The relationship between lexicon and syntax in texts written in Catalan by school children and adolescents

Anna Llauradó Singla

The relationship between lexicon and syntax in texts written in Catalan by school children and adolescents

by

Anna Llauradó Singla

Dissertation presented within the doctoral program

Theoretical, computational and applied linguistics

Department of General Linguistics

Universitat de Barcelona

In Partial Fulfillment of the

Requirements for the Degree of

Doctor of Linguistics

Under the supervision of

Dra. Liliana Tolchinsky

Universitat de Barcelona

Universitat de Barcelona

September 2012

Abstract

In the life long process of language development, the school years, and their vast range of literacy activities, play a major role. The linguistic knowledge of schoolers can hardly be characterized without taking into account their performance in the written modality. Writing becomes the necessary platform for the remarkable changes that occur at the lexical, morphosyntactic and discursive levels, all of which are key to the successful attainment of literacy. Literate speaker/writers not only show an advanced sophisticated linguistic repertoire but, most importantly, they also show ability to use such repertoire flexibly for communicating a diversity of purposes.

In order to characterize the pathways of language development of Catalan schoolers ranging from 5 to 16 years of age, a period that covers all the compulsory school years in the Catalan educational system, we compiled the CesCa (Català Escolar Escrit a Catalunya) corpus. The CesCa includes written vocabularies of 5 different semantic fields and texts of 6 different types produced by 2,436 school children and adolescents attending 32 state and semi state schools in Catalonia. The participants were grouped into 5 separate groups according to their home languages. Only two groups spoke Catalan at home as their only language or in a bilingual condition along with Spanish. The sample thus notably represents the multilingualism of the school population at present, and renders an updated picture of authentic (written) language productions by that school population. All the written productions have been digitalized and prepared for computational processing in the studies presented in the thesis but also in future research. Using a corpus-based approach, we have examined different domains of development: the lexicon, key in later language development, the syntax (and the relation between these both domains) and spelling, as a problem solving space in which different levels of language are involved. We have also examined the influence of multilingualism on lexical development.

First, the domain of lexical development accounts for the acquisition through (linguistic) experience and interaction with their environment, of new lexical items and constructions that become better interconnected and that better represent the child's knowledge-base. We have found the lexicon to grow markedly throughout gradeschool in size as well as in quality, to include longer morphologically complex words, a higher proportion of adjectives (a later developing category) and more advanced, specialized, sophisticated terms or multiword constructions. Against similar research in other languages we have not found text lexical density, a wide spread used measure of later lexical development, to grow with age, adding evidence to the importance to include as many languages as possible in cross linguistic studies and suggesting future studies intended to confirm the present results.

Home language arose as a relevant variable for lexical outcome. However, multilingualism was not necessarily damaging for later lexical development. In fact, bilingual and multilingual children (children who speake neither Catalan nor Spanish at home and with a time of stay in Catalonia of 4 or more years) outperformed children who use only Spanish for out-of school purposes. Thus, being instructed in a language different from one's home language is more a handicap for monolingual children than for those other children who speak more than one language (in addition to using Catalan at school) out of school.

Both the vocabularies and the texts yield evidence that different semantic fields and types of texts trigger different types of lexicon. The different semantic fields triggered different grammatical categories and some primed more frequent words and other less frequent, more sophisticated terms. However, it is by the analysis of the text-embedded lexicon that we can best assess how, with age, children learn to fine tune their lexical uses to the type of text they are producing.

Next, the domain of syntactic development is related to the acquisition of more complex, low frequency, structures deployed for an increasingly varied range of purposes. We have analyzed the texts regarding the pattern(s) of growth of syntactic complexity in two different sites: the noun phrase and the clause level. Compared to lexical development, acquisition of syntactic complexity is a more protracted, and in the case of clause complexity, late process. Only 10th graders

produced significantly more complex syntactic architectures and explanations, the most school like type of text, arouse as the preferred site for this increased complexity. We have found some significant correlations between the lexical and syntactic uses in the written texts. Word length, use of nominalizations and level of lexical formality correlated with both clause complexity mainly, and with noun phrase complexity more moderately. Interestingly enough, performance on complex syntax was found to be related on the whether the rate of lexical growth preceding the stage at which syntax was assessed was or not sustained.

Finally, the domain of spelling regards the way linguistic information is mapped onto orthographic segments in a particular language. We have examined the developmental pattern of spelling from 1st through 5th grade, with a particular regard on he different types of knowledge necessary for rendering orthographic spelling in Catalan. We have found children to make fewer mistakes when they can turn to phonographic and morphological analysis of the words than when they need to use orthographic or lexical knowledge. Morphologically based spellings increase substantially between 1st and 2nd grade pointing to a possible effect of the salient rich morphology in Catalan.

Every domain of language undergoes developmental changes during the school years. But through actual language use, linguistic units/patterns are (re)organized under the constant environmental pressure resulting from the inevitable variation in each instance of perceived/processed input. Hence, developmental changes in the lexicon, for example, the acquisition of morphologically complex nominalizations, affect this domain clearly but also the syntactic domain, fostering longer denser architectures, for instance, longer heavier noun phrases. Overall, underpinned by cognitive and social maturations and pushed by participation in the literate practices of their linguistic community, school children's linguistic repertoire expands and children learn to deploy it flexibly in different contexts and for different purposes.

This thesis contributes to the field of later language development by covering a sample of children and adolescents wide-ranging in age and linguistic background and by applying a combination of well established language variables

with other not so well known yet, on a so far not well researched romance language such as Catalan.

Acknowledgements

Several people have played an important part in the making of this thesis. My most sincere gratitude to all of them:

To Dr. Liliana Tolchinsky, her guidance was intellectually enlightening and stimulating; her support and personal example priceless.

To Dr. Maria Antonia Marti, for generously sharing her time and suggestions.

To Manu, for his help on many important details.

To Oriol, Ana and Xavi, without their contribution I would still be on the road.

To Mila, Naymé and Cristina, for many days of cheerful support.

To the other members of the GRERLI.

To Max, for the many hours he was entitled to, with the promise that I will make up to him.

Contents

List of tables

List of figures

CHAPTER 1

Introduction

This thesis is about later language development in Catalan. More particularly, it is about the way(s) Catalan school children and adolescents ranging from 5 to 16 years of age, with a diversity of home language(s) backgrounds, but all being instructed in Catalan immersion programs, use their developing repertoire of lexical forms and morphosyntactic constructions when they are required, in their habitual school context, to write down, in Catalan, words and several texts addressing a variety of communicative purposes.

1.1 Point of departure

For decades, linguists from different domains have contended about the necessity that linguistics research be grounded on empiric methodology, in sharp contrast with other positions urging that empirical data is not as essential for linguistics as it is for other disciplines, and that it gets on faster and more efficiently by relying on speaker's intuitive knowledge of their language (Sampson, 2005). However, wider consensus has been reached today regarding the desirability that in language studies, as in other sciences, assumptions are testable, that is, they are contrasted by real language data, that is language as is used by real speakers. Also true, unlike other sciences, there are valid areas of linguistics such as literary stylistics and word semantics, where empirical method may not apply (Sampson, 2001).

While the members of all language communities may be born with the same language faculty, languages are constantly in a state of flux, and the changes they

undergo are not fully predictable. In consequence, languages are as much the result of analogy, idiosyncrasy and anomaly as they are a manifestation of a universal language faculty (Teubert, 2005). Within this frame, corpus linguistics is not an end in itself but is one source of evidence leading for improving descriptions of the structure and use of languages. Also it can lead to new theoretical frames which can explain phenomena which cannot be explained by recourse to general rules and assumptions.

Corpus linguistics is empirical and its object is real language data. As noted by Leech (1992), its focus is on performance rather than competence, and on observation of language in use leading to theory rather than vice versa. The boundaries, therefore, between corpus-based description and argumentation and other approaches to language description are not rigid, and linguists of varied theoretical positions may use corpora for evidence, which is complementary to evidence obtained from other (equally valid) sources. This is the corpus-based approach.

From a different, evolved, perspective, corpus linguistics can also offer a perspective on language that sets it apart from received views relying heavily on categories gained from introspection rather than from the data itself (Tognini-Bonelli, 2001). Thus, while corpus linguistics may make use of the categories of traditional linguistics it does not take them from granted. It is the discourse itself and not a language external taxonomy of linguistic entities which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus driven approach (Teubert, 2005). From this perspective, using corpus data merely to find out more about what we know already, or rather what we think we know, since this body of knowledge comes often from pre-corpus study, falls too short from exploiting the full potential of corpus data. Corpus data provide insights that were not previously available and corpus linguistics is not grounded on language universals understood as ontological features. Rather, corpus linguistics is concerned with the contingencies of language use and looks to explain phenomena than cannot be explained by recourse to general rules and assumptions

Corpus linguistic approach to language is not psychological but rather social. Its primary focus is on meaning understood as what is communicated between the members of a discourse community. Each discourse (or text) segment: word, multiword construction, proposition among others is observed for both form and meaning. Form represents meaning and there is no possible meaning without form. Discourse is not an ontological reality, it is a construct, an object of research set by the researcher. It is the researcher's task to define and delimit his object of research, his linguistic corpus. Then a corpus is a sample of texts whether written or transcribed from orality, compiled according to pre-established principles suitable for empirical quantitative and qualitative analysis by means of available computational resources. This analysis must allow the researcher to identify the linguistic features of the corpus and also complex associative patterns, that is, the systematic way in which these linguistic features are used in relation to other non-linguistic features.

In this study, the CesCa corpus was planned and compiled for the purpose of obtaining a multifaceted characterization of the development of language use in texts of different types written by children and adolescents attending compulsory school in Catalan whether or not this is their preferred language for out-of-school communication.

In the three past decades, and particularly since the recognition of the linguistic bases of reading and writing (Kamhi & Catts, 1989), the field of language development during the school years and beyond has gained increasing attention. At the period encompassing late childhood and adolescence, language development at different linguistic levels — lexical, grammatical and discursive— intensively interacts with increasing metalinguistic awareness and acquisition of literacy and is shaped by extralinguistic cognitive and social factors.

It is of utter importance that studies on later language development include a wide range of different languages and cultures. On the one hand too many of the available findings regarding (later) language development have been established on the basis of a single language –typically English–, and therefore the claim, for universality of a developmental phenomenon remains to be confirmed. On the other hand, the research community needs to determine which aspects of the developmental processes are governed by formal linguistic constraints, and by

those other non linguistic factors, such as the cognitive and the social development of the child (Slobin 1973, 1982).

So far, very few studies have focused on later language development in Catalan, a Romance language spoken by some 7,5 million people in Northern Spain. Most important for the purpose of this thesis, Catalan is the language used in school instruction on every subject matter in every school in Catalonia. However, after several marked waves of immigration, the proportion of children who have Catalan as their primary language for communication at home and socially represents a bare 25% of the overall population.

For this thesis we aimed at obtaining an updated picture of the Catalan uses by a representative sample of the school children and adolescents attending Catalan schools not only because this is linguistically relevant but also because of the important educational implications of such a endeavour. With this purpose we compiled a corpus and planned five corpus-based studies with the aim of analysing three developmental domains. First, the (written) lexicon, on the basis of production of isolated words and text-embedded lexicon each serving a different set of goals. Next, syntax, with a focus on the developmental acquisition of text-embedded syntactic complexity as well as on the relationship between syntactic and lexical uses. Finally, spelling, with a focus on the different types of linguistic knowledge children of different ages resorts to for solving orthographic difficulties.

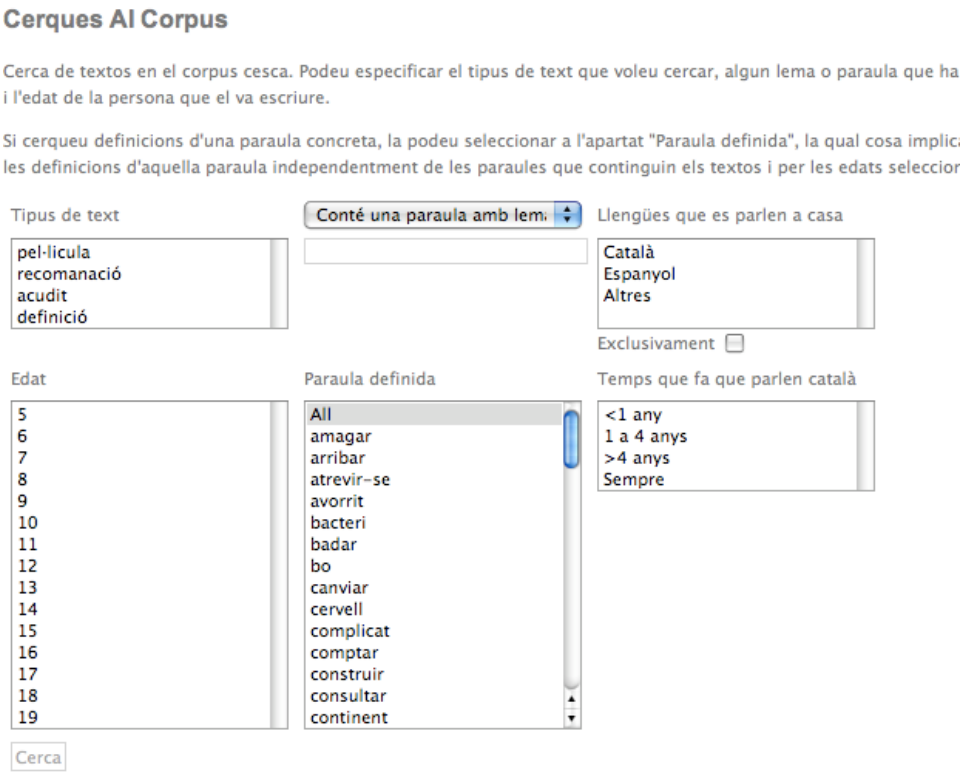
Our approach to the endeavour of characterizing the Catalan uses in the written productions of school children and adolescents is corpus-based. Corpus linguistics offers linguists, psycholinguists and educationists the possibility to focus on real language productions on a quite large scale in order to describe them by means of applying theoretically grounded research paradigms but also to bring out linguistic phenomena that have not been explored yet.

