

Corpus CesCa

Compiling a corpus of written Catalan produced by school children

Anna Llaurado, Antonia Marti and Liliana Tolchinsky

University of Barcelona

This paper outlines the compilation of a corpus of Catalan written production. The CesCa corpus presents a picture of the Catalan written language throughout compulsory schooling. It contains two kinds of data: *Vocabularies* of five semantic fields comprising 242,404 lexical forms and *Textual data* of four different discourse genres consisting of 207,028 tokens. Both vocabularies and the textual data have been morphologically analyzed and lemmatized. The corpus is freely available. This paper will outline the main features of the corpus and make some suggestions as to the uses to which the corpus can be put.

Keywords: written Catalan, lexical development, vocabularies, discourse genres

1. Introduction

Corpus linguistics makes it possible to obtain samples of authentic language uses in different contexts. It can help reveal developmental changes in language use such as lexical enrichment, the use of increasingly complex syntax, discourse attunement to genre specific features, progressive acquisition of collocations, preference of language and choice of language and register. Researchers and educationists interested in language development in Catalan, however, have never had a corpus of authentic language production at their disposal that would enable the study of language development. Existing corpora in Catalan are compilations of texts written by expert professional writers (AnCora-CA) or texts written in ancient Catalan (CICA). There are a few corpora of child language which, however, comprise very small samples or oral productions only (e.g. CCCUB). Our purpose was to fill this gap by compiling a corpus of vocabularies and texts written in Catalan from late childhood and throughout adolescence. Thus, none of the existing Catalan corpora is comparable with CesCa.

Analyses of such language production would provide an updated picture of the development of linguistic knowledge in Catalan beyond early childhood, the period of life during which speakers-writers turn into expert users of their language (Berman & Slobin 1994). We decided to compile written language production serving a variety of communicative purposes. Two main features of later language development formed the basis for this decision: mastery of the written modality and diversification of discourse development. While the oral productions of 3rd and 5th graders may not differ remarkably, the differences are much stronger when written markers are added to the comparison. While early language development centers on the acquisition of phrase structure and discourse uses are mostly limited to intimate/familiar contexts, later language development increasingly expands toward a variety of addressees and communicative functions. We decided to reflect these two features by sampling written language uses, produced in response to a variety of communicative purposes. Thus, we gathered texts representing different genres of discourse; that is, texts that serve different communicative purposes and are therefore expected to present different global text structure and organization and distinctive linguistic features. We gathered narrations of a film storyline as a representation of the narrative genre, recommendations of a film as a representation of the argumentative genre, definitions of words (a noun, a verb and an adjective) as a representation of the definition genre, and a joke as a representation of colloquial/contextualized use, closer to the spoken modality than the other three. As for the vocabularies, we included five different semantic fields: food, clothing, leisure activities, traits of personality and natural phenomena. Some fields, e.g. food and clothing, foster more everyday-like terms whereas others, e.g. natural phenomena, allow for a more specialized, advanced lexicon. Some semantic fields, e.g. food, clothing and natural phenomena, prime the use of nouns whereas the field of leisure activities fosters the use of verbs and that of traits of personality prime the use of adjectives. In all, such a configuration allows us to explore both text-embedded and isolated lexical uses in vocabularies. In order to prepare the collected data for future research, all tokens in the corpus were stored in a database and lemmatized and annotated for POS morphological features.

In Section 2 we describe the process by which the corpus was compiled and provide information on the participants and elicitation procedures. In Section 3 we comment on the storage of the corpus data. We then proceed to show some details concerning the processing of the data in both the vocabularies and the texts (Section 4) and the configuration of linguistic units in the corpus (Section 5). Finally, we suggest some directions for future research in which the corpus can be put to use (Section 6).

2. Collection of data and corpus compilation

CesCa is a corpus of Catalan written vocabularies and texts composed by school children in 2006. The corpus is freely available at <http://clic.ub.edu/es/cesca>. It was created with the general purpose of obtaining a realistic picture of the state of knowledge of written Catalan throughout compulsory schooling. Catalan is the (vehicular) language of schooling throughout compulsory schooling. Thus, children and adolescents attending school in Catalonia must develop literacy in Catalan regardless of their home language(s).

