine Plural)

As we have seen in this section, the characteristics of the GRERLI-CAT1 corpus are as follows: it contains 316 texts, which include a total of 84,081 tokens, 40,612 types and 31,811 lemmas. The main features that identify this corpus are: (i) Language: the corpus contains texts in Catalan; (ii) Developmental dimension: the corpus reflects later, school age language development, because it contains data produced by informants from nine years old to adulthood; (iii) Genre: it contains texts in two discourse genres: narrative and expository; (iv) Modality: it includes texts of two modalities of production: spoken and written; and finally, (v) Corpus storage: the corpus is stored in five different versions that can serve as input for different computational platforms. Moreover, because the corpus is part of a cross-linguistic project, in which the same data elicitation procedure was used for the different languages, it can be compared with the corpora in the other languages that took part in the study.

3. Dimensions of analysis

In order to identify the characteristics of the lexicon that can be considered as indicators of age, discourse genre and modality of production differentiation, the selected measures are: (3.1.) lexical diversity, calculated using type-token, lemma-token and lemma-type ratios; (3.2.) lexical density, i.e. the proportion of content words

relative to the total number of words; (3.3.) word length, i.e. the number of letters of each lexical word; and (3.4.) productivity of verbs, characterised by the diversity of verbal lemmas used and the productive use of inflectional morphology of verbs, i.e. the number of verb tenses associated to each verbal lemma.

3.1. Lexical diversity

Lexical diversity (Malvern et al. 2004) is a widely used measure of language development, because it can gauge the range of vocabulary use. Although the type-token ratio is by far the most commonly used measure of lexical diversity, it is much less useful in measuring vocabulary of inflectional languages like Catalan than another measure, the lemma-token ratio, which provides more accurate information about lexical diversity, as Granger / Wynne (2000: 251) say:

> A learner who uses five different forms of the verb go (*go/goes/going/ gone/went*) in one and the same text has a less varied vocabulary than the one who uses five different lemmas (such as *go/come/leave/enter/return*).

Moreover, a parameter that calculates the correlation between lemmas and types, the lemma-type ratio, can also be useful to study lexical diversity of Catalan. The type-token ratio is obtained by dividing the number of types by the number of tokens, the lemma-token ratio by dividing the number of lemmas by the number of tokens; and finally, the lemma-type ratio by dividing the number of lemmas by the number of types.

3.2. Lexical density

Lexical density (Read 2000) is considered to be a good indicator of textual richness. If content words convey the vast bulk of semantic content, then the relative proportion of lexical words used in a text can provide an idea of its informativeness. Lexical density is obtained by

dividing the number of lexical words (nouns, main verbs, adjectives and adverbs) by the total of words in a sample.

3.3. Word length

Word length (Strömqvist et al. 2002) is regarded as an indicator of lexical complexity. The longer words are supposed to be derivationally more complex words, so word length is a suitable measure of lexical complexity. Function words (auxiliary verbs, determinants, pronouns, prepositions, conjunctions and interjections) have been removed from this count because of their grammatical status. Means for word length are obtained by counting the number of letters in each lexical word over the total number of words.

3.4. Productivity of verbs

Productivity of verbs is seen here as an indicator of lexical richness within the verbal domain. In the context of the present study we consider the diversity of verbal lemmas (number of different lexical verbs that appear in the corpus) apart from the productive use of the inflectional verb morphology (number of different verb forms – marking of tense/aspect or mood– for each specific verbal lemma). Diversity of verbal lemmas is obtained by dividing the total of verbal lemmas by the number of different verbal lemmas used in a sample. Productive use of inflectional morphology (Serrat et al. 2004) is calculated by counting the number of different tenses used for each verbal lemma. For these counts, differences in person/number inflection have not been considered, and the non-finite forms of the verb (infinitive, participle and gerund) have only been considered as part of a compound verb. That is, tenses formed by an auxiliary verb (*haver* 'to have' or *anar* 'to go', and *ser* 'to be', for passive forms) plus a non-finite form of the main verb.