CesCa: *el català escolar escrit a Catalunya*

The CesCa corpus consists of the vocabularies of 5 different semantic fields and texts of 4 different genres produced by 2,436 children and adolescents ranging from 5 to 16 years of age that were attending last year of preschool and all compulsory school (grades 1st to 10th) in 32 different schools throughout

Catalonia (chapter 2). The participants fulfilled the writing tasks in the context of their habitual language class in their schools, as part of their daily school activities. Participants had diverse home language backgrounds. 32% of the participants declared to speak Catalan primarily at home in addition to school, 28% declared to use either Catalan or Spanish quite indistinctly out of school, 20% declared to speak Spanish primarily for family and social interaction, 9% declared to speak neither Catalan nor Spanish at home but to have been familiar with Catalan for more than 4 years (long staying immigrant children) and finally 7% of the participants declared to speak neither Catalan nor Spanish at home in addition to having been familiar with Catalan for less than 4 years (recent arrival immigrant children). Thus the corpus accurately represents the current multilingual population of Catalan schoolers. The data were digitalized and prepared for computational processing. Given that the written material had been produced by non-expert writers, we had to deal with and work out a set of difficulties, ranging from unintelligibility to unconventional word segmentation and to non-normative uses of lexicon and grammar with consequences on the lemmatization process, before we could actually proceed with data processing. Currently, the corpus is of public access for researchers and education provisioners and provides them with the only public database with these characteristics.

As presented below, the corpus can be browsed and searches can be filtered by semantic field/type of text (window on the top left corner) , age (window on the bottom left corner), home language (window on the top right corner) and time of familiarity with Catalan of the participant (window on the bottom right corner), lemma (midddle top) and word defined (if the search concerns the definition type of text) (middle bottom).



The vocabularies in the corpus allowed us to examine what children's lexical productions were like when they are asked to write down (isolate) words, of their own choice, of different semantic fields (chapter 3). We set out to examine whether age of participants and semantic field had an effect on a series of dimensions such as prevalence of a single word or multiword constructions, morphological complexity of the construction, degree of sophistication and specialization of the terms. We found that a semantic field such as Natural phenomena (1) elicits increasing specialization of the terms (N) produced (in 1 *moon* → *planet* → *artificial satelit*), most likely as a consequence of sustained exposure to school based practices.

(1) Natural Phenomena semantic field

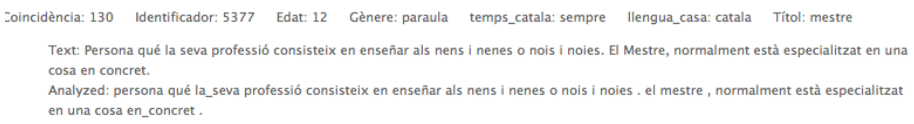
The Traits of personality semantic field (2) triggered the production of adjectives, and these became more (morphologically) sophisticated with school grade (in (2) , *(he) repeates more than a parrot* → *bother* → overwhelming).

(2) Traits of personality semantic field

Certain that text-embedded is the most appropriate context for analysis of lexical usage, we decided to characterize the growth of text-embedded written lexicon in the school years as shown in texts of different types (chapter 4). Below we present an example of two types of text (1) Word Definition and (2) Recommendation of a film as they appear in the corpus interface.

The first line displays the text metadata (Identifier, Age, Type of text , length of stay in Catalonia, home language and title). Next, under Text: it appears the text as produced by the child and finally, under Analyzed: the text corrected for word separation in order to make analysis possible.

(1) Definition


Coincidència: 130 Identificador: 5377 Edat: 12 Gènere: paraula temps_catala: sempre llengua_casa: catala Títol: mestre

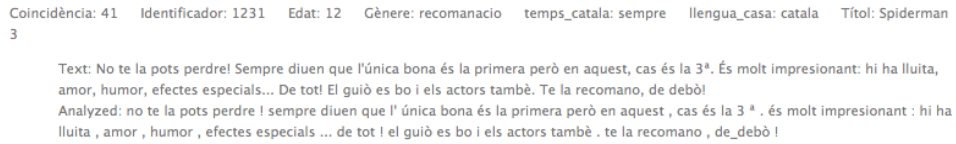
Text: Persona qué la seva professió consisteix en enseñar als nens i nenes o nois i noies. El Mestre, normalment està especialitzat en una cosa en concret.
Analyzed: persona qué la_seva professió consisteix en enseñar als nens i nenes o nois i noies . el mestre , normalment està especialitzat en una cosa en_concret .

'person whose profession consists in teaching to kids (masc) and kids (fem) or boys and girls. The teacher, habitually, is specialized in something in particular.'

Children produced formal texts, well adjusted to the conventional patterns, including both hyperonims and specialized qualifyiers, when they were required to

provide a definition, a highly school-based type of text.

(2) Recommendation


Coincidència: 41 Identificador: 1231 Edat: 12 Gènere: recomanacio temps_catala: sempre llengua_casa: catala Títol: Spiderman 3

Text: No te la pots perdre! Sempre diuen que l'única bona és la primera però en aquest, cas és la 3ª. És molt impresionant: hi ha lluita, amor, humor, efectes especials... De tot! El guiò es bo i els actors també. Te la recomano, de debò!
Analyzed: no te la pots perdre ! sempre diuen que l' única bona és la primera però en aquest , cas és la 3 ª . és molt impresionant : hi ha lluita , amor , humor , efectes especials ... de tot ! el guiò es bo i els actors també . te la recomano , de_debò !

'You can't miss it! They always say the first is the only good one but in this case the it's the third. It's really impressive: there is fight, love, humor, special effects… everything! The script is good and so are the actors. I truly recommend it!'

In contrast, their recommendations were more informal and they vividly involved the addressee by intentioanlly including some spoken-like features such as direct speech.

With the goal of characterizing the lexicon, we used a range of distributional measures for characterizing text-embedded lexical usage (word length, use of morphologically complex words, use of adjectives, lexical density) whose validity has been established by a series of previous similar research in a variety of languages. To the best of our knowledge, research on later lexical growth in texts written in Catalan is extremely scarce. In addition, we tested other less well researched measures (*F*-measure for level of text formality).

Given the particularities of the sociolinguistic characteristics of the Catalan school population, we tapped the corpus for the presence of non-Catalan words or constructions in the written productions of school children (chapter 5). Although mixing between languages, Catalan and Spanish primarily, has been researched before, it has never been done on the basis of a corpus of authentic productions. Also, code switching is a primary concern of teachers dealing with large proportions of students who do not use Catalan habitually out of the classroom. Quite unexpectedly however, we found that children had included a markedly lower amount of forms in other languages than Catalan or hybrid forms than we had expected, quite against what is commonly observed in their oral interactions.

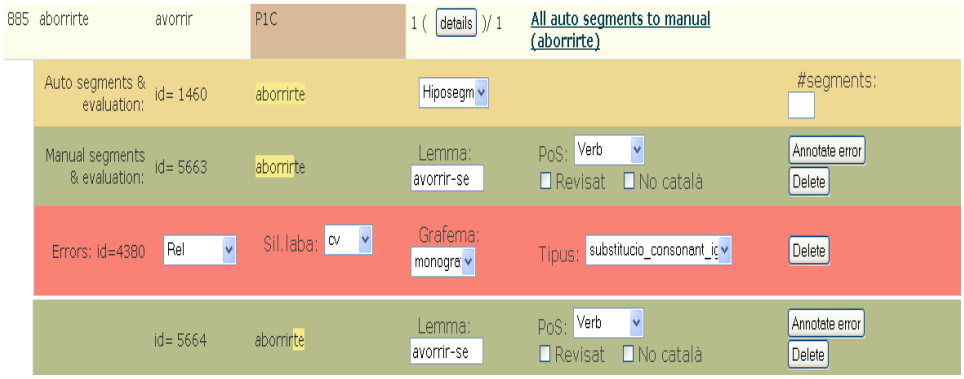
Next, we set to examine the syntactic component (chapter 6), which has also

been shown to undergo significant changes during the school years (Scott 1988, Nippold, 1988). Discourse drives changes occurring at the syntactic level, that is children learn to deploy more marked grammatical structures in a discourse embedded context (Scott, 2004) and syntax becomes the necessary tool for participating in the full array of genres that characterize proficient literate language users. The lexicon and syntax have been shown to correlate at the preschool years in a variety of languages. More recently, research have been found correlations between the two domains for school age children in English and Hebrew in narratives and expository texts. The fact that the CesCa corpus includes some extensively researched genres (i.e., narratives) but also others less well explored (i.e., recommendations and jokes) only pushes the relevance of examining the syntactic level. Also, the corpus-based approach is a suitable platform for addressing different but related aspects of language productions such as here, were we tap the relation between syntactic and (formerly explored) lexical uses.

The last study of this thesis is concerned with the developmental pattern of spelling (chapter 7), a key component of writing and one of the most significant challenges of a child's early academic life. Learning to spell involves understanding the relation of the graphic elements with the different levels of language: phonology, morphology, syntax, and the lexicon. While a great many studies have explored the developmental patterns of spelling both in more transparent orthographies (Spanish or Finnish) and more opaque ones, the relation between spelling and other linguistic domains has been less examined for languages with more transparent orthographies (but see Gillis & Ravid 2006). Catalan, a language with a rich morphology, is represented by a moderately transparent orthography (less than Spanish but more than French). Morphological recognition of the different morphemes in the word is useful for avoiding misspellings.

Analyzing spelling on a corpus database represents a huge task in terms of coding. Therefore we designed an interface (see screen image below) that allowed us to 1) correct the words for segmentation errors and 2) lemmatize the words, assign them a grammatical category (on the manual segment evaluation level, in the screen image), and code separately each spelling error in the word for information concerning the (3) morphological segment of the word containing the misspelling, (4) the syllable structure, (5) whether it was on a monograph or a

bigram and (6) what type of error it was (on the error level, on the screen image).



1.2 Thesis outline

The present thesis consists of a collection of 6 studies wrapped between this introduction and a concluding chapter (chapter 8) that glue them together in a subject whole. The six studies are the following:

1. Llauradó, A., Martí, M. A., & Tolchinsky, L. (2012). Corpus Cesca*:* Compiling a corpus of written Catalan produced by school children. *International Journal of Corpus Linguistics* (in press).
2. Tolchinsky, L., Martí, M.A., & Llauradó, A. (2010) The growth of the written lexicon in Catalan from childhood to adolescence. *Written language and literacy*, 13 (2), pp- 206-35.
3. Llauradó, A. & Tolchinsky, L. (2012). Growth of text-embedded lexicon in Catalan: from childhood to adolescence. *First Language* (in press).
4. Llauradó, A. & Tolchinsky, L. (2012). Developing a written lexicon in a multilingual environment. In Grommes, P. & Hu, (eds.): *Plurilingual education: Policies, Practice, language development*. Hamburg Studies in Linguistic Diversity (HSLD) Amsterdam, NL: John Benjamin (in press).
5. Llauradó, A. & Tolchinsky, L. The development of syntactic complexity and its relation to lexical growth in the Catalan written language (submitted)
6. Llauradó, A. & Tolchinsky, L. The developmental pattern of spelling in Catalan from 1st to 5th grade (submitted).

The first 4 chapters include studies that have been or soon will be published in peer reviewed journals or books. The two last ones are currently under review. They are co-authored by this thesis advisor. The papers are reprinted here reformatted to make the typography of the thesis consistent, and the references and appendices are integrated in a single bibliography section at the end of this volume.