2.1 Participants

A total of 2,396 children/informants produced the corpus of vocabularies: 1,106 males and 1,290 females. The corpus of textual data included productions by 2,161 participants. Differences between the two samples are due to two facts. First, 42 children attending kindergarten were excluded from the text writing tasks due to their notable difficulties with writing. Second, for a number of reasons, 193 children at different school levels did not proceed with the task battery.

At the time of the study the participants were attending 32 schools (25 state schools, 5 semi-state schools and 2 private schools) spread across the four provinces of Catalonia. By means of a sociolinguistic questionnaire, information was obtained about sex, age, school level, home language or languages and how long

Table 1. Distribution of participants by school level and home language(s)

School level	Distribution of participants according to home language(s)				
	Catalan	Catalan/ Spanish	Spanish	Other lan- guages	Total par- ticipants
Kindergarten	33	58	21	20	132
1st grade	64	80	42	39	225
2nd grade	40	22	30	10	102
3rd grade	59	71	62	27	219
4th grade	49	34	45	9	137
5th grade	55	86	137	22	300
6th grade	34	55	55	20	164
7th grade	50	115	136	6	307
8th grade	56	122	89	12	279
9th grade	43	176	102	8	329
10th grade	30	87	73	12	202

participants had been familiar with Catalan. Four groups were identified using informants' answers to the question about the languages spoken at home, ranging from those stating that they speak only Catalan at home to those who spoke neither Catalan nor Spanish at home (see Table 1).

We sampled a minimum of one hundred participants per school level. We obtained permission for the elicitation procedures from the schools' principal. At the time we were gathering our data, parental consent was not a requisite. As for participants' home language, a majority of participants declared Catalan to be their home language(s) but other home languages were also featured in line with the current level of linguistic diversity in Catalonia (Llaurado & Tolchinsky, forthcoming).

2.2 Elicitation procedure

The 32 schools involved had been individually informed of our aim to build up a corpus of texts and vocabularies and had consented to participate. Participants' teachers were trained by the researchers in data gathering procedures (cf. Section 2.4 below). They held a meeting with the research team and were informed about the goal of the project. They were instructed to provide their students with a general explanation about the task and then with the specific instructions by reading them. The instructions included in the elicitation procedures were piloted so as to ensure that they provided the participants with adequate guidance regarding the communicative purposes of their writings. Teachers were allowed to assist students with possible doubts about the procedure. However, they were requested not to assist any child with the writing task itself. Teachers sent us all the texts and vocabularies produced by the participants.

2.3 Tasks

Five different tasks for eliciting lexical and textual productions of different kinds were presented accompanied by the following instructions:

- T.1. For obtaining the vocabularies in the five semantic fields — i.e. food, clothing, leisure activities, personality traits and natural phenomena — participants were asked to “write down all the words you can remember”. An example was provided for each semantic field.
- T.2. For obtaining a narrative text participants were asked to narrate the plot of a film with the instruction: “Tell the story of a film or TV series that you like”.
- T.3. For eliciting an argumentative text, participants were asked to recommend a film or TV series with the instruction: “How would you recommend (the film, or TV series) to a friend? Write it down.”

T.4. For telling a joke they were requested to “think of a joke or a funny story that you know and tell it”.

T.5. For providing word definitions they were asked to “define these words” (a noun, a verb, and an adjective).

2.4 Data gathering procedure

Vocabularies and texts were produced collectively in participants’ habitual classrooms at the request of their Catalan language teachers. Both the vocabularies and texts were written by hand. At the time of data gathering a number of participants were not familiar with text processing. Therefore handwriting was preferred in order to avoid possible graphic, spelling and textual deviations due to this lack of word processing skills. Exceptionally, kindergartners were seated in small groups (5 children) so that the teacher could help them with technical aspects of writing. Although there was no explicit time limit, the task did not take more than one class session. Completion of the sociolinguistic questionnaires was conducted in the same way as the production of vocabularies and texts. Sociolinguistic questionnaires were always completed before moving on to the vocabulary and text writing tasks.

In sum, this corpus differs from other corpora of written language in that (i) it contains a large amount of non-normative forms due to the characteristics of participants; (ii) it reflects a process of language development throughout different age groups, from age 5 up to age 17; (iii) it offers data about participants’ preferred home language or languages and about how long they have been using Catalan for, and (iv) the original writing has been preserved in digitalized form as one of the versions of the corpus.

3. Data Storage

The original version of both vocabularies and texts, written by hand, was digitalized in plain text format and organized in a database (CesCa), which has a relational format (in MySQL) that facilitates the retrieval of the information pertaining to each participant for both the vocabularies and texts.