4. Results

4.1. Quantitative description of the corpus

Table 2 presents the distribution of tokens, types and lemmas by text type. There are some differences between both the two discourse genres and the two modalities of production. Expository texts are significantly longer than narrative texts, both spoken and written: (F (1, 74) = 19.024, p = .000) for tokens, (F (1, 74) = 22.661, p = .000) types, (F (1, 74) = 24.078, p = .000) and lemmas, respectively. Spoken texts in both genres have significantly more tokens, types and lemmas than written texts: (F (1, 74) = 25.669, p = .000) for tokens, (F (1, 74) = 12.080, p = .001) types, (F (1, 74) = 12.556, p = .001) and lemmas, respectively. By text types, spoken expository has more tokens than spoken narrative and written expository than written narrative. Types and lemmas present a different pattern: expository texts, both spoken and written, have more types and lemmas than narrative ones. A significant interaction between genre and modality is also found for tokens (F (1, 74) = 11.423, p = .001), types (F (1, 74) = 11.113, p = .001) and lemmas (F (1, 74) = 8.451, p = .005).

Table 3 presents the distribution of tokens, types and lemmas by age group. There is a significant effect of age: (F (4, 74) = 13.910, p = .000) for tokens, (F (4, 74) = 24.176, p = .000) types, (F (4, 74) = 24.047, p = .000) and lemmas, respectively. High school group has the highest number of tokens, types and lemmas, followed by the older groups, university and teachers. The next group is secondary school and elementary school is the age group with the lowest number of tokens, types and lemmas. So the counts of tokens, types and lemmas increases gradually from elementary to high school, and then decreases progressively from high school to university and then from university to teachers. However, the means of tokens, types and lemmas show a different pattern: there is an increase from elementary school to university, but then the means of tokens, types and lemmas decrease in the teachers' group. An interaction is found between genre and age for tokens (F (4, 74) = 4.660, p = .002), types (F (4, 74) =

3.748, p = .008) and lemmas (F (4, 74) = 3.732, p = .008). The number of tokens also shows an interaction between modality and age (F (4, 74) = 2.574, p = .045).

		Counts	M	SD
Spoken Expository	Tokens	30,315	383.73	406.69
	Types	12,269	155.30	107.77
	Lemmas	9,479	119.99	79.64
Written Expository	Tokens	16,929	214.29	143.39
	Types	9,828	124.41	65.19
	Lemmas	7,791	98.62	51.38
Spoken Narrative	Tokens	22,151	280.39	275.56
	Types	9,730	123.16	77.97
	Lemmas	7,658	96.94	59.63
Written Narrative	Tokens	14,686	185.90	104.56
	Types	8,785	111.20	53.98
	Lemmas	6,883	87.13	41.98

Table 2. Tokens, types and lemmas distribution by text type.

		Counts	M	SD
Elementary School	Tokens	11,223	140.28	57.63
	Types	6,356	79.45	25.62
	Lemmas	5,013	62.66	18.63
Secondary School	Tokens	12,200	160.52	71.70
	Types	6,592	86.73	28.19
	Lemmas	5,137	67.59	19.55
High School	Tokens	26,082	326.02	192.00
	Types	12,059	150.73	56.26
	Lemmas	9,252	115.65	41.66
University	Tokens	19,627	490.67	282.00
	Types	8,241	206.02	69.98
	Lemmas	6,460	161.50	53.38
Teachers	Tokens	14,949	373.72	75.37
	Types	7,364	184.10	30.50
	Lemmas	5,949	148.72	25.90

Table 3. Tokens, types and lemmas distribution by age.

4.2. Lexical diversity

Figure 1 illustrates the results by text type of the three measures selected for characterising lexical diversity: types per token, lemmas per token and lemmas per type ratios. Firstly, the results of the type-token ratio are presented. Regarding the differences between the two discourse genres, no significant effect of genre is found. As for modality, type-token ratio is significantly higher in the written than in the spoken texts (F (1, 74) = 175.278, p = .000). By text types, written texts, both expository and narrative, have the highest type-token ratio, followed by spoken narrative and spoken expository. There is an almost significant interaction between genre and modality (F (1, 74) = 3.568, p = .063). Secondly, the lemma-token ratio results show that, like in the type-token ratio, genre has no significant effect. The lemma-token ratio is significantly higher in the written than in the spoken modality (F (1, 74) = 100.689, p = .000). By text types, the lemma-token ratio is higher in spoken narrative texts than in spoken expository, but in written narrative is lower than in written expository. This ratio shows an interaction between discourse genre and modality of production (F (1, 74) = 4.349, p = .040). Finally, the results for the lemma-type ratio are presented. This ratio is not affected by genre or modality. Spoken narrative texts have the highest lemma-type ratio, followed by written expository, spoken expository and written narrative. An almost significant interaction between genre and modality is found (F (1, 74) = 3.529, p = .064).