In chapter 2 we present the corpus that has served as database for all the studies included in the thesis. Following, chapters 3 to 5 discuss questions concerning lexical development in the corpus. Chapter 6 presents the analysis of the acquisition of complex syntax at three different stages of compulsory school and its relation with selected measures of lexical development. Finally, chapter 7 is concerned with the developmental patterns of spelling through gradeschool.
In the remainder of the chapter, I review previous studies to put my work into perspective.

1.3 Related work

1.3.1 Later language development

From our very first utterances in early childhood and throughout lifespan, users of any language experience an ongoing process of reorganization of our linguistic structure and knowledge, which are affected by every instance of language use (Bybee, 2010). Speakers develop linguistic knowledge and language use by means of and in order to produce and understand different types of texts – conversational utterances, jokes, stories, essays, or articles. Linguists, psychologists and psycholinguists have given broad attention to linguistic attainments that occur during infancy and early childhood since the 1960s. Though to a fairly lesser extent, language development in the grade school years and adolescence has received mounting attention in the last decades. Recognition of the linguistic bases of reading and writing has doubtless added to the interest in later language development. As more is known about the close connection between the development of literacy and continued growth in language, and consequently between language development and the academic success (or risk to failure) of

school-age children and adolescents, the importance given to the field may continue to improve.

Language development beyond early childhood is a gradual and protracted process extending throughout adolescence and well into adulthood (Nippold, 1988, 2004; Berman, 2008). The traditional emphasis on early language development, as well as the critical period hypothesis set forth by Lenneberg back in 1967 may have brought some people to question the possibility of significant growth in language during the school years. True, by age 5 children have acquired the vast majority of the phonological, morphological, syntactic and semantic regularities of the target languages irrespective of the language or languages to be learned (Weissenborn & Höhle, 2000), and it is also true that they can speak in long and complex utterances, produces clear articulation, uses a wide variety of words and actively contributes to the conversation. However, a lot more remains to be learned (Owens, 2008). More complex oral narrative and conversational skills (Goetz & Shatz, 2000) and the ability to comprehend figurative language are developing from the mid primary years. Hence the linguistic productions of a child of this age hardly match and adult's or even a 12 year old's (Berman, 2007, Berman & Slobin, 1994, Ravid 2004). As children progress through school they exhibit new and more extended repertoires of linguistic items, categories, and constructions and show increased proficiency and flexibility in the use of these forms in a wide range of communicative settings, both in the spoken and the written modalities (Berman, 2004) as well as increasingly more efficient and explicit ways of representing the language and thinking about it and for accessing its functions and structures including higher-order, non-literal aspects (Berman, 2004, 2008; Berman & Ravid, 2008; Karmiloff-Smith 1986; Nippold, 1998; Ravid & Berman, 2010, Tolchinsky, 2004).

From a usage-based language stance, such increasing repertoire of linguistic forms constitutes a dynamic system constantly evolving as a consequence of one's own experience with language (Bybee 2007, Goldberg 2005, Tomasello, 2003). There is little point in considering these linguistic forms as abstract, isolated elements but rather they must be observed in relation to how people use them in different kinds of texts (Berman & Verhoeven, 2002; Berman, 2006) and under the constraint of differing communicative circumstances, goals and audience.

Language is not used in a void and, as Tolchinsky puts it: "There is no such thing as *neutral* use of language: speakers learn to attune their speech to specific intentions, purposes, and interlocutors" (2004: 235). Changes in the speaker's grammar are driven by discourse, that is, discourse provides children with a developmental mechanism for the acquisition of linguistic devices (Hickmann, 2003). In other words, the organization and re-organization of linguistic forms is considered as embedded in discourse (Du Bois, 2003).

This period of intense linguistic change, driven and shaped by language experience, taking place during the school years, provides linguists and psycholinguists with a dearth of information about the nature of such linguistic changes across all linguistic domains, but also about the way use and knowledge of the language interplay with and crystallize at cognitive and social maturity. For decades, oral language outshined writing as a focus of interest/research. Speech was seen as primary whereas writing was considered to be secondary, almost a mere reflection of speech. However, to a fairly important extent, the process from moving beyond emergence to becoming a proficient speaker takes places hand in hand with the school-centred process of literacy acquisition. And let's not forget, becoming a proficient writer is a major accomplishment, a fundamental requisite, for academic, and today post-academic, success. Through schooling children receive increased exposure to and practice with the written modality of their language. From this moment on, the linguistic knowledge of school age children and adolescents can hardly be characterized without taking into account their performance in the written modality (Ravid & Tolchinsky 2002; Tolchinsky 2004). Clearly, when linguistic markers are broadened to include written forms in addition to spoken forms of language the distinction between first grade an third grade schoolers becomes much more evident. In order to acquire the language of literacy children need to quickly move beyond learning the notational aspects of writing and become skilful with writing as a discourse style. They need to gain quick and informed access to a wide range of linguistic varieties, discourse genres and registers of use. And they need to gain ability at producing monologic pieces of discourse wrapping them in appropriate register and genre specific features. The developmental path leading to such integrated control over all the aspects involved in text composition appears to be an effortful, at times thorny, task.

Composing a text is an extraordinarily complex strategic activity (Berninger & Winn, 2006) done for a purpose in a sociocultural context (Nystrand, 2006) that involves orchestrating many cognitive abilities in addition to accessing and manipulating, through graphomotor processes, representations at a variety of linguistic levels –phonologic, lexical, morphologic, syntactic and rhetoric. Whereas a variety of studies have shown that children from very early on and before being able to write conventionally distinguish between narratives and descriptions (Tolchinsky 1992), and reproduce some genre specific linguistic forms (Pontecorvo & Zucchermaglio, 1988; Spinillo, 2001; Teberosky, 1992), learning to transcribe this discourse knowledge into the notational restrictions of the printed language, however, is not exempt of difficulty. Low level components such as spelling and punctuation need to be fluently coordinated with high level ones ranging from clause packaging to planning, organizing and revising the text. Understanding in depth the developmental path followed by typically developing children is necessary also so as to inform what does and what does not deviate from the norm. Currently, research is underway on understanding writing skills of children with oral language disabilities (Dockrell & Connelly, 2009; Nelson, Roth, & Van Meter, 2009).

In this frame, written language constitutes the core of literate language use. This is not to claim a complete separation between the oral and the written modalities of a language. Rather, learning to read and write requires and facilitates the active analysis of certain aspects of languages that were largely ignored or passively experienced during the preschool years. If orality has an expressive power adequate to interpersonal conversational communication (Olson, 1994), writing forces the writer to control and shape the flow of information through linguistic means and to see the text as a whole (Stromqvist, 2004). In fact, true command of both modalities and skilful ability to move in both and between the two can be seen as the milestone of linguistic literacy. Literacy related activities foster and are enhanced by the acquisition of advanced metalinguistic skills, from about the age of 9 years and throughout adolescence. Such skills enable children to acquire advanced vocabulary including the literate lexicon by analyzing the words contained in the expression as well as the word's morphological structure. Also,

children's processing and production of more complex syntactic structures in a range of genres including fictional and expository texts, is ongoing from this stage on (Nippold 2002; Larson and McKinley 2003; Nippold, Hesketh, Duthie & Mansfield, 2005; Nippold 2007; Nippold and Sun 2008; Ravid and Berman 2010).

Development at different domains can not be seen as a fully autonomous process and rather development in one domain relates to development at others as show by the critical interactions between different components of language; for example between word learning and syntactic knowledge (Gleitman, 1990), between semantics (word meaning) and pragmatics (discourse) (Cain, Towse & Knight, 2009). Understanding the precise nature of the reciprocal interactions between the different language blocks is relevant for typically developing children but key where the process of language development is somehow disordered (Byrne, MacDonald & Buckley, 2002; Philofsky, Fidler & Rogers, 2008) given the implications it holds for the acquisition of literacy (Kendeou & van den Broek, 2007; Nation, Snowling & Clarke, 2007; Myers & Botting 2008; Nation, Cocksey Taylor & Bishop, 2010)

The protracted and complex nature of language development, which cannot be explained by one single mechanism, sheds light on the cognitive and social underpinning of developing language use (Berman & Katzenberger, 2004, Reilly & Anderson, 2002, Stromqvist, Nordqvist & Wengelin, 2004). As children progress through school grades, they advance their ability to use language for school-related purposes as evidenced, for instance, by the late appearance of metacognitive verbs, or by their improved ability to handle abstraction and deal with analogical thought (Ginsburg & Opper, 1988), or to understand figurative meanings of constructions and expressions, or else to represent and manipulate operations in order to formulate and test hypothetical deductive reasoning and to consider a problem from a variety of perspectives. Cognitive development causes changes in problem-solving behaviour and in gathering and storage of information (Rumelhart & McClelland, 1986). With age, children gain attentional memory and resources for processing information and have better executive control processes. Psycholinguistic studies have established a strong relationship between cognitive (measured by tests of intelligence) and language (vocabulary) development (Anderson & Freebody, 1981).

The role and the contribution of social factors to development of language in toddlers and preschoolers has been extensively documented (Aronson & Thorell, 2002; Corsaro, 1985: Goodwin, 1990). In studies about the nature of such contribution to language development through the school years (and beyond) there is a shift to issues of language and identity (Hoyle & Adger, 1998). However, variability concerning the communicative context creates in the speaker writer the necessity to adjust his expression –to make choices overarching from phonology to syntax and pragmatics — to the non-linguistic requirements of the situation. These requirements include among others the goal, the relationship between the speakers (or the writer and the reader) and the circumstances surrounding the communicative situation and determine, for instance, the choice of register and/or modality. Thus, social interaction, as undertaken, for instance, in peer talk, provides an adequate scenario for exercising the speaker's sensitivity to the social needs of interlocutors, a responsibility to communicate a message that will be understood, and an awareness that such understanding depends on the way the message is produced (Barbieri et al., 1990; Hasan, 1992). Such explanatory skills are of relevance since explanatory discourse, through its association with decontextualized modes of thinking, helps children gain membership into a literate community (Blum-Kulka, 2010).

1.3.1.1 Central components of Later Language Development

From a usage-based perspective, grammar is the cognitive organization of one's experience with language. Usage feeds into the creation of grammar just as much as grammar determines the shape of usage (Bybee, 2006). Grammar cannot be thought of as pure abstract structure that underlies language use, just as there can be no discrete separation of grammar and lexicon and between open and closed class items, but rather items and constructions lay separated by a matter of gradience on a same continuum. Grammar is thus seen as emergent from experience, mutable, and ever coming into being rather than static, categorical, and fixed. In other words, language is a complex dynamic system, where grammar is built up from specific instances of use that marry lexical items with constructions (Bybee, 2010).

1.3.1.1.1 Later lexicon development

Later language development is a dynamic process that affects every language domain from phonology to pragmatics. The lexicon nonetheless plays a particularly key role in the development of language during the school years and it provides a unique domain for studying the interaction between context and cognition, and the ways in which this interaction changes with development (Dockrell & Messer, 2004).

Through schooling children's core lexicon –the basic vocabulary acquired in the preschool years mostly through spoken interaction increases exponentially to become a literate lexicon, that is, a mental dictionary of thousands of complex and low frequency words, coexisting in a dense semantic network organized for flexible access and use (Baayen & Renouf, 1996; Ravid 2004). Unlike most words learned through early childhood which commonly denote concrete familiar entities, words encountered in literate contexts often express multiple, abstract or figurative meanings, include multiword expressions and idioms and metaphors and refer to internal, cognitive and affective states. The literate lexicon has an encyclopaedic nature, includes words belonging in a wide range of knowledge domains and requires specialized school like knowledge of the world (Biber, 1995).

Lexical development in the school years entails, in first place, an enlargement of the number of words a child knows. A child entering 1st grade knows about 10,000 words and following Nagy and Anderson's (1984) that child acquires 3000 words per year between third and ninth grades; that is his lexicon grows at the impressive rate of several words per day. Importantly, a preschooler's vocabulary size at the point of entering school, which in turn conditions his capacity to develop it further has been shown to depend on whether the environmental opportunity he has had before starting formal schooling was more or less rich (Weizman & Snow, 2001). The differences between children from either upper and lower socioeconomic strata are maintained through elementary school and affect the rate of acquisition of both basic and complex terms.

Whereas young preschool children learn most of the words they know through oral interaction, the educational setting becomes a main context for word learning for school age children and what factors may be involved in promoting vocabulary growth (as well as individual differences in it) remains an open debate. Three main courses of vocabulary teaching seem to be: direct, explicit instruction (McKeown, Beck, Omanson & Perfetti, 1983; Stalh & Fairbanks, 1986), learning words and their meanings from contexts, especially during reading activities (Nagy & Herman, 1987) or growing ability to infer the meanings of words through morphological knowledge (Tyler & Nagy, 1983; Anglin 1993). The two last possibilities are of particular importance since, with school grade, children encounter most of the new words they learn through autonomous reading of text books or other types of printed language.