We organized a relational database taking into account three basic elements: the texts, the lexical forms of vocabularies and the file (the total amount of words in vocabularies and texts produced by each participant). Each text and each lexical form of vocabularies is related to one file, to one age (from 5 to 16), to one school (i.e. one of 32), to the language or languages participants identified as their home language (i.e. 1. only Catalan, 2. only Spanish, 3. both Catalan and Spanish, 4. other languages), and to the length of time they had been familiar with the Catalan language (1.

Catalan L1, 2. more than four years, 3. less than four years but more than one year, 4. less than one year). Each lexical form in the vocabularies is related to one of the 5 semantic fields (food, clothing, leisure activities, personality traits and natural phenomena) and texts are specifically related to one of the six types of texts (narration, recommendation, joke, definition of a noun, definition of a verb and definition of an adjective). One file consists of lexical forms belonging to one of the five semantic fields or of one text belonging to one of the six types and the associated participant's metadata. The relation between each word and the text it appears in is never lost.

4. Processing the corpus

4.1 The vocabularies

One digitalized version was produced based on the first handwritten original. Next, a mirror version transcription of vocabularies reproduces the lexical forms as written by participants with total accuracy. No spelling corrections have been introduced. Due to the nature of the participants, the corpus obviously contains many Catalan forms but it also has a large variety of graphic variants, orthographic errors, creative forms of derivation, creative forms of hybridization, other languages, multiword constructions and segmentation errors. Lexical forms were classified by selecting a variant that represented all the existing variants of that form in the corpus of vocabularies. For example, Table 2 shows the different variants (flective, orthographic, etc.) subsumed under the canonical form *malaltia* "illness".

Table 2. Example of different types of variants under one canonical form

Canonical form	Lexical form	Type of variant
<i>malaltia</i> "illness"	<i>malalties</i> "illnesses"	Flective
	<i>malalties</i> "illnesses"	Orthographic
	<i>maLaLtia</i> "iLLness"	Graphic
	<i>mal altia</i> "ill ness"	Segmentation
	<i>que està malalt</i> "that he/she is ill"	Multiword construction
	* <i>enfermetat</i> — Spanish stem for <i>enfermo</i> "ill" + - <i>etat</i> Catalan suffix	Hybrid
	* <i>antisa</i> — <i>anti</i> Catalan prefix + <i>sa</i> "healthy"	Creative coinage
	<i>enfermedad</i> "illness"(Spanish)	Other language

4.2 The corpus of texts

The texts are available in three different formats following the first handwritten original: a mirror version in digital format, a morpholexical normalized version and a morphologically tagged version.

First, in the mirror version, the transcription of texts mirrors — reproduces with total exactitude — texts as written by participants. No spelling corrections have been introduced at all:

- a. Use of capital or low case letters has been respected as well as use of other graphic signs employed by participants in their original texts (Example 1).
 1. *dos Botjos s'estan escapANT eN cotxe i UN d'ells diu [...]*
 “two Crazy men are escapING bY car and ONE of them says [...]”
- b. No attempt has been made to either split or bring together graphic units in order to obtain a word in normative spelling (Example 2).
 2. *unsenyor va alas palucaria [...]*
 “aman goes tothe hairdresser [...]”
- c. Illegible characters have been transcribed as an asterisk (Example 3).
 3. *Hi ha un centres al bosc ******
 “there is a center in the forest *****”

Second, the normalized version was set up in order to prepare texts for automatic morphological analysis. As the morphological analyzer uses the graphic word as the unit of analysis (i.e. strings between blank spaces) it cannot process a text in which lexical words have been wrongly split or joined. Here, orthography was manually standardized only with regard to aspects concerning the conventional separation of graphic words in orthography. Thus:

- a. Orthographic words segmented in more than one written pseudowords have been joined by an underscore (Example 4).
 4. *pati llas* “side burns” → *pati_llas* “sideburns”
- b. Chaining of more than one orthographic word in only one written pseudoword has been split into the corresponding orthographic words (Example 5).
 5. *unsenyor* “aman” → *un senyor* “a man”

No other graphic or orthographic alteration has been made with respect to the original text.