Figure 2 shows type-token, lemma-token and lemma-type ratios by age group. The type-token ratio is significantly affected by age (F (4, 74) = 10.423, p = .000), it decreases from elementary school to university but it increases in the teachers' group. This result indicates a higher lexical diversity in the oldest age group, so an interaction both between genre and age (F (4, 74) = 5.349, p = .001) and between modality and age (F (4, 74) = 4.608, p = .002) is found. Regarding the lemma-token ratio, it trends similarly: there is a significant effect of age (F (4, 74) = 7.310, p = .000), there is an interaction between genre and age (F (4, 74) = 4.070, p = .005) and between modality and age (F (4, 74) = 4.422, p = .003). Finally, the lemma-type ratio only shows an almost significant effect of age (F (4, 74) = 2.252, p = .071).

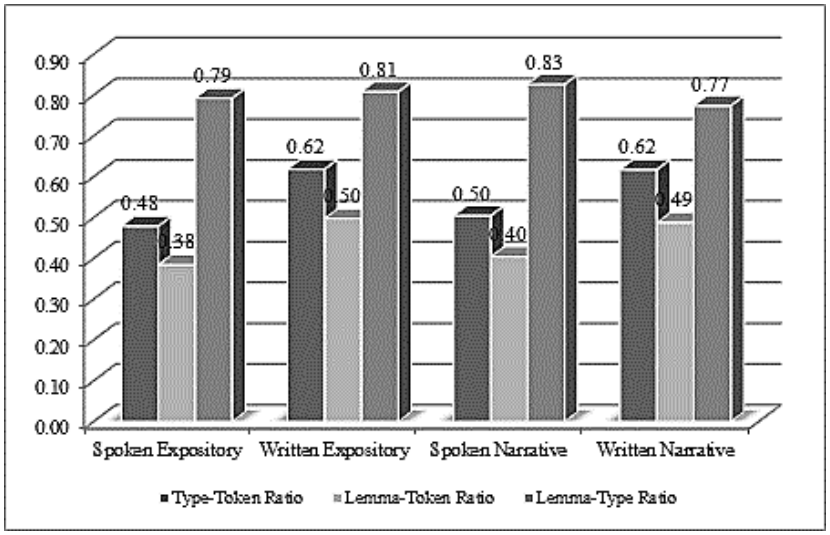

Figure 1. Type-token-lemma ratios by text type.

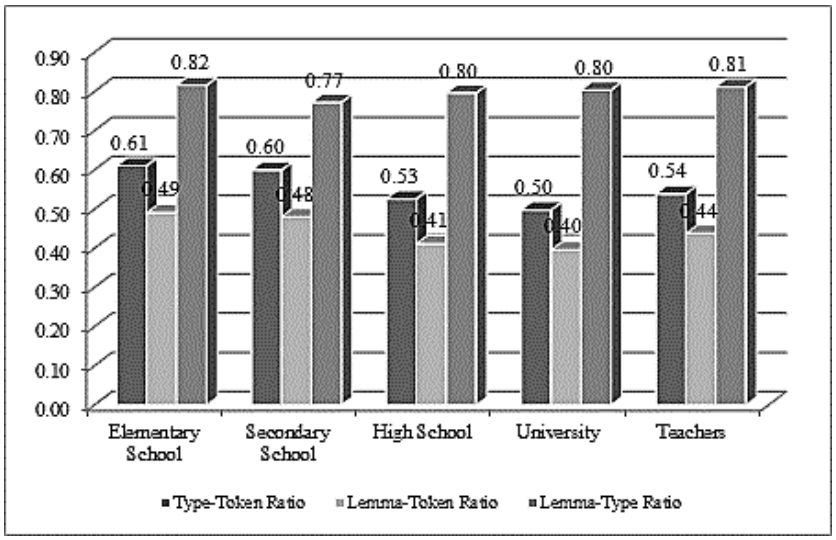

Figure 2. Type-token-lemma ratios by age group.

4.3. Lexical density

Table 4 presents lexical density by text type; roughly, almost half of the tokens used on each text type are lexical tokens. Expository texts have significantly higher lexical density than narrative ones, (F (1, 74) = 30.151, p = .000). No significant effect of modality is found. By text types, lexical density is higher in spoken expository texts than in spoken narrative, but in written narrative is higher than in written expository texts. There is a significant interaction between genre and modality (F (1, 74) = 4.524, p = .037).