However, growth in word knowledge not only occurs through addition of new words but through the development of an organized semantic network. This type of semantic organization is reflected in the syntagmatic-paradigmatic shift that takes its most crucial turn between the ages of 5 and 9 years. Also, the increase of connection between the lexical items promote changes in the meaning of words that make possible understanding of abstract, double-function (Schecter & Broughton, 1991) or polysemous (Durkin, Crowther & Shire, 1986) words, as well as metacognitive and metalinguistic verbs (Astington & Olson, 1987). Also, lexical attainments become intrinsically related to other aspects of language development such as verbal reasoning or understanding of figurative uses of language.

How, then, word knowledge should be measured is critical to educators, clinicians, parents, and researchers. Frequently used measures of vocabulary estimates have included definition, that the word be used meaningfully in a sentence, provision of a synonym or paraphrase o selection of either on multiple-choice tests. Research has devoted many endeavours to examining the role word knowledge plays in word definition. Obviously, an individual cannot define a word unless he has some knowledge of it. However, such individual can have knowledge of the word that does not show in his definition. Thus, word definition requires reflection on the lexicon and is related to cognitive and linguistic development,

literacy and academic achievement (Nippold, 1998). Word definition mastery therefore is a developmental matter that gradually improves during school age years in terms of formality, presence (and quality) of a specific category term and number of characteristics mentioned (Snow, 1990). It should not be the only, nor the main, via of vocabulary assessment. Dockrell and Messer (2004) claimed that word knowledge should be assessed with different measures and should consider the quality and quantity of children's vocabulary knowledge. These authors also claimed that both comprehension and production of vocabulary must be assessed. A well-developed receptive vocabulary is a prerequisite for fluent reading, a critical link between decoding and comprehension (Joshi, 2005). In general, children with a larger vocabulary tend to continue to expand their word sets faster, and to understand texts more easily, than children with a smaller vocabulary. However, understanding some aspects of a word does not necessarily indicate understanding of the word's meaning in a more complex context. Therefore, it is necessary to assess both the receptive and productive dimensions of lexical knowledge.

One major feature of this protracted process of lexical acquisition is progressive access and command of morphosynatctic forms, such as morphologically derived terms, which are rather rare in everyday oral input (Anglin, 1993). Conventional words, that is lexical items between blank spaces in the printed language tend to grow longer (in number of letters) as a consequence of their multimorphemic form. This (recursive) affixation is used to create words expressing semantically complex deverbal and deadjectival attributes and states. Complex derived words become increasingly important throughout the school years in content area reading, writing, textbooks, and literature (Nippold, 2007). In other words, acquisition of derived nominals is at the heart of developing a literate lexicon. It includes a wide spectrum of nouns relating to verbs and adjectives whose meanings range from semantically transparent morphologically compositional to more semantically opaque morphologically blended terms. In all cases they demand a solid command of the language morphology and play a key role in construction of syntactically dense noun phrases and subordinating constructions (Ravid & Cahana Amitay, 2004). It is not surprising then that derived nominals be rather uncommon in everyday spoken language whereas, instead,

they have been found to be are profusely used in mature literate written texts, most particularly in texts belonging to the expository genre (Ravid, 2003, 2004).

Later lexical development has been show to trigger growth of adjectives in a very special way (Ravid & Levie, 2010). Adjectives are a less primary lexical category than nouns and verbs and they denote attributes or properties of nouns (Lyons, 1968), that is, they narrow down the identification of nouns and NP's. Adjectives, are less dense in meaning and have a less correlated structure than nouns, and they are more prone to adjusting not only their form but also their meaning according to the modified noun. Research has found that both adults and children rely on the contrastive functions of adjectives in the interpretation of NPs (Prasada & Cummins, 2001). Size and array of adjectives has been found to coincide with the consolidation of an advanced, high-register, literate lexicon and its cognitive correlates (Dockrell & Messer, 2004) both in English (Bar-Ilan & Berman, 2007) and Hebrew (Ravid, 2010). Morphosemantic and syntactic distribution of adjectives, has been found to be affected by modality of production. Rate of adjectives has been found to be higher in written texts than in spoken texts.

In addition to the above presented measures (word length, the use of nominalizations, the use of adjectives) all of them accounting for intraword characteristics of the lexical pieces, other measures determined by proportion of words within a text (lexical density and lexical diversity) have also been used as descriptors of later lexical development. Lexical density, the term most often used for describing the proportion of content words (nouns, verbs, adjectives, and often also adverbs) to the total number of words, and lexical diversity, the term describing the proportion of different words over the total number of words produced, have been both used as measures of later lexical development showing differences between elementary and junior-high school on the one hand and children above 17 years of age and adults on the other hand (Johanson, 1999, Ravid, 2008). Several studies have found lexical density and lexical diversity to discriminate by genre but this results have not been replicated else were (Johanson, 1999) suggesting that usefulness of these measures may be language specific. Lexical diversity poses additional difficulties as it depends on text length, an issue worth considering in texts produced by elementary school children.

One final, but worth noting, consideration concerns the types of words having shown marked growth. Considering all the above, the development of a later lexicon can be tracked as affecting primarily to open class grammatical categories. However, such enrichment of words and constructions --belonging in these open-class lexical categories-- is provided through use of items lying between the open and the closed class categories such as adverbials, connectives and discourse markers. This border-like lexical elements can be seen as syntactic constructions with functional alterations (Ravid & Schlessinger, 1995). By using them, textual cohesion and coherence can be enhanced (Hickmann, 2003), and they further contribute to improving the texts by providing a way for personalizing the writer's stance (Schiffrin, 1994).

We cannot conclude this section without adding some of the educational implications of later lexical development, to those of linguistic relevance stated above. Improved knowledge of derivational morphology plays an increasingly important role in the interface between lexicon and syntax (Ravid, 2004). Assessment of vocabulary knowledge at 1st grade explains 30% of the reading comprehension variance in 11th grade (Cunningham & Stanovich, 1997; Leong & Ho, 2008) and explains individual variance in reading comprehension (Leong & Ho, 2008; Laufer & Nation, 1999). Frequency of use of nouns and verbs plays an important role in reading speed (Holmes, Stowe & Cupples, 1989) and in reading comprehension as well. Children with reading difficulties usually exhibit a poorer vocabulary than their more skilled peers. Moreover, educational interventions on lexical aspects entail progress in reading comprehension (Nation, Snowling, & Clarke, 2007). Further research aiming at improved understanding of the complex process of lexical acquisition (Anglin, 1993) continues to be in place and is crucial if we aim at providing disadvantaged learners with tailored instructional practices.

In the work we are presenting here, the CesCa corpus was designed so that it would provide us with the possibility to examine the development of the written lexicon from a two-fold perspective: as produced in a word writing task in which they were required to produce as many *words* as possible for 5 different semantic fields, on the one hand, and text-embedded, on the other hand. First, the corpus was designed to examine how lexical growth is realized in different lexical

categories, similarly to the way semantic weight is distributed in the mental lexicon. The children had to retrieve decontextualized lexical units: entities, qualities and activities, that is, N for entities, Adj. for qualities and V for activities. Three semantic fields primed N since this category predominates in the lexicon too. However, the three semantic fields targeting N, elicited noun words (or constructions) varying in the frequency and sophistication pattern of the primed words. Second, confronting the children with a word writing task was a suitable way to track the developmental pathway of the word construct. From a usage-based perspective, through repeated instances of experience with language use, the lexical items are married to constructions and available for access thanks to rich memory storage (Bybee, 2010). Thus, in the child's environment, the common tokens belonging to the different semantic fields represented in the CesCa vocabularies may be words but also constructions bigger in size. Reducing these multiword construction to a single word production might rather be the outcome of increased familiarity with this unit as the building block of written texts.

In a second, but fundamental, line, the corpus allowed for text-embedded characterization of the lexical uses. With age, and familiarity with the written language, children's lexicon experiences growth in size and becomes increasingly fine tuned to the different genre specific features serving different communicative goals. The CesCa corpus was planned to provide us with a fairly complete reservoir of data. Unlike many previous studies that leave children younger than 9 out of their research, the CesCa corpus includes children attending each grade of compulsory schooling, from 1st grade (5 years) to 10th grade (15 years) and therefore allowed us to track the process of development and growth in detail. Also, it includes six different types of text, some of a conversational, informal nature such as joke telling, and recommendations, but also others far more formal and detached such as explanations and definitions. Asking a schooler to provide a definition of a noun is not uncommon in school setting. The CesCa in addition includes two other types of definition: of a verb and of an adjective, both far less common in daily school activities. It is important that a range of measures tapping on different aspects of lexical growth be used in research on lexical development. For instance, although the growth in size and diversity of the lexicon during the school years is considered a major feature of later lexical development, it is also

important also to use measures suitable for tapping into the morphologica composition of the lexical pieces, given the typological characteristics of the (rich) Catalan morphology.

1.3.1.1.1.1 The growth of text-embedded lexicon in a multilingual environment

In the context of the present work, it is important to take into account that the linguistic background of the Catalan schoolers is highly heterogeneous. Repeated waves of immigration in the 50s and 60s brought to Catalonia many families from the southern regions of Spain. It is not uncommon that the descendants of these immigrant families continue to live in Catalonia to date and that they use Spanish more often than Catalan for social communication. More recently, Catalonia has experienced a new major wave of immigration with many families coming from many different countries in the South American, African and Asian continents. The proportion of non-Spaniard immigration rose form 3% in 2000 to a 13% in 2008. In this context the designation L1, L2 and so on, does not correspond exactly to the ecological situations in which these languages are acquired by children and adolescents at present. Catalan is the only language used at school through a national program of immersion, and Spanish has a massive presence both in the media and socially. Therefore it is very unlikely to find a strictly monolingual school-age child or adolescent neither in Catalan nor Spanish or another language. All children must use Catalan in school-based tasks but many do not use Catalan for family or social communication, some degree of bilingualism in Catalan and Spanish, or in Catalan and some other language is the norm. Although both first and second language learners face the same problem, that is how to map form and function to produce meaningful utterances based upon their language experiences (Ellis 2002; Lieven & Tomasello 2008), there are certain fundamental differences between L1 and L2 acquisition. Some phenomena such as code switching are restricted to multilingual speakers (they have been shown to exhibit it in oral interactions for a variety of meta-communicative purposes (Myers-Scotton, 1993; Poplack, 1987). In general, bilingualism appears to have both benefits and costs. Regarding costs, bilinguals typically have lower formal language proficiency than monolinguals do; for example, they have smaller vocabularies and weaker access to lexical ítems (Bialystock, 2001). The benefits, however, are that bilinguals

exhibit enhanced executive control in nonverbal tasks requiring conflict resolution (Bialystock & Martin, 2004, Costa, Hernandez, & Sebastian-Galles, 2008).

Given the key role of the lexicon in the process of later language development, and given the marked multilingualism of the Catalan school population, it is important to have research grounded data on the effect of this multilingual environment on the pathways of lexical development shown by schoolers throughout compulsory schooling. Instances of code switching and code mixing are habitual in classroom peer, as well as in child-teacher, interactions. When the child switches codes, he uses a word, or string of words, in a language that is not Catalan, within an interaction being held in this language. Code switching may be due to different reasons: lack of knowledge of a particular term to expressive preference, among many others. By mixing we refer to producing hybrid instances, that is a word made up of elements of Catalan and also elements of any other language known by the speaker (although not necessarily by the other interlocutor). These two phenomena are common in oral interaction and many studies have been conducted to research. Our work, instead, was set to find out whether this linguistic behaviour is also found, an if so to what extent, in school writing tasks. We think that the child may perceive writing as being more formal, and may hold back from mixing different codes. This might have relevant implications regarding the role of writing practices when teaching to Catalan L2 children.

1.3.1.1.2 Later syntactic attainments

Syntax it is the structural foundation of sentences (Crystal, 1996) and it is due to syntactic competence that a speaker writer can generate an infinite number of sentences to express an inexhaustible supply of ideas (Chomsky, 1965). Since early on and certainly by the time children enter formal school, they produce grammatically well-formed multi-clausal sentences in their conversational interaction with family and peers (Brown, 1973) containing all types of subordinate clauses, including nominal, relative, and adverbial clauses (Diessel, 2004). Because of this, syntactic development beyond the childhood years might appear to be less obvious that lexical development. However, inquiry into later language development concerned with syntactic development in late childhood

and adolescence has revealed that great deal of syntax remains to be learned in this period (Scott, 1988), both at the intra-sentential and inter-sentential level (Karmiloff-Smith, 1986). The use of simple coordination and of marked juxtaposition decreases with age, whereas coordination with ellipsis of subject, finite and non-finite subordination increases in children's production with age and in the course of schooling therefore (Mazur-Palandre, 2006). This lengthy development of syntax beyond the preschool years has been hypothesized to rely on cognitive stimulation, expanding knowledge base, and acquisition of abstract thought all of them considered key factors contributing to this process (Loban, 1976; Moffett, 1968). From a usage-based perspective, "discourse provides children with a developmental mechanism for the acquisition of linguistic devices" (Hickmann, 2003, p. 335)– in the present case, complex syntax, whether in conversational interaction or in monologic productions, whether spoken or written. However, increasing levels of experience with the written language as a discourse style play a very special role in the development of linguistic complexity.