Third, in the morphologically tagged version, all the tokens contained in the normalized version were automatically lemmatized and morphologically labeled by the HS-morpho tool. The HS-morpho tagset is based on EAGLES recommendations (Civit 2003). For Catalan, twelve categories are coded (noun, verb, adjective, adverb, pronoun, determiner, preposition, conjunction, interjection, punctuation marks, numbers and abbreviations). Each label consists of a specific number of slots each of which expresses a predetermined segment of information. For example, in a noun label, the main category is expressed in the first slot, the subcategory (common or proper) in the second, genre in the third, and number in the fourth (Example 6).

6. *bebè* “baby” ncms000 [noun common masculine singular]

As in any lexicographic study, lemmas refer to the canonical form of the word, that is, the form representing all the possible flective variants (including graphic and orthographic variants). Again, due to the particularities of the corpus, specific criteria were adopted for lemmatization:

- a. Words in Spanish or in other languages were lemmatized in the language in which the word is written. For instance, the Spanish word *catalejos* “spyglass” used in an otherwise Catalan text was lemmatized in Spanish (Example 7).

7. *Doncs que el protagonista esta mirant per el catalejos al dolent [...]*
 “so the principal character is looking through the spyglass [...]” → lemma:
catalejo “spyglass”

- b. Hybrid forms were lemmatized by their form but inflective features were not kept (Example 8).

8. *enfermetats* Spanish stem for *enfermo* “ill” + *-etat* Catalan suffix → lemma:
enfermetat

- c. Illegible forms were lemmatized by the form as it was (Example 9).

9. *m***** → lemma: *m*****

Finally, each word in the text was labeled for language (1. Catalan, 2. Spanish, 3. Other language, 4. Hybrid, 5. Unknown). Because morphological analyzers are designed to process normative texts, an in-depth manual revision was carried out after the automatic process in order to correct mistakes in labeling.

5. Configuration of the corpus in terms of linguistic units: Tokens, types and lemmas

The vocabularies are constituted by lexical forms, that is by single word or multiword units, from five semantic fields: food, clothing, leisure activities, personality traits and natural phenomena. There are 242,404 lexical forms comprised of 44,049 types. The lexical forms were submitted to a manual process of classification that grouped all the occurrences that were considered to refer to the same entity under a canonical form representing all of them (see Table 2 above for an example). In previous research, we set out all the particulars regarding the criteria applied in this process (Tolchinsky et al. 2010). Table 3 shows the distribution of lexical forms, types and canonical forms by semantic field.

Table 3. Distribution of lexical forms and types by semantic field in vocabularies

	Food	Clothing	Leisure activities	Personality traits	Natural phenomena
Lexical forms	72,014	50,226	51,234	34,995	33,935
Types	9,842	7,095	12,561	8,795	6,930
Canonical forms	1,856	791	2,235	2,471	1,864

The textual data consist of a total of 11,332 texts (rather than the expected 12,966 texts) since not every participant produced all the required types of text. The process of morphological analysis and lemmatization according to the criteria detailed above yielded a total of 207,028 tokens, 169,257 types and 157,652 lemmas. Each of these three different linguistic units of analysis allows us to interpret the corpus in terms of lexical variety (types), morphological richness and orthographic/graphic variation (tokens) and conceptual underpinning (lemmas). The distribution of these units by types of text appears in Table 4.

Table 4. Distribution of tokens, types and lemmas by types of text

	Narration of a film	Recommendation of a film	Joke telling	Definition of nouns	Definition of verbs	Definition of adjectives
Tokens	55,290	31,229	58,561	23,200	20,300	18,480
Types	43,053	27,281	42,326	21,089	18,621	16,887
Lemmas	39,041	25,765	38,229	20,337	17,839	16,441

Jokes yielded the largest number of tokens followed closely by the narrative data. That is, the most colloquial and the narrative genres appear as the two wordiest texts. The rather reproductive character of joke telling is a possible explanation for its wordiness, while the fact that narrative is the earliest acquired genre

(Karmiloff-Smith 1992) and also abundantly practiced throughout schooling may account for its wordiness relative to other genres. The argumentative data comes third in the total number of tokens. The definitional data, in contrast, yielded the lowest results both for tokens and types.

A look at the ratios between the three established units of description for this corpus — i.e. tokens, types and lemmas — provides interesting insights in terms of both the lexical and conceptual richness of texts. In Table 5 we present the distribution of token/type, token/lemma and type/lemma ratios by type of text.