Table 5 presents the results by age group. Lexical density is significantly affected by age (F (4, 74) = 10.193, p = .000). No interactions between genre and age or between modality and age are found. We can observe from the counts of lexical density that, except for the secondary school group, lexical density shows a developmental pattern. However, since the mean lexical density of each age group increases gradually, we run Bonferroni post-hoc comparisons in order to determine between which age groups significant differences are found. Regarding discourse genres, significant differences in expository texts are between the youngest and the oldest groups, that is, between elementary school (M = .484, SD = .042) and teachers (M = .527, SD = .032) (d = .043); and also between secondary school (M = .470, SD = .036) and both university (M = .509, SD = .020) (d = .039) and teachers (d = .057). For narrative texts, Bonferroni post-hoc analyses show that significant differences are to be found between elementary school (M = .431, SD = .044) and the three oldest groups, that is, high school (M = .469, SD = .029) (d = .038), university (M = .493, SD = .018) (d = .061) and teachers (M = .488, SD = .021) (d = .056); and also between secondary school (M = .446, SD = .048) and both university group (d = .046) and teachers (d = .041). As for modality, significant differences in spoken texts are found between the youngest and the oldest groups, that is, between elementary school (M = .460, SD = .028) and both university group (M = .506, SD = .021) (d = .045) and teachers (M = .505, SD = .025) (d = .044); and also between secondary school (M = .456, SD = .038) and both university group (d = .049) and teachers (d = .048). Finally, for written texts, significant differences are found between elementary school (M =

.455, *SD* = .040) and the two oldest groups, that is, university group (*M* = .496, *SD* = .021) (*d* = .041) and teachers (*M* = .509, *SD* = .025) (*d* = .054); and also between secondary school (*M* = .459, *SD* = .035) and teachers (*d* = .049).

		Counts	M	SD
Spoken Expository	Tokens	30,315	383.73	406.69
	Lexical tokens	15,071	190.92	210.85
	Lexical density	.497	.487	.040
Written Expository	Tokens	16,929	214.29	143.39
	Lexical tokens	6,960	107.06	73.19
	Lexical density	.411	.495	.044
Spoken Narrative	Tokens	22,151	280.39	275.56
	Lexical tokens	10,527	133.53	138.20
	Lexical density	.475	.463	.042
Written Narrative	Tokens	14,686	185.90	104.56
	Lexical tokens	6,897	86.25	52.78
	Lexical density	.469	.455	.053

Table 4. Lexical density by text type.

		Counts	M	SD
Elementary School	Tokens	11,223	140.28	57.63
	Lexical tokens	5,182	64.23	27.36
	Lexical density	.461	.457	.031
Secondary School	Tokens	12,200	160.52	71.70
	Lexical tokens	4,122	73.96	34.42
	Lexical density	.337	.458	.031
High School	Tokens	26,082	326.02	192.00
	Lexical tokens	12,627	157.72	97.27
	Lexical density	.484	.481	.022
University	Tokens	19,627	490.67	282.00
	Lexical tokens	9,902	247.32	142.20
	Lexical density	.504	.501	.016
Teachers	Tokens	14,949	373.72	75.37
	Lexical tokens	7,590	190.82	41.95
	Lexical density	.507	.507	.021

Table 5. Lexical density by age group.

4.4. Word length

Figure 3 shows the results of word length by text type. Word length shows significant differences between the genres and modalities studied. Regarding the differences between genres, expository texts have a higher word length than narrative texts, there is a significant effect of genre (F (1, 74) = 19.614, p = .000). As for modality, written texts have a higher word length than spoken texts, and significant differences are found (F (1, 74) = 68.432, p = .000). By text type, written expository texts have the longest words, followed by spoken expository, written narrative and spoken narrative; there is a significant interaction between genre and modality (F (1, 74) = 4.478, p = .038).

Figure 4 illustrates the word length results by age group. Word length is significantly affected by age (F (4, 74) = 3.728, p = .000). There is also a significant interaction between genre and age (F (4, 74) = 4.707, p = .002) as well as between modality and age (F (4, 74) = 7.633, p = .000). This measure shows a clear developmental pattern, because it increases throughout the subjects' age, presenting a highest increase between university and teachers groups.