Writing activities allow the child writer to operate without the constraints of spoken production, and therefore facilitates the production of planned, formal (written) discourse. As a consequence of reading and writing activities, children increase their efficiency at accessing and processing complex structures. Thus the concurrence of more advanced stages of cognitive development with extensive experience with literacy based activities and with reading and writing of different types of (academic) discourse would enable the speaker writer to integrate new information into existing knowledge systems, supporting and promoting the production of, progressively more complex, tighter more cohesive monological pieces of text. In contrast, spoken texts continue to hold a more interactive, expressive orientation. When speaking, children tend to produce loosely connected texts, including a high proportion of juxtaposition, and parenthetical asides, showing higher reliance on discursive connectivity than on strict syntactic linking of clauses. Writing serves as a platform for constructing text-embedded complex structures that might eventually be translated to children's spoken language (Jisa, 2004).

The way a speaker/writer uses syntax is a major contributor of linguistic complexity (Crystal, 1996), certainly a milestone of later language development.

The process of attaining syntactic maturity (a process that extends into adulthood) has been characterized by gradual increases in the length and complexity of spoken and, most particularly, written productions (Nippold, 2007). Thus, using a high number of clauses is characteristic of more embedded text structures such as subordinate and relative clauses, which makes possible the expression of more complicated relationships among ideas (Coirier, 1996). Clearly, clause and not sentence, is the most appropriate unit of analysis. In fact, the very notion of sentence as a viable unit of writing, as opposed to other units such as a fragment or run-on sentences, both perfectly admissible in oral conversation, would be acquired through reflective experience on text writing (Berman & Ravid, 2009).

Also, the use of an increasingly higher number of words per clause would be related to the intra-clause level of complexity. A high (mean value of) clause length has been related with syntactic structures associated with linguistic literacy, such as nominalizations, attributive adjectives, non-finite subordination (using infinitives, participles, or gerunds), passives, conjoining, and prepositional phrases. All these devices, allow the (speaker) writer to compress several propositions into a single clause (Chafe & Danielewicz, 1987; Scott, 2004). The use of these types of constructions increases with age, and so increases consequently the number of words per clause in written texts produced by school graders. This pattern of increase, however, shows more prominence in high school and adult writing (Hunt, 1970) than in middle school (and younger) writers. This is seen as supporting the view that language development is a process that spans throughout life.

Notwithstanding the wide variety of parameters accepted as indicators of grammatical complexity, hypotaxis, that is, the ability to express hierarchical relationships between clauses by embedding one into another, unlike the more linear-like chaining produced by parataxis and juxtaposition, is the most profusely used. The ability of producing complex subordinate sentences makes possible the expression of more complicated relationships among ideas (Coirier, 1996). Expressing manner, temporal, conditional and cause-and-effect relationships among others, for example, often require the use of subordinate clauses (and conjunctions). Likewise, mental state and speech act verbs that characterize a person's attitude toward a proposition typically take subordinate clauses (Olson & Astington, 1990). Research has documented how the ability to express increasingly

abstract ideas in longer sentences containing more marked, less frequent constructions such as passive, middle-voice, impersonal constructions and non-finite subordination, and multiple and embedded subordinate clauses continues to develop throughout the school-age years, and into adulthood (Berman, 2004; Berman & Verhoeven, 2002; Friedmann & Novogrotsky, 2004; Jisa, Reilly, Verhoeven, Baruch, & Rosado, 2002; Jisa &Viguié, 2005; Loban, 1976; Nippold, Mansfield, & Billow, 2007; Ravid & Saban, 2008; Ragnarsdóttir & Strömqvist, 2005; Scott, 2004). In similar lines to what has been shown regarding length of clause, the use of complex syntax takes a lengthy developmental pattern. Research on extended text writing and consistently across genre (narrative and essay) from 1st to 7th grade, the most frequent syntactic construction was the single independent clause. In contrast, single independent clauses introduced with a coordinating conjunction, considered an immature form of writing, occurred far less often. Constructions involving two independent clauses were more frequently connected by coordinating than by correlative conjunctions. Among syntactic constructions involving an independent clause and a dependent clause, relative clauses occurred the most often in essays. Subordinate clauses occurred the most often in narratives, and even adverbial clauses, a type of subordinate clause that occurred rather rarely, was more often used in narratives than in essays (Berninger, Beers & Nagy, 2010). Interestingly, these same authors found that whether children were asked to write an extended piece of writing or just a single good sentence was related to use of complex syntax. Thus, extended writing elicited use of less complex (multi sentence) constructions whereas single sentence writing elicited more instances of an independent clause plus a dependent clause.

Construction of a subordination index (an index that computes the number of main and subordinate clauses, per T-unit or clause package) is a frequently used measure of syntactic complexity. However, according to Scott (1988) there have been division of opinions as to what exact structures to include among the subordinate clauses. An indirect problem with the subordination index is that it has been acknowledged to overshadow the more complex picture of syntactic development that includes several other important features, for instance, adding discourse-structuring devices like adverbial connectives to the repertoire (Scott, 1988), noun and verb phrase expansion, and usage of the expanded noun phrase in

new grammatical roles (other than as post-verbal elements). Loban (1963) contended that syntactic development during the school years is concentrated at the phrase level rather than at the clause level (cited in Scott 1988: 68). However, to date this statement remains debatable due to scarcity of data on how development and discourse type influence use and frequency of different subordinate clause types.

The development of noun phrases as a relevant feature of development of complex syntax has also been examined by research, although to a lesser extent than use of subordination. Noun phrases have been seen as a platform for constructing broader, discourse embedded syntactic architecture. Although the grammar of noun phrases, mid-level in size, smaller than a clause or sentence, but easily extendable beyond a single word. in different languages is basically in place by 3 years of age (Radford, 1990; Slobin, 1985), new elaborations in both the noun phrase head and its associated modifiers, take place during the school years reflecting related developments in lexical repertoire, syntactic proficiency, and communicative competence. Research has established that reliance on elaborated lexical noun phrases emerges as a relevant means for evaluating the increased complexity of language use during the school years. From middle childhood on, noun phrases grow longer in words, they include more, and more varied types of modifiers, they reveal greater syntactic depth, and they employ semantically more abstract nouns as heads. Importantly, the study of the development of noun phrase structure during the school years sheds light on syntactic acquisition from middle childhood to adolescence in typologically distinct languages (Ravid & Berman, 2010).

In a usage-based perspective, the acquisition of the wide repertoire of later developing syntactic structures is considered to be driven by discourse, that is a child's motivation for using one particular structure or construction derives from the broader discourse context as well as from the child increased awareness of text function and sense of audience (Nippold, 2007; Scott, 2004). Schoolers as young as 9 have shown some genre sensitivity at the level of syntactic usage, for instance, they produced longer clauses when writing an expository than a narrative text (Scott & Windsor, 2000). In addition to producing longer clauses, older student writers used more complex noun phrases (Malvern, Richards, Chipere & Duran,

2004; Ravid & Berman, 2010), more nominalised forms (Schleppegrell, 2004) and more relative and adverbial clauses (Scott & Windsor, 2004). That is, the true hallmark of later syntactic development is not just the production, in itself, of longer sentences containing heavy noun phrases and/or multiple and embedded subordinate clauses, but the increasing ability to integrate these utterances into organized and sustained pieces of more or less formal pieces of discourse, in genre appropriate ways (Bates, 2003; Berman & Verhoeven, 2002; Hunt, 1970; Jisa, 2004; Loban, 1976; Nippold et al., 2005; Nippold, Mansfield, & Billow, 2007; Ravid & Tolchinky, 2002; Verhoeven, Aparici, Cahana-Amitay, van Hell, Kriz & Viguie Simon, 2002). Thus, effective use of (complex) syntactic structures is one of the many requirements of a well written text. The relation between writing quality and complex syntax is not, however, a straight one, both because there are many other factors that contribute to quality, and also because the contribution of syntactic complexity to text quality is only measurable in terms of genre-dependent criteria (Beers & Nagy, 2009).

The corpus CesCa consists of more than two thousand texts of different types written by children and adolescents ranging from 5 to 16 years, that is in the period of compulsory school. Some of the texts, i.e., the explanation of a film, represents the narrative genre that is commonly practiced at school. Some other texts, i.e., the recommendation of a film and the joke telling are not as habitually practiced in writing. These differences between the types of text yielded different patterns of text-embedded lexical growth and development.

The children produced more complex and sophisticated lexical items in their explanations than in the other two types of texts. Recommendation was a somewhat border type of text, ranging from very formal detached texts to highly involved spoken-like ones. Overall it did not contain as high a proportion of complex words as explanation or definition. However, it was the preferred type of text for use of adjectives, a lexical category that indicates later lexical development and that can foster denser noun phrases and complex syntactic frames therefore. Joke telling persistently figured as the type of text involving less complex lexical uses. The next step was to examine whether similar patterns would arise for syntactic uses. In addition to analysing syntactic subordination, probably the most habitual measure of syntactic complexity, we also analyzed noun phrase

complexity and mean clause length given that both these measures have show higher reliability for the characterization of written texts produced by children in elementary and middle school. In fact, given that the lexical characterization had revealed three peaks of lexical growth in 2nd 6th and 10th grade, we decided to analyse the text-embedded syntactic constructions in the same texts in order to examine the relationship between lexical and syntactic uses.

1.3.1.1.3 The contribution of spelling to later language development

In contrast to spoken language, written language is a recent cultural invention, which did not exist until some 5000 years ago (Rayner & Pollatsek, 1989) and which, again unlike spoken language required explicit instruction to emerge. If a few centuries ago, only a few percentage of the population had the privilege of being taught to read and write, nowadays the ability to produce and understand written language is crucial for successful participation in our technology-based society (Snowling, 2000). In absence of such skills one is sure to drop out at school, as well as likely to experience problems obtaining and retaining a job with the subsequent emotional distress.

Spelling is a critical aspect of written language proficiency and is, without doubt, one of the most significant challenges of a child's early academic life. A wide, deep knowledge base underlies what on the surface may seem like a 'simple' skill" (Joshi, Treiman, Carreker & Moats 2009) as established by cognitive theories of the development of the writing process (Fayol, 1991, 1999, 2004; Hayes, 1996; Hayes & Chenoweth, 2006, Hayes & Flower, 1980) pointing that the translation of ideas into writing involves several levels of language. A part of this process requires that a child writer draw on transcription processes at the word level (spelling) and also at the subword level (handwriting) (Berninger, Yates, Cartwright, Rutberg, Remy, & Abbott, 1992, Richards, Berninger, & Fayol, 2009). Lack of automaticity of transcription processes seemingly affects productivity (Graham, Berninger, Abbott, Abbott & Whitaker, 1997). Transcription abilities have been found to explain performance at the sentence level. In second grade, only spelling, which shares common variance with morphological knowledge (Carlisle, 1994; Carlisle & Nomanbhoy, 1993; Garcia, 2007; Garcia, Abbott, &

Berninger, 2010) and marks parts of speech (grammar) (Nunes & Bryant,2006; Nunes, et al., 1997, 2006 ; Tyler & Nagy, 1989 , 1990), explained unique variance in sentence combining. In third grade, subword- (handwriting automaticity) and word- (spelling and morphological signals) levels explained unique variance in sentence combining. Spelling difficulties have been shown to have detrimental effects on the quality of a text. Poor spellers may exhaust their cognitive resources merely figuring out the correct spelling of words, while paying little attention to other aspects of text construction. Conversely, writers who are not preoccupied with their spelling may devote more time and energy to developing text content, and to revision and rewriting.

Spelling is a complex skill that interacts with other knowledge necessary for efficient written communication. In particular, learning to spell involves understanding the relation of the graphic elements with the different levels of language: phonology, morphology, syntax, and the lexicon. The way a particular language is spelled is the orthographic system of that language. Learning the conventions of the particularities of the (alphabetic) orthographic
system of a language entails different types of progress.

Orthographies of different languages can be put on a depth continuum according to their degree of phoneme-grapheme consistency. More transparent orthographies have almost perfect mapping of phonemes onto graphemes, whereas in more opaque orthographies the same letter can represent more than one phoneme and the same phoneme can be represented by several letters, depending on its context (Frost, Katz & Bentin, 1987).