Table 5. Token/type/lemma ratios by types of text

	Narration of a film	Recommendation of a film	Joke telling	Definition of nouns	Definition of verbs	Definition of adjectives
T/T	.78	.87	.72	.91	.92	.91
T/L	1.42	1.21	1.53	1.14	1.14	1.12
T/L	1.10	1.06	1.11	1.04	1.04	1.03

The more colloquial-like data yield the lowest type/token ratio (.72), followed by narrative (.78) and argumentative (.87) data, while definitional data score the highest type/token ratio (.91). In other words, more colloquial-like texts are expressed by means of fewer different words whereas more academic-like data require a greater diversity in lexical repertoire. The token/lemma and type/lemma ratios show precisely the reverse pattern. Thus, definitional data produce the lowest ratios (1.14 and 1.04 for token lemma and type/lemma, respectively), meaning that definitions yield a high concentration of different lemmas while the most colloquial-like data produce the highest ratios (1.53 and 1.11 for token lemma and type/lemma, respectively), meaning that jokes concentrate fewer different lemmas.

5.1 Some applications of linguistic units: Lemma/inflectional and orthographic variants ratios by school level

The ratio of lemmas to the number of inflectional variants provides a measure of morphological richness. It provides information on the productivity of lemmas by school level in the corpus. On the other hand, the ratio of lemmas to the number of orthographic variants provides information on deviance in spelling throughout compulsory schooling. Distribution of morphological richness and orthographic variance by school level is presented in Table 6.

As shown, the number of inflected variants of one lemma increases by school level. On the other hand, the number of incorrectly spelt words decreases as school level increases.

Table 6. Morphological richness and orthographic variants by school level

	kinder	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
	gr.	gr.	gr.	gr.	gr.	gr.	gr.	gr.	gr.	gr.	gr.
L/Inflectional variants	1.13	1.40	1.48	1.44	1.46	1.53	1.51	1.50	1.52	1.53	1.46
L/Orthographic variants	3.08	2.09	1.98	1.93	1.88	1.98	1.88	1.87	1.83	1.86	1.64

6. Possible directions for research using CesCa

The CesCa corpus is a valuable database for linguistic, psycholinguistic and educational research. On the one hand, it is unique in the Catalan language, a Romance language that so far remains less well researched than others such as French, Italian and Spanish; on the other hand, it covers all the compulsory school levels, allowing for a detailed tracking of the path through which the written language is progressively mastered while including productions meant to address different communicative purposes, both in terms of vocabularies and texts. Specifically, the vocabularies cover five different semantic fields (food, clothing, leisure activities, traits of personality and natural phenomena) and the texts represent four different genres (narrative, argumentation, joke and definition). Hence, the CesCa corpus goes beyond the habitual narrative versus expository division and incorporates less well researched types of discourse ranging from typically school-based ones such as definitions to more informal ones such as jokes.

So far, the CesCa corpus has served as a database for a variety of research endeavors focusing on the development of language use in Catalan with regard to lexical and discourse uses. First, the developmental trends of a crucial component such as the lexicon have been analyzed both in text-embedded contexts (Llaurado & Tolchinsky, forthcoming) and in isolated vocabularies (Tolchinsky et al. 2010) revealing the extent of the impact of both school grade and communicative purpose (defined by type of text in the text-embedded lexical uses and by semantic field in isolated vocabularies). Also, the characterization of the lexicon has provided data based information about the reliability of using measures considered to be crosslinguistically suitable for assessing the development of the lexicon in Catalan. Further research is needed, however, in order to gain a better understanding of the developmental path of acquisition of vocabulary depth. Future analysis of the use of synonyms and antonyms, or of the particular intricacies of the verbal paradigm — for instance, the uses of the subjunctive and conditional verbal modes — will

contribute additional relevant information. The CesCa corpus certainly allows this type of research.

Second, the CesCa subcorpora of jokes and narratives have been used in an Automatic Humor Detection System. The purpose of the system was to detect the underlying mechanisms of humor by means of comparing neutral texts (narratives) with humor texts (jokes). These two CesCa subcorpora were compared on the grounds of their level of complexity and their vocabulary. The results show that sequences of words are less predictable in the case of jokes, meaning that they have a higher level of complexity. In terms of vocabulary, jokes contain more unknown words; therefore, the use of neologisms appears to be higher in this type of text.