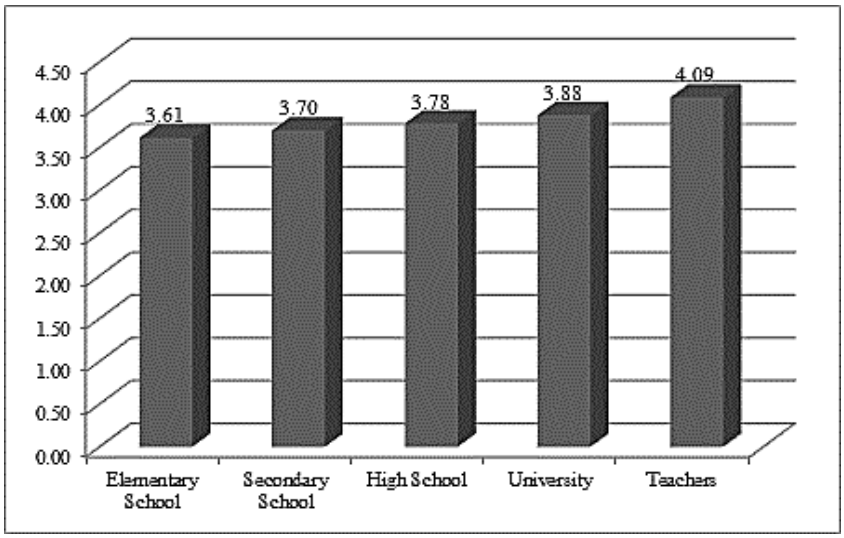

Figure 3. Word length by text type.

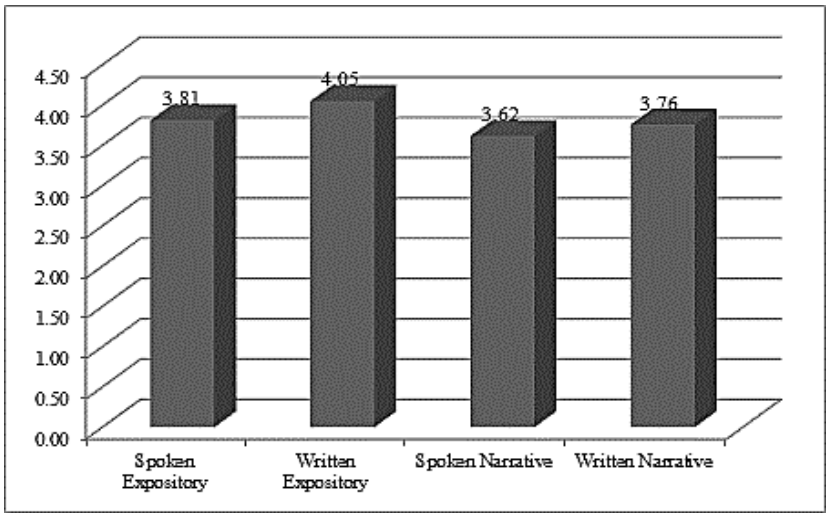

Figure 4. Word length by age group.

4.5. Productivity of verbs

4.5.1. Diversity of verbal lemmas

Table 6 shows the diversity of verbal lemmas by text type. No significant effect of genre is found. Written texts have higher diversity of verbal lemmas than spoken ones (F (1, 74) = 114.586, p = .000). By text type, written expository is the text type with the highest diversity of verbal lemmas, .658, then spoken expository, with a diversity of .516, closely followed by written narrative, .508, and finally, spoken narrative is the text type with the lowest diversity of verbal lemmas, .463.

Table 7 shows the diversity of verbal lemmas by age group. There is a significant effect of age, (F (4, 74) = 4.117, p = .005). The results show that diversity increases gradually, except for the university group. There is a significant interaction between genre and age (F (4, 74) = 3.659, p = .009) as well as between modality and age (F (4, 74) = 5.549, p = .001).

		Counts	M	SD
Spoken Expository	Verbal lemmas	635	77.61	83.78
	Different verbal lemmas	328	28.63	18.64
	Diversity	.516	.471	.157
Written Expository	Verbal lemmas	407	40.59	25.77
	Different verbal lemmas	268	22.76	12.24
	Diversity	.658	.597	.116
Spoken Narrative	Verbal lemmas	621	61.09	53.90
	Different verbal lemmas	288	24.27	14.65
	Diversity	.463	.454	.125
Written Narrative	Verbal lemmas	584	39.92	20.64
	Different verbal lemmas	297	22.16	10.71
	Diversity	.508	.575	.115

Table 6. Diversity of verbal lemmas by text type.