During recent years much research has been dedicated to comparing the acquisition of written word processing mechanisms in deep and shallow orthographic systems (Seymour, Aro & Erskine, 2003; Snowling & Hulme, 2005, Treiman & Kessler 2005; Ziegler, Bertrand, Toth, Csépe, Reis, Faisca, Saine, Lyytinen, Vaessen & Blomert, 2010). To compare different Systems is of great relevance due to the fact that previous models of reading and spelling were primarily based on studies carried on in English, which possesses a markedly dee orthographic system (Seymour et al., 2003). There is reasonable doubt about the advisability of generalizing English-based models to other systems (Share, 2008). The dual-route model (Tainturier & Rapp, 2001, for a overview of this theoretical

framework and relevant evidence) assumes the existence of at least two processes for spelling: A lexical process, which relies on accessing word-specific memory (Barry, 1994) and may be semantically mediated (Hillis & Caramazza, 1991) or may involve direct connections between phonology and orthography (Patterson, 1986), and a sublexical process, based on phonological-to-orthographic conversionrules (Tainturier & Rapp, 2000). According to this model, word identification in shallow orthographic systems is based on phonological prelexical computation (Frost, 2005; Kats & Frost, 1992, 2001; Ziegler & Goswami, 2005), a procedure which is absolutely insufficient for identifying most of the words in a deep orthographic system. From a different perspective, neurological studies comparing languages with contrasting orthographies in terms of transparency (Italian and English) have shown that reading generates activation of different brain areas. Thus, redaing in Italian produced greater activation of areas associated with phonological processing whereas Redding in English caused greater activation of the areas involved in semantic processing and word naming (Paulesu, Demonet, Fazio, McCrory, Chanoine et al., 2001). However, some authors have pointed out that the contribution of lexical access in word reading might be in fact masked by the rapidity of sub-lexical phonological word decoding in a transparent orthography (Dehaene, 2007).

Although phonological skills have been assumed to play a very important role in the initial phases of learning to spell (Ehri, 1997), full reliance on grapho-phonemic knowledge, however, would render adequate spelling in very few, if any, orthographies. Learning to spell is a linguistic process that involves understanding and learning to perceive, integrate and map onto orthographic segments linguistic information at different levels: phonology, morphology, syntax, and the lexicon .Certainly, the spelling of most words can be resolved by applying linguistic knowledge in relation to one or more levels of language, in accordance to the orthographic conventions specific to each system. However, in some cases, the correct spelling can solely be resolved through rote learning. Thus, learning to spell goes the mere acquisition of school-learned skill, but rather consists of uilding knowledge about the nature of the particular orthography as a notational system in a number of dimensions, integrating grapho-phonemic links, orthographic–internal consistencies, and aspects of morphological units encoded

in the system (Ravid, 2001).

1.3.1.1.3.1 *Spelling and phonology*

In alphabetic orthographies, the graphic units represent the consonants and vowels in a language, in a more or less consistent way as seen above. This is different than saying that these systems transcribe the *sounds* of speech, as the sounds people produce while speaking are highly variable and subject to personal and regional variations that are not captured by spelling. Rather, orthographies represent categories of sounds, *phonemes*, which are abstract entities. Thus the connection between phonological awareness and literacy acquisition is not unidirectional. As Olson (1996) suggests, literacy acquisition involves the learner in learning to hear, and to think of, the sounds of their language in a new way. Although it is commonly said that alphabetic systems are based on letter-to-sound correspondences or phonographic correspondences, they are not the sounds that speakers writers hear but the sounds they learn to conceptualize. Phonological awareness measured before the children start to learn to read is a strong predictor of children's progress in reading and writing (Bradley & Bryant, 1983) And learners of alphabetic systems show significantly higher levels of awareness of phonemes than learners of non-alphabetic systems (Read, Zhang, Nie & Ding, 1986). That is why the orthographic representation of the words of a language is kept stable, in spite of speaker's different accents, voices, and intonations.

1.3.1.1.3.2 *Spelling and morphology*

Orthographies do not represent only phonology, but also other levels of language such as morphology. Languages represented by deep orthographies include many examples of orthographic regularities accounting for representation of morphological segments. For example, in English the string of letters *-ing* does not represent just a string of phonemes; rather, it altogether represents the progressive aspect. French is another language with a deep orthography in which orthographic representation of morphological segments is particularly challenging for beginner spellers since quite often the orthographic segment is mapped onto a

silent morpheme. Thus, there is no difference between 'chante' and 'chantent' in spoken French, but the plural is marked by –nt in the written form (Totereau, Thenevin, & Fayol, 1997). But orthographic representation of morphologic segments happens also in languages with less deep orthographies. For instance in Catalan, letter -r is used for spelling the phonologically empty infintive morpheme. The writer needs morphological knowledge, or at least this type of knowledge will aid him in spelling any word with the same infinitive function correctly. Spelling thus requires that children grasp morphological information, over and above knowledge of phoneme-to-grapheme mappings.

1.3.1.1.3.3 *Spelling and syntax*

The writer may also need some level of syntactic awareness in order to render correct spelling. This holds for languages with deep orthographies such as French in which many word endings are not pronounced but they still need to be written down so that correct syntactic agreement between, for instance, determinant and noun is realized. For example, the difference in number in the noun phrases although both the singular and the plural form of *la pomme,* 'the car' and *les pommes,* 'the cars' is realized in oral language in the determinants *la* /la/ and *les* /le/ the difference in number is not realized in the noun *pomme* /pom/ and *pommes* /pom/. Therefore, the writer needs to use his knowledge that the noun and the determinants must agree in number. Otherwise, he is likely to misspell the form inflected for number *les pomme. As we have seen with spelling and morphology, the writer may need syntactic awareness in order to render correct spelling in a language with a moderately transparent orthography such as Catalan. Let's consider, for instance three different orthographic representations *s'hi pot anar* /si'pota'na/ 'it-there can be gone to - it can be accessed', *si hi pot anar* /si'pota'na/ 'if you can go there', *sí pot anar* /si'pota'na/ 'yes you can go'. In all three phrases, *s'hi*, *si* hi and sí are mapped onto the same phonological segment /si/. Therefore, phonographic mapping does not suffice and the writer needs to be aware of the syntactic function of the segment he is writing in order not to misspell it.

1.3.1.1.3.4 *Orthographic patterns and rules*

Each alphabetic writing systems may also have a set rules governing the combinations of letters that are legal in each orthography. These patterns establish the legal combinations of the letters in the alphabet at the syllable, word, and inter-word level. For example, in Catalan, the letters <n> and <s> may appear together within a word, as in *co<u>ns</u>ell,* 'advice,' or at the end of the word, as in *avio**ns**,* 'planes'. In contrast, Spanish allows <n> and <s> to appear together only in the first context, but not at the end of a word, as in *co<u>ns</u>truir, '*to build'. Orthographic rules determine restriction of use of a phonographically legitimate letter in a particular intraword context. Thus, in Catalan –s correspond with both /s/ and /z/. However, when the sound /s/ occurs in intervocalic position within the word, letter –s no longer represents /s/ which must be represented by –ss. Clearly, the phenomenon of letter use depending on word context is not restricted to transparent orthographies. Non-phonemic orthographic constraints determine that, in English, the sound /k/ can be written with <c>, <k> or <q> at the beginning of a word, e.g. *cake, kite, quiet,* but only <k> and <ck> are possible at the end of a word, e.g., *sock, book*. In short, orthographic conventions determine the correct spelling of words in all orthographies, beyond the application of phoneme-to-grapheme correspondences. Although native children may develop knowledge about the legal letter strings in their language implicitly, knowledge of the context dependence rules must be taught to children through explicit instruction.

1.3.1.1.3.5 *Spelling and the lexicon*

Finally, sometimes the correct spelling of words, especially in very inconsistent orthographies but also quite frequently in consistent ones, may only be rendered if the writer has established the orthographic word in his orthographic lexicon by memorizing the specific items. The writer must know that the word *verd* /bɛr/ 'green', in Catalan, is spelled with –v, instead of the phonographically possible –b, He must also know that *vermell* /bðrmɛλ/ 'red' is spelled with –e in the root instead of it phonological counterpart /*ð*/. This phenomenon is particularly

marked in deep orthographies. In English, *muscle,* and *tongue* cannot be resolved by applying phonographic, orthographic, morphological, or syntactic knowledge; it is only the knowledge of the lexical item itself that leads to the correct spelling of such words.

In addition to the importance of spelling as a component of writing, it also has important social implications, as written expression sometimes serves as the first or sole means of contact among people. Spelling gives a decisive first impression and is the one aspect of writing that is readily evaluated and judged by others. For these reasons, spelling also receives a lot of attention on the part of educational research. Thus, teachers will sure benefit from a deeper understanding of differences across orthographies, and the implications for literacy instruction. In particular, teachers and practitioners should become aware of research findings that show that "orthographic differences do make a contribution in the acquisition of literacy skills and that certain orthographies can slow down literacy acquisition in beginning readers" (Joshi, 2010). Also, teachers and practitioners should learn of research findings that show what types of instructional practices have or have not reported significant benefits to children's development of spelling abilities Rieben, Ntamakiliro, Gonthier & Fayol (2005).

Unlike many of the experimental works focussing on spelling and developmental command of the different orthographic systems, we take a corpus-based approach to tapping the developmental pattern of spelling through grade school. Notwithstanding the limits of this type of methodological approach to such an object of study as spelling, we think the many relevant insights it can contribute make it worth it, at least as a first approximation to the field study. Spelling is not, or at least is far more, than a varnish with which to give the written product a conventional surface. Spelling is part of the writing process, it involves punctual segmentation of the string of speech, or thought, into the words it consists of, consideration about the morphosyntactic role of the word being written, attention to the constraints imposed by both the spelling and the orthographic system. For all this, assessing spelling on written products actually produced by the child makes plenty of sense. In the present work, we examine how spelling develops following different patterns depending on the type of knowledge the child needs to

resort to in order to produce accurate spelling of the words he has decided to write down. This method produces some interesting insights: for instance, children seem to be quite aware of the salient morphological features of the Catalan language, although, to the best of our knowledge, they are not a fundamental part of the spelling instructional practices. Instead, phonographic errors persist across the board, sometimes because children are tackling spelling through phonological analyses only, a strategy that does not suffice for coming up with the required conventional form of the word. Or lexical knowledge remains poor through grade, suggesting, perhaps, that reading and writing are seen as separate independent practices.

1.3.1.1.4 Some considerations on the cross sectional role of morphological knowledge

In everyday language use, we know (store, retrieve, and use) words, just as simple one-piece forms, not as sums of their parts. Yet, although we know words as wholes, we also know them by their parts (Bender, 1968). Understanding when and how do children come to acquire the internal structure of words, what knowledge they acquire or how well they are able to use such knowledge will make an important contribution to a more overall understanding of later language development.

According to Dressler, the degree of (inflectional) morphological richness of a language constitutes the most important typological characteristics of that language (Dressler, 2004, Laaha, 2007). Acquisition of inflectional morphological processes in L1 is related to the typological characteristics of the language. That is, children are sensitive to the typological properties of the language they are acquiring; they are sensitive to the relative communicative importance and structure of morphology in their verbal interactions. Children's patterns of acquisition show that they can process some kind of information more readily than others, thus children learning different languages types typically follow similar timelines (Peters, 1995). Several studies have examined the impact of the saliency (or lack thereof) of morphology on spelling acquisition, and suggest that awareness of the morphological features of a morphologically rich language is

pushed forward by spelling acquisition in an alphabetic language (Ravid & Gillis, 2006).

Research suggests that derivational words might be in general acquired somewhat later than inflected (and compound) words (Berko, 1958, Clark, Hetch, & Mulford, 1986). In fact, there is evidence that in the earliest period of language acquisition, the child neither engages in morphological analysis nor combines morphemes when producing words, that is that all words are psychologically monomorphemic (Miller, 1991). Later on the child's vocabulary begins to incorporate multimorphemic words such as nouns or verbs marked with inflection. With age, lexical development becomes increasingly characterized by growth in morphemic complexity with increasingly complex forms being incorporated to vocabulary as children learn more (and more about) language (Clark et al., 1986). Superior word learners show particular skilfulness at analyzing derived words into morphological components and they particularly apt at using derivational knowledge to learn new words. Anglin (1993) establishes the importance of such ability in order to extend one's lexicon through the incorporation of what he call the potentially knowable words. That is, morphologically complex words whose meaning need not be learnt by rote memory and can, instead, be deciphered by what Angling named the morphological problem solving.

By its very nature morphology cuts across formal boundaries laid down by linguists (McClelland et al., 2010). Thus, notwithstanding the importance of the role played by morphological awareness in vocabulary acquisition throughout the school years, or maybe as a consequence of it, such role pervades other domains of later language development. As early as in first grade, morphological signals, a word-level feature, was the only that explained unique variance in a sentence combining task. Thus, morphological affixes that mark grammar functions (Nunes, Bryant, & Bindman, 1997 , 2006) were found to be related to sentence combining. It has thus been argued that morphology serves a unique scaffolding function within and across levels of language and creates a bridge across the word and syntax levels for relating word-level suffixes marking grammatical function to sentence syntax. Additionally it creates a bridge across spoken and written words at the word -level where morphemes correspond to both phonology and

orthography (Berninger et al., 2010). Later on, highschoolers and specially adults substitute dynamic, concrete verbs and basic adjectives by more abstract and morphologically complex nouns derived from those same verbs and adjectives. Because of such shift in lexical perspective the noun phrases making up the clauses become more heavy, more complex therefore. As head of large and complex noun phrase architectures, derived nominals, which are almost always modified in adult discourse, take on the textual role of promoting the flow of information in the text. In sum, extended school age vocabulary gains morphological complexity and attracts syntactically elaborated constructions in service of discourse functions (Ravid, 2012, Ravid & Levie, 2010). In the dynamic, ever evolving system that is grammar under the constant reorganization driven by one's ongoing experience with language, this grammar built up from specific instances of use that marry lexical items with constructions, and it is interesting to note that almost all constructions contain some explicit morphological material, tying them fairly concretely to specific words or morphemes. In use, grammar is routinized and entrenched by repetition and schematized by the categorization of exemplars (Bybee, 2010).