Furthermore, the CesCa corpus has widened the research possibilities in definitions and definitional skills, a field that has mainly focused on noun definition, since it provides ample material for the developmental tracking of definitions of words from other grammatical categories (verbs and adjectives). Initial explorations in this regard show that despite the strong effect of schooling in all grammatical categories, definitional patterns yield clear-cut differences between them (Albert & Tolchinsky 2010).

Also, the corpus is being analyzed regarding the developmental patterns of syntax in texts serving different communicative purposes. Information regarding the pattern of increase of syntactic complexity with schooling will be obtained. Together with the studies regarding text-embedded lexical uses, this sort of research will enable psycholinguists and educationists to obtain useful data on the relationships between lexicon and syntax. Complexity is certainly a yardstick of language development. Future research is needed in order to obtain additional information regarding further measures of syntactic complexity, and also patterns of global structure and content in the texts.

The corpus has been partially analyzed with the aim of defining the developmental path of spelling in Catalan. In the transparency/opacity continuum defined for alphabetic orthographies, Catalan lies approximately halfway. Thus, it is less orthographically transparent than Spanish but less opaque than French and English. This is the first corpus-based study of a not so well researched orthography and therefore makes a valuable contribution to crosslinguistic research on spelling development. Within this line of research our results, for instance, point that Catalan school graders writing in a relatively transparent orthography use morphological knowledge when writing words (Llaurado & Tolchinsky, forthcoming). The corpus will make it possible for future research on spelling development to include participants from secondary school. Furthermore, analysis of the texts provides a suitable context for exploring, for instance, word segmentation in written texts. Although most research on spelling uses rather (semi)experimental tasks, using a corpus-based approach for exploring spelling allows us to obtain an authentic picture of the relationship between spelling and knowledge of other language domains.

Given the remarkable development affecting all the language components, from spelling to pragmatics, the relevance of research on developmental discourse uses through late childhood and adolescence goes beyond the interest of linguists and psycholinguists and has important educational implications. Hence, the importance that the corpus is of public access. The analyses performed so far, as well as all the future ones that this corpus makes possible, provide the educationist community with a guideline of what aspects (local and global) are relevant for writing quality, as well as with specific examples of both good and bad text-embedded realization of each of these aspects. Such corpus-based research provides examples that are not abstractions on language production but excerpts of texts actually produced by schoolchildren. Certainly, this is much more tangible and specific than theoretical elaborations on the elusive notion of text quality, and most likely teachers will feel more safely lead in tracking down the diversity of components contributing to a good text, the extent of that contribution, and the types of difficulties the child writer can encounter.

One last, and important, feature of the corpus is that it includes the written productions of children and adolescents with very diverse linguistic backgrounds, that is some are Catalan native speakers, others speak other languages at home, some have been familiar with Catalan from birth, while others have learned it through schooling at different ages. All this information is retrievable through the database. Therefore, the corpus allows for relevant research on developmental literacy skills in Catalan both as L1 and L2 in a multilingual environment.

References

- Albert, M. & Tolchinsky, L. 2010. "Un robón no es lo mismo que un señor que roba: El desarrollo de la estructura formal y semántica en la definición de diferentes categorías morfológicas de palabras". Paper given at the *VI Language Acquisition International Meeting, Barcelona, 10–12 September*.
- Berman, R. A. & Slobin, D. I. 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, NJ: Lawrence Erlbaum.
- Civit, M. 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*. Procesamiento del Lenguaje Natural, Colección Monografías, número 3. Barcelona: Universitat de Barcelona.
- Karmiloff-Smith, A. 1992. *Beyond Modularity*. Cambridge, MA: MIT Press.
- Llauradó, A. & Tolchinsky, L. (Forthcoming). "The growth of text-embedded lexicon in Catalan from childhood and adolescence". *First Language*.
- Tolchinsky, L., Marti, M. A. & Llaurado, A. 2010. "The growth of the written lexicon in Catalan from childhood to adolescence". *Written Language and Literacy*, 13 (2), 8–22.

Authors' addresses

Anna Llaurado
Linguistics
University of Barcelona
Gran Via 585
08007 Barcelona
Spain
anna_llaurado@yahoo.es

Liliana Tolchinsky
Linguistics
University of Barcelona
Gran Via 585
08007 Barcelona
Spain
ltolchinsky@ub.edu

Maria Antonia Martí
Linguistics
University of Barcelona
Gran Via 585
08007 Barcelona
Spain
amarti@ub.edu