		Counts	M	SD
Elementary School	Verbal lemmas	323	33.02	14.57
	Different verbal lemmas	154	16.76	5.87
	Diversity	.476	.573	.096
Secondary School	Verbal lemmas	301	35.14	16.81
	Different verbal lemmas	151	17.23	6.41
	Diversity	.501	.555	.089
High School	Verbal lemmas	640	68.35	37.14
	Different verbal lemmas	327	27.91	8.67
	Diversity	.510	.529	.082
University	Verbal lemmas	597	96.12	54.21
	Different verbal lemmas	285	37.92	14.13
	Diversity	.477	.529	.035
Teachers	Verbal lemmas	546	67.30	11.05
	Different verbal lemmas	293	33.17	6.10
	Diversity	.536	.605	.033

Table 7. Diversity of verbal lemmas by age group.

4.5.2. Productive use of inflectional morphology of verbs

The productive use of inflectional morphology by text type presents no significant differences between discourse genres and modalities of production. Verbal lemmas with the least productive use of inflectional morphology, that is, verbs which appear in only one tense,

represent more than a half of the total in all text types. Spoken narrative is the text type which shows a more productive use of inflectional morphology, verbs like *anar* 'to go' and *dir* 'to say' are used in 11 different tenses. Written expository is the text type which has less productive use of inflectional morphology, verbs like *fer* 'to do' and *veure* 'to see', are used in seven different tenses.

The productive use of inflectional morphology by age group presents no significant differences. More than a half of verbal lemmas have the least productive use of inflectional morphology, they appear in only one tense. Elementary school is the age group that shows a less productive use of inflectional morphology, for example, the verb *dir* 'to say' is used in eight different tenses. Secondary school group has a similar productive use of inflectional morphology, for instance, the verb *tenir* 'to have got' is used in nine tenses. The oldest age groups have more productive use of inflectional morphology, for example, the verb *dir* 'to say' is used in 11 tenses in all these three groups.

Verbal lemmas which have a higher productive use seem to be high frequency verbs (e.g. *dir* 'to say', *fer* 'to do', *tenir* 'to have got'). By contrast, verbal lemmas with a lower productive use seem to be low frequency verbs (e.g. *traumatitzar* 'to traumatize', *senyorejar* 'to dominate', *violentar* 'to embarrass'). We have looked up in the DdF (Diccionari de Freqüències, 'Frequency List') (Rafel 1998) of the Institut d'Estudis Catalans, which is based on the CTILC (Corpus Textual Informatitzat de la Llengua Catalana) and includes literary and non-literary written texts in Catalan, in order to compare the most frequent verbal lemmas of the GRERLI-CAT1 corpus with the frequency of the same verbal lemmas in the CTILC corpus. Table 8 shows the 10 most frequent verbal lemmas (from highest to lowest) of the GRERLI-CAT1 corpus on the first column; the second represents their absolute frequency, that is, the number of occurrences of tokens of this verbal lemma related to the total of tokens of the GRERLI-CAT1 corpus (84,081). The third column shows their relative frequency, that is, the percentage of the lemma representation to the total of tokens of the corpus. In the fourth column, the same verbal lemmas are presented from highest to lowest frequency, according to the DdF; the fifth column represents their absolute frequency related to the total of tokens of the CTILC corpus (52,375,044); and, finally,

the sixth column shows their relative frequency. As we can see, the results are quite similar, if we take into account the huge difference between the number of tokens of both corpora. This confirms that verbs with more productive use are the most frequently used.

Verbal lemma GRERLI-CAT1	*Absolute frequency*	*Relative frequency*	*Verbal lemma CTILC*	*Absolute frequency*	*Relative frequency*
dir 'to say'	48	3.967	*ser* 'to be'	1,000,352	1.951
fer 'to do'	44	3.636	*fer* 'to do'	338,070	.659
ser 'to be'	37	3.058	*tenir* 'to have got'	236,795	.462
anar 'to go'	31	2.562	*dir* 'to say'	222,380	.433
haver 'to have'	26	2.149	*poder* 'can'	211,688	.413
passar 'to happen'	23	1.901	*anar* 'to go'	114,340	.223
poder 'can'	23	1.901	*estar* 'to be'	111,112	.216
tenir 'to have got'	17	1.405	*passar* 'to happen'	64,499	.125
estar 'to be'	14	1.157	*començar* 'to start'	31,259	.060
començar 'to start'	8	.661	*haver* 'to have'	1,984	.003

Table 8. Comparative between the 10 most frequently used verbal lemmas of the GRERLI-CAT1 corpus and the results of the DdF.