In sum, the lexical, syntactic, spelling –and morphological-- nexus illustrates how skilfully mature language users assemble and incorporate different linguistic constructions and items with the purpose of generating a richly weaved discursive texture. Although the present work does not include a specific study on morphological development throughout compulsory school, morphological knowledge has revealed itself a key component of the overall process of later language development in Catalan. Increase in word length and in use of nominalizations, both of them related in Catalan to morphological processes of derivation, have proven to be major indicators of later lexical development. Use of nominalizations, in turn, involve literate abstract terms supporting dense complex syntactic architectures, or heavy noun phrases (a site we have found to be a powerful platform for development of complex uses of syntax). Finally, our work on spelling shows that children become aware of the morphological status of derivational and flexional morphemes and use this knowledge in their spellings from very early on.

1.4 Major contributions of this work.

Overall, this thesis aims at contributing to the research field of later language development, a field devoted to better understanding how the human capacity for language development remains active well beyond the school years, most likely throughout the life span. The relevance of which is solidly established by the breadth of the international research community working on it.

Against this frame the primary contributions of this thesis are:

1. To provide with a corpus of written language in Catalan (CesCa) and subsequently with corpus-based studies on later (Catalan) language development. The corpus is of public access at http://clic.ub.edu/corpus/cesca. A main advantage of disposing of a (public) corpus is that it provides the research and educationist community with an authentic picture of language as it is used by its (speakers) writers for different purposes. The CesCa corpus is unique in Catalan: no comparable sampling was available since other existing corpora in Catalan are compilations of texts written by expert professional writers (AnCora-CA)) or texts written in ancient Catalan (CICA).

2. To provide a characterization on later language development in *Catalan*, a romance language so far little examined from this research perspective. Since focus was mostly on English since not far ago, the interest of expanding research over other languages has gained weight over the years. The typological characteristics of the language have been shown to have an effect on some of the language development processes. For instance, Catalan's morphology is far richer than English and the child's keeping up with flexional and derivational processes plays an important role in word learning, intra-sentence agreement (and coherence therefore), and spelling. Instead, compounding has little interest for a language as Catalan.

3. To provide a characterization of a language, Catalan, that exhibits an unusual peculiarity, among the European languages. It concerns the bilingualism or multilingualism exhibited by almost every single speaker of Catalan, since this language is co-official with Spanish in Catalonia in northern Spain. In this background, Catalan is used exclusively for instructional purposes through an immersion program in every school in Catalonia. Children therefore develop their literate uses in Catalan although many of them do not use Catalan almost at all for social or family communication out of school.

4. To expand later language development research beyond the narrative/non-narrative divide since it focuses on four different genres: narrative, argumentation, colloquial (represented in the corpus by a joke telling) and definition (including definitions of three different lexical categories: a noun, and the more rare of a verb and an adjective). The participants therefore produced texts serving quite distinct communicative goals and ranging from very formal, detached school based texts such as explanations and definitions to informal, involved, somewhat oral like texts such as recommendations and joke telling.

5. Finally it contributes to the research on the spelling domain with research grounded data on Catalan orthography, which holds differences in terms of transparency with other better know ones, i.e., French, Spanish and Italian. The contrast between transparent (Spanish and Italian) and opaque (French) orthographies has been exploited in several studies. Catalan orthography is less transparent than Spanish and Italian but also less opaque than French. It this offers the possibility of testing the presumed effect of transparency, for instance on issues related to the role of morphological awareness on spelling, on a moderately different context.

CHAPTER 2

Corpus Cesca:

Compiling a corpus of written Catalan produced by school children

---

Anna Llaurado, Mª Antonia Marti and Liliana Tolchinsky

University of Barcelona

In press in *International Journal of Corpus Linguistics*

Abstract: This paper outlines the compilation of a corpus of Catalan written production. The CesCa corpus 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 presents a picture of the state of knowledge of the Catalan written language throughout compulsory schooling and is of public access at http://clic.ub.edu/es/cesca, Possible uses of the corpus for future research are suggested.

Key words: written Catalan, lexical development, vocabularies, narrative, argumentative, jokes, word definitions.

2.1. Introduccion

Corpus linguistics makes it possible to obtain samples of authentic language uses in different contexts. It might 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 register. Researchers and educationists interested in language development in Catalan, however, have never had a comparable sampling of authentic productions at their disposal. Other existing copora 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 (CCCUB). Our purpose was to fill this gap by a compiling a corpus of vocabularies and texts written in Catalan from late childhood and throughout adolescence. Thus, none of the existing Catalan corpus is comparable with CesCa.

Analyses of these productions 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 productions responding to a variety of comunicative purposes due to two main features of later language development: 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 stonger when written markers are added to the comparison. While early language development centers on the acquisition of phrase structure and discurse uses are mostly limited to intimate/familiar contexts, later language development increasingly expands toward a variety of adressees and communicative functions. We decided to reflect these two features by sampling written language uses responding to a variety of comunicative purposes. Thus, we gathered texts representing different genres of discourse; that is, texts that respond to different communicative purposes and are therefore expected to present different global organizations 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 leisure activities fosters the use of verbs and 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 the tokens in the corpus were stored in a database and lemmatized and annotated for POS morphological features.

In the following, we describe the process by which the corpus was obtained (Section 2). This section includes a description of participants and elicitation procedures. Next, we comment on the storage and processing of the corpus (Section 3) regarding both database creation and data analysis. We then proceed to show some details concerning the general characteristics of the corpus (Section 4) and the configuration of its linguistic units (section 5). Finally, we suggest some directions for future research in which the corpus can be put to use (section 6).

2.2 Obtaining the corpus

CesCa is acorpus of Catalan written vocabularies and texts composed by school children in 2006. 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 .

2.2.1 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 separate facts. Firstly, 42 children attending kindergarten were excluded from the text writing tasks due to their notable difficulties with writing. Secondly, 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 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).

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

School level	Distribution of participants according to home language/s				
	Catalan	Catalan/Spanish	Spanish	Other languages	Total participants
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

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 et al., 2012, in press).

2.2.2 Elicitation procedure

The 32 involved schools 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. 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.

2.2.3 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: 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 a film with the instruction: *"Tell the story of a film or TV series that you like and tell it"*

T.3. For eliciting an argumentative text, participants were asked to recommend a film, or TV series following 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).

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

2.3. 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 (one of 32), to the language or languages participants identified as their home language (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 the five semantic fields and one element of each category of text. The relation between each word and the text it appears in is never lost.

2.4 Processing the corpus

2.4.1 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 exactitude. 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.

Table 2 shows the different variants (flective, orthographic, etc.) subsumed under the canonical form *Malaltia* 'illness'

Table2. Example of different types of variants under one canonical form

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

2.4.2 The corpus of texts.

The texts are available in three different formats following the first handwritten original: a mirror version in digital format, a normalized 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 3)

3. *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 4).

4. *unsenyor va alas palucaria (…)* 'aman goes tothe hairdresser'

c. Illegible characters have been transcribed as an asterisk (Example 5).

5. *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 has been 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 6).

6. *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 7).

7. *unsenyor* 'aman' → *un senyor* "a man"

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

Third, in the labeled 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 8).

8. *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 characteristic 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 9)

9. *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 10).

10. *enfermetats* Spanish stem for *enfermo* 'ill' + -etat Catalan suffix → lemma: enfermetat

c. Illegible forms were lemmatized by the form as it was (Example 11).

11. 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.

2. 5 Configuration of the corpus in terms of linguistic units: tokens, types and lemmas.

The *vocabularies* were constituted by lexical forms, that is, word or multiword units, from five semantic fields: food, clothing, leisure activities, personality traits and natural phenomena. The vocabularies comprise 242,404 lexical forms and 44,049 different lexical forms (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 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	72014	50226	51234	34995	33935
Types	9842	7095	12561	8795	6930
Canonical forms	1856	791	2235	2471	1864

The textual data consist of a total of 11,332 texts (out of 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 lemmata. 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 lemmata 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
Lemmata	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, probably due to different reasons. The rather reproductive character of joke telling may explain its wordiness. The fact that narrative is the earliest acquired genre (Karmiloff- Smith, 1992) and also profusely practiced throughout schooling may, in turn, 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: tokens, types and lemmas, provides interesting insights in terms of both the lexical and conceptual richness of texts. In Table 5 we present 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 obtain the lowest type/ token ratio (.72), followed by narrative (.78) and argumentative (.87) data and, finally, definitional data, which 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 ratio patterns perform in precisely the opposite manner. 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 lemmata 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 lemmata.

2.6 Lema/inflectional and orthographic variants ratios by school level.

The ratio of lemmata to the number of inflectional variants provides a measure of morphological richness. It provides information on the productivity of lemmata by school level in the corpus. On the other hand, the ratio of lemmata 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.

Table 6. Morphological richness and orthographic variants by school level

School level	Kinder	1st. gr.	2nd gr.	3rd gr.	4th gr.	5th gr.	6th gr.	7th gr.	8th gr.	9th gr.	10th 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

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.

2.7 Possible directions for research using CesCa

The CesCa corpus presents some particularities that make it a valuable database for linguistic, psycholoinguistic and educational research. On the one hand, it is unique in the Catalan language, 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, for both vocabularies and texts. In particular, 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.

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 including lexical and discourse uses. First, the developmental trends of a crucial component such as the lexicon have been analyzed both in text-embedded contexts and in isolated vocabularies revealing the extent of the impact of both school grade and

comunicative purpose (defined by type of text in the text-embedded lexical uses and by sembantic 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.

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 perplexity 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 perplexity. In terms of vocabulary, jokes contain more *unknown words*; therefore, the use of neologisms was considered to be higher in this type of text. In another vein, the CesCa corpus has widened the research possibilities in definitions and definitional skills, a field that has mainly focused on noun definition. The CesCa corpus also 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.

In the near future, the corpus will be analyzed regarding the developmental patterns of syntax in texts serving different comunicative purposes. Information regarding the pattern of increase of syntactic complexity with schooling will be obtained.Also, 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.

The corpus is currently being analyzed with the aim of defining the developmental path of spelling in Catalan. In the transparency/opacity continuum defined for alphabetic orthographies, Catalan lays 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 inspelling development.

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 acces.

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, other speak other languages at home, some have been familiar with Catalan since birth others have learned it, through schooling, and 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 L1 and L2 in a multilingual environment.

CHAPTER 3

The growth of the written lexicon in Catalan from childhood to adolescence

---

Anna Llaurado, Mª Antonia Marti and Liliana Tolchinsky

University of Barcelona

Published in Written Language and Literac, *13* (2), 8-22.

Abstract: The lexicon, a complex device of storage of units of language use, is a central component of linguistic knowledge closely tied to grammar and has a strong influence on demanding cognitive tasks and academic achievement. This study aims at tracking the growth of the Catalan written lexicon of children and adolescents throughout compulsory schooling, a time when the lexicon is assumed to experience an exponential growth. There were 2,436 participants from 5 to 16 years old attending compulsory school in Catalonia at the moment of the study. They were asked to produce in writing as many names as they could remember in five different semantic fields: Food, Clothing, Leisure activities, Personality traits and Natural phenomena. Although both the task and the provided examples primed production of single words, participants produced a variety of constructions in the five semantic fields. The 242,404 *lexical forms* that were produced were lemmatized into 8,498 different lemmas and coded according to different linguistic dimensions. The size and the conceptual underpinning of the lexicon grow significantly throughout compulsory school and show an increase in the use of Catalan correct forms, a reduction in deviant forms and a steady use of words and constructions in other languages. The use of multiword constructions as a mechanism for word generation questions the separability between lexicon

and syntax. The corpus, which is of public access (http://clic.ub.edu/es/cesca), provides a picture of the state of a language developing in a multilingual environment in terms of frequency of use of words and constructions and range of orthographic and linguistic variants in five semantic fields.