5. Conclusion

Our study enables us to distinguish the specific contribution of four lexicon-related measures for the corpus-based study of lexical development. The results show that word length offers the best diagnosis of lexical development, genre and modality differentiation. Lexical density is a good indicator of developmental changes and genre differentiation, but not of modality differentiation. Lexical diversity, by contrast, is a good indicator of developmental changes and modality differences, but not of genre differentiation. Similarly,

diversity of verbal lemmas is also an indicator of developmental changes and modality differences. Finally, productive use of inflectional morphology is not appropriate for characterizing later, school age language development of lexicon.

Word length appears as a valid measure to differentiate between school level, discourse genres and modalities of production. There are significant differences between age groups, effects of genre and modality and also an interaction between genre and age, modality and age and also between genre and modality. Word length shows a clear developmental pattern from childhood to adulthood. Moreover, word length is significantly higher in the expository genre than in the narrative one, and in written compared to the spoken modality. In line with other studies, such as Stömqvist et al. (2002), for Swedish language, or Llauradó / Tolchinsky (in press), for Catalan language, word length appears as the best diagnosis of development, genre and modality differences.

Lexical density is a valid measure to differentiate between school level and discourse genres. Significant differences between age groups and an effect of genre are found, and there is also an interaction between genre and modality. Lexical density yielded no clear developmental pattern, like the ones found in previous studies (Llauradó / Tolchinsky in press), though we found differences between the youngest and the oldest groups. As for differences by genre, expository texts are denser than narrative ones. However, and unlike other studies in which lexical density showed significant differences between written and spoken texts (Strömqvist et al. 2002), we found no significant effect of modality of production.

Lexical diversity, measured by type-token, lemma-token and lemma-type ratios, is a valid measure for revealing developmental changes and modality differences. However, lexical diversity does not account for genre differentiation. In both type-token and lemma-token ratios, there is a significant effect of age and modality, and interactions between genre and age, modality and age and also between genre and modality. Regarding the lemma-type ratio, no significant differences between genres, modalities and age groups are found. The highest lexical diversity is found in the oldest age group, in line with the findings of other studies (Berman / Verhoeven 2002;

Strömqvist et al. 2002). Written texts have a higher type-token and lemma-token ratios than spoken texts. This finding also corroborates those found for type-token ratio for Swedish by Strömqvist et al. (2002).

Finally, productivity of verbs functions halfway to our purposes. As for diversity of verbal lemmas, there is a significant effect of age and modality of production, and interactions between genre and age and between modality and age are also observed. Diversity of verbal lemmas increases gradually through age groups, except for the university group. Written texts have higher diversity than spoken ones. As for the productive use of inflectional morphology of verbs, it does not seem as a valid measure to differentiate between school level, discourse genres and modalities of production. Nine-year-olds and adults display a similar use in the production of inflectional morphology of verbs.

In sum, for our purposes, the most suitable measure is word length, because it can characterise differences in all variables: age, genre and modality. Lexical density is appropriate to differentiate between ages and discourse genres, and lexical diversity reveals as useful measure to differentiate between ages and modalities of production. Finally, diversity of verbs can serve to differentiate between ages and modalities, but productive use of inflectional morphology of verbs is not a valid indicator of text construction development in different genres and modalities.

The characteristics of the GRERLI-CAT1 corpus will allow us to explore the type of lexicon used by different age groups, as well as the syntactic organization of different types of texts, the textual components of each discourse genre or the effect of the order of text production (Cutillas in press), among others. Cross-linguistic comparisons with the other languages of the main project will enable us to analyse the similarities and differences between these languages, as well. Moreover, comparisons with the same corpus for Catalan as L2 (GRERLI-CAT2) will provide us with valid data to characterise Catalan language development (L1) and language learning (L2).

Acknowledgements

Data for this paper were gathered within the I+D research project *Hacia el dominio experto de la lengua: estudio comparado del desarrollo del repertorio lingüístico nativo y no nativo en castellano y catalán* 'Becoming an expert user of language: a comparative study of the development of native and non-native linguistic repertoire in Spanish and Catalan' funded by the Spanish Ministry of Science and Innovation, reference code EDU2009-08862.

References

AnCora-CA corpus. <http://clic.ub.edu/corpus/ancora/>

Aparici, Melina 2010. *El desarrollo de la conectividad discursiva en diferentes géneros y modalidades de producción [The Development of Discursive Connectivity in Distinct Genres and Modalities of Production]* (Unpublished PhD dissertation). Barcelona: Universitat de Barcelona.

Berman, Ruth A. / Slobin, Dan 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Mahwah: Lawrence Erlbaum Associates.

Berman, Ruth A. / Verhoeven, Ludo 2002. Cross-linguistic Perspectives on the Development of Text Production Abilities: Speech and Writting. *Written Language and Literacy*. 5/1, 1-44.

Boix, Emili / Alturo, Núria / Perea, Maria P. 2002. Corpus del Català Contemporani de la Universitat de Barcelona (CUB). A General Presentation. *Romanistische Korpuslinguistik. Korpora und gesprochene Sprache*. Tübingen: Gunter Narr, 155-170. <http://www.ub.edu/ cccub/>.

CesCa corpus. <http://clic.ub.edu/corpus/corpus/cesca/>

CICA corpus. <http://www.cica.cat/>

Civit, Montserrat 2003. *Criterios de desambiguación morfosintáctica de corpus en español [Morphosyntactic Disambiguation Criteria of Corpora in Spanish]*. Alicante: Sociedad Española para el Procesamiento del Lenguaje Natural.

CTILC corpus. <http://ctilc.iec.cat/>

Cutillas, Laia In press. Parlar per escriure o escriure per parlar? [Talk for Writing or Write for Talking?]. In L. Tolchinsky (ed.) *Cap a una explotació didàctica dels corpus lingüístics [Towards an Educational Explotiation of Linguistic Corpora]*. Barcelona: Horsori.

EAGLES 1996. *Recommendations for the Morphosyntactic Annotation of Corpora.* <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>.

Granger, Sylviane / Wynne, Martin 2000. Optimising Measures of Lexical Variation in EFL Learner Corpora. In Granger, Sylviane / Wynne, Martin (eds) *Corpora Galore: Analyses and Techniques in Describing English*. Amsterdam: Rodopi, 249-257.

Llauradó, Anna / Tolchinsky, Liliana In press. The Growth of the Text-embedded Lexicon in Catalan from Chilhood to Adolescence. *First Language.*

Llauradó, Anna / Martí, Maria Antònia / Tolchinsky, Liliana 2012. Corpus CesCa: Compiling a Corpus of Written Catalan Produced by School Children. *International Journal of Corpus Linguistics.* 17/3, 428-441.

MacWhinney, Brian 2012. *The CHAT Transcription Format.* <http://childes.psy.cmu.edu/manuals/chat.pdf>.

Malvern, David / Richards, Brian / Chipere, Ngoni / Durán, Pilar 2004. *Lexical Diversity and Language Development. Quantification and Assessment.* Hampshire: Palgrave MacMillan.

Nesselhauf, Nadja 2004. Learner Corpora and their Potential for Language Teaching. In J. Sinclair (ed.) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 125-152.

Rafel, Joaquim 1998. *Diccionari de freqüències. [Frequency List]*. Barcelona: Institut d'Estudis Catalans.

Rafel, Joaquim / Solanellas, Josep M. 1986. El corpus textual automatizat de la llengua catalana. *Actas de las II Jornadas Españolas de Documentación Automatizada*. 147-161.

Read, John 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Serrat, Elisabet / Sanz-Torrent, Mònica / Bel, Aurora 2004. Aprendizaje léxico y desarrollo de la gramática: Vocabulario verbal, aceleración morfológica y complejidad sintáctica [Lexical Learning and Grammar Development: Morphological Productivity, Syntactic Complexity, and Verb Vocabulary Learning]. *Anuario de Psicología*. 35/2, 221-234.

Strömqvist, Sven / Johansson, Victoria / Kriz, Sarah / Ragnarsdóttir, Hrafnhildur / Aisenman, Ravid / Ravid, Dorit 2002. Toward a Cross-linguistic Comparison of Lexical Quanta in Speech and Writing. *Writen Language and Literacy*. 5/1, 45-67.

Taulé, Mariona / Martí, Maria A. / Recasens, Marta 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of 6th International Conference on Language Resources and Evaluation*, 96-101.

Tolchinsky, Liliana 2004. The Nature and Scope of Later Language Development. In R. Berman (ed.) *Language Development across Childhood and Adolescence*. Amsterdam: John Benjamins, 233-248.

Torruella, Joan 2009. Los ejes principales en el diseño de un corpus diacrónico: El caso del CICA [The Principal Axes in the Design of a Diachronic Corpus: the case of CICA]. In Cantos, Pascual / Sánchez, Aquilino (eds) *A Survey on Corpus-based Research / Panorama de Investigaciones Basadas en Corpus*. Asociación Española de Lingüística del Corpus, 21-36.