Key words: lexicon, lexical growth, written Catalan, later lexical development, lexicon/syntax delimitation

3.1 Introduction

The linguistic knowledge of school age children and adolescents can hardly be characterized without taking into account their performance in the written modality (Ravid & Tolchinsky 2002; Tolchinsky 2004). While few significant differences may be found between a 6 year old and an 8 year old in markers of linguistic development, a clear distance appears between the two when the written modality is compared (Nippold, 2007). This study focuses on the development of the written lexicon in Catalan, from childhood throughout adolescence. Certainly, written language furnishes speakers not only with new vocabulary but with new ways of syntactic and rhetoric organization as well as with different means to think about language which have an overall effect on speakers' linguistic competence. But, the lexicon is central to speaker/writer linguistic competence; it is a complex device of some sort of storage of available elements for creating and understanding messages, for responding and making sense of linguistic input both in real and referred time and has a strong influence on highly demanding cognitive tasks such as reading and writing, and on students' academic achievement.

Within generativist and other projectionist approaches lexical entries are conceived as a skeleton on which syntax builds up is built up). A lexical entry is not just an arbitrary pairing of sound/meaning, but includes a variety of formal diacritics that translate into a set of instructions for syntax (Borer, 2005). These views assume a level of representation with well-defined formal characteristics that can be accessed directly from the information in lexical entries along with

combinatorial principles of some other kind. Lexicon and syntax are conceived as two distinct components of linguistic knowledge supported by different learning mechanisms and types of memory. Lexical learning would take place through association mechanisms and be stored in declarative memory whereas learning of syntax would be caused by the triggering of UG or by procedural learning and stored in the procedural memory. The selective dissociation between lexicon and syntax that characterize some kinds of aphasia supports the idea of separability between lexicon and syntax. In the field of computational linguistics some theoretical proposals assume that structural information is obtained from information stored in the lexicon (Pustejovsky, 1995; Pollard & Sag, 1994).

From a usage-based approach (e.g., Bybee, 2007; Goldberg, 2005), the approach taken in the present study, there is not such a clear-cut distinction between lexicon and syntax. A unique way of learning is postulated for both and it is assumed that units of differing size and nature such as morphemes, lexemes, phrases, idioms, etc, work as processing units –i.e., as units of planning, production and perception –. The fact that in early language development vocabulary size is the best predictor of grammatical knowledge, that the relationship between lexical and grammatical knowledge has been established for different languages and even for aphasic patients and for language processing in real time supports this position (Bates & Goodman, 1997).

In sum, though diverging as to the degree of separability between lexicon and syntax, lexicalist and projectionist perspectives agree upon the amount and diversity of information encoded in the lexical entries including information about the syntactic context where lexical items can occur, their combinatorial preferences with other lexical elements and the way they can function in discourse. Lexical knowledge is not restricted to storage in memory of sound/meaning pairs but embraces global language knowledge. Moreover, for usage-based approaches there is no strict distinction between lexicon and syntax because units of distinct complexity are accessed and used depending on the communicative situation requirements. Besides its purely linguistic interest, the study of the lexical component is of importance for a number of educational reasons. Lexical growth protracts beyond childhood and adolescence and well into

adulthood and it constitutes a key facet of later language development (Anglin, 1993; Nippold, 2005). Throughout schooling vocabulary is extended and allows for greater lexical diversity and encoding of more specific concepts secondary meanings of polysemous terms are added to rapidly mastered primary meanings and comprehension of figurative language is developed (Tolchinsky, 2004). Improved knowledge of derivational morphology plays an increasingly important role in the interface between lexicon and syntax (Ravid, 2004). Vocabulary knowledge predicts academic success (Cunningham & Stanovich, 1997; Leong & Ho, 2008) and explains individual variance in reading comprehension (Leong & Ho, 2008; Laufer & Nation, 1999). Frequency of use of nouns and verbs plays an important role in reading speed (Holmes, Stowe & Cupples, 1989) and in reading comprehension as well. Children with reading difficulties usually exhibit a poorer vocabulary than their more skilled peers. Moreover, educational interventions on lexical aspects entail progress in reading comprehension (Nation, Snowling, & Clarke, 2007). Overall, the lexical domain shows in a very unique manner the ways in which context and cognition interact as well as the changes of such interaction with development (Dockrell and Messer, 2004).

3.1.1 Goals of the study

The first goal of the study is to track the development of the written lexicon from the age of 5 to 16 years old. During this period a diversification of children's linguistic circumstances occurs. The family environment is enriched with the introduction of new interlocutors –peers and adults beyond the family– who bring with them different registers and styles. And, a more sustained experience *with writing and written language* takes place from a twofold stance: *written language as discourse style* –the kind of language used for writing as essentially different from the one used for speech; and *written language as a notational system* –the perception and growing command of the representational system that is used in the written modality. Written language becomes a significant additional source of learning about the world and about language (Ravid & Tolchinsky, 2002; Pinker, 1999) and the more children improve their writing skills the more fluently they write and vice versa. All of these should have an impact on vocabulary growth,

most particularly on written vocabulary growth. It is our aim to assess this growth by computing the number of *lexical forms* –i.e., expressions as written by the subject– in different semantic fields. These forms were submitted to a lemmatization process (see "lemmatization criteria" section). Two were the purposes of this process: primarily, to make possible an initial exploration of the conceptual vocabulary. A diversity of lexical forms may be underpinned by the same concept. Thus, we deemed it important to examine separately the diversity of lexical forms and the diversity of lemmas – i.e., the conceptual underpinning of one lexical form or more than one –. Secondarily, as a consequence of the lemmatization process the search through the corpus by outsiders to the study (linguists, psycholinguists, educators) was eased.

A second goal of the study is to explore how this growth is realized in semantic fields that prime different grammatical categories. In the mental lexicon semantic weight is distributed through the range of syntactic categories (Pustejovsky & Boguraev, 1993) and therefore the developmental picture would be distorted if nouns or verbs alone were taken into account. In our study, semantic fields were selected that would each prime use of different syntactic categories both through the instructions and through the provided example. We aimed at having our participants access their stored lexicon in order to retrieve decontextualized lexical units: entities, qualities and activities. Thus we devised a divergent naming task and asked them, for instance, "*write down as many clothing items as you can think of.*" Three of the five chosen fields were represented by the noun category (Clothing, Food, and Natural Phenomena), one field was represented by the adjective category (Personality traits) and one field was represented by the verb category (Leisure activities). The noun category has been favored nonetheless as it is the most represented in the lexicon. Verbs account for some 12% of the total in works of reference such as dictionaries, lexical databases, semantic nets (such as Wordnet), and adjectives account for some 10% and nouns for some 78%. Here, an additional consideration regarding development of adjectives is to be taken into account: studies and surveys of natural language acquisition show that adjectives appear later in child speech than do nouns and verbs (Caselli, Bates, Casadio, i Fenson, Fenson, Sanderl, Weir, 1995) and constitute a low-frequency class in children's early lexicons. Use of adjectives has

proved diagnostic in studies of later language development during school years (Ravid, Levie & Avivi-Ben Zvi, 2003). Moreover, many experimental and simulation studies account for the psycholinguistic processes beyond this sort of representation. It is well established that in lexical recognition and retrieval tasks there exists an important frequency effect: we are faster at naming high frequency words or objects (Harley, 1995; Jescheniak & Levelt, 1994) and sensitive to the phonological neighbors of a word. The more phonological neighbors a word has, the more chances it gets activated (Burke, MacKay, Worthley & Wade, 1991; Harley, 1995). There also exists an important word category effect: generally nouns are better remembered and more easily retrieved than adjectives and verbs.

Our third goal was to tap the linguistic configuration of the corpus regarding the size of the production units. The way elements are stored as well as their size (morphemes, words, sublexical units, lemmas, idioms and units not necessarily corresponding to standard linguistic units) is highly controversial and has been the object of different models (Ravid, 2004), all of which however attest a key characteristic: lexical elements coexist in a sophisticated net of relationships that responds both to form and content motivations. According to Bybee (2007), the representation in memory of a phonetic form is a categorization of the tokens of use and, therefore, it presents a range of variations subject to the linguistic experience/environment of the user. In this view, conventionalized lexical entries for activities, items of clothing or food might be an outcome/result of an abstraction process from common tokens in the informants' environment which may or may not be single words. Following this reasoning, the diversity and complexity of descriptions would tend to decrease with age and increasing experience with conventionalized written language.

Our fourth goal was to explore the impact of the multilingual situation in Catalonia and in Catalan schools in the linguistic configuration of the corpus. Catalan is a language spoken by 27.5% of the Spanish population[i]. It is co-official to Spanish in Catalonia, where it is the language of instruction. Spanish has a strong presence in the environment and the mass-media and the notable increase of immigration during the past decade has turned Catalonia into a truly

multilingual community. Thus, it is relevant to explore the relative permeability of the written lexicon to words in other languages.

Finally, the study has also a fifth -more practical- goal which is to provide psycholinguists and educationists with an authentic and updated picture of the state of the Catalan language as depicted by the vocabulary used throughout compulsory schooling. Unlike frequency dictionaries often based on professional edited texts, the picture provided by this study is extracted from authentic productions by actual non-professional speakers/writers. There are very few similar works for the Catalan language. Borromba (1976) is limited to the basic vocabulary of 3- to 6-year-old children, more recently Cordero (2002) tracks the basic vocabulary of 6- to 14-year-old children using, however, a very small sample.

3.1.2 Predictions

The study was guided by 5 main predictions:

(1) As due to the enrichment of linguistic interactions, and the increase of experience with writing and written language, vocabulary continues to grow throughout childhood and adolescence, we predicted an increase in the number of *lexical forms* or *variants* produced by subjects in every semantic field.

(2) *Lexical forms* are behavioral realizations whereas lemmas capture the conceptual underpinning of a semantic field. We expected an increase in the number of lemmas throughout school for the same reasons that would cause an increase of lexical forms.

(3) Given the way syntactic categories are distributed through the lexicon, we predicted an effect of syntactic category in the amount of variants produced in the different semantic fields. In particular, we supposed there would be more *lexical forms* for nouns than for adjectives or verbs taking into account the priming effect of nouns in processes of lexical retrieval and the fact that in specialized argots, as school discourse basically is, terminology is essentially nominal. Also, adjectives should be fewer not only out of a low proportion in the lexicon but also due to late appearance and slow development in language acquisition (Ravid & Nir, 2000).

(4) Although full consensus as to the degree of separability between lexicon and syntax is still lacking, the amount and diversity of information encoded in the lexical entries is agreed upon from several approaches. If the mental lexicon disposes of a diversity of processing units that are activated when a particular task is confronted, we expect, regarding the size dimension of the linguistic configuration of the corpus, to find a variety of constructions with which to name entities and not just isolated words in spite of the fact that isolated words were primed and exemplified in each semantic field. Descriptions would tend to decrease with schooling, to be substituted by simpler lexical forms. Also, due to increasing exposure to written texts and to classroom discourse, we envision a reduction of the orthographic variants used in naming entities of the different semantic fields.

(5) Finally, regarding the language dimension of the linguistic configuration of the corpus, we expect that the multilingual situation in Catalonia will translate into a notable presence of non-Catalan terms.

3.2 Method

3.2.1 Participants

A total of 2,436 texts were gathered but due to different formal problems (e.g., lack of personal data) the final corpus was constituted by 2,396 texts. They were produced by children and adolescents ranging from 5 to 16 years of age who were attending 31 schools (25 public schools, 5 semi-private schools and 2 private schools) in the four different provinces of Catalonia at the moment of the study. Literacy teaching starts in Catalonia very early. At age five preschoolers are engaged in different kinds of literacy activities including writing without a model and reading of different types of text. The linguistic background of participants was defined according to their responses to open questions about the language/s they speak at home and with their friends. Four groups of responses were identified: participants who declared that Spanish or Catalan was their home language; participants who declared that they use both Spanish and Catalan and participants whose home language (s) are other than Catalan or Spanish. Table 1 shows the

distribution of texts per school level from preschool (P5) up to last year of junior high school (JH4) and the distribution of participants according to their home language/s.

Table 1. Distribution of participants by home language and distribution of texts by school level.

School level	Distribution of participants according to home language/s				Total texts
	Catalan	Catalan + Spanish	Spanish	Other languages	
Preschool (P5)	33	58	21	20	132
1st grade elementary school (ES1)	64	80	42	39	225
2nd grade elementary school (ES2)	40	22	30	10	102
3rd grade elementary school (ES3)	59	71	62	27	219
4th grade elementary school (ES4)	49	34	45	9	137
5th grade elementary school (ES5)	55	86	137	22	300
6th grade elementary school (ES6)	34	55	55	20	164
1st year junior high school (JH1)	50	115	136	6	307
2nd year junior high school (JH2)	56	122	89	12	279
3rd year junior high school (JH3)	43	176	102	8	329
4th year junior high school (JH4)	30	87	791	184	202

The current linguistic situation in Catalonia is represented in our sample. Since 1978, both Catalan and Spanish have had official recognition. Catalan is spoken in Catalonia (and in a few other provinces in Spain, southern France, and the island of Sardinia as well). Spanish is spoken throughout Spain. Importantly, Catalan has been the only language of instruction across all levels and in all public and semi-private schools in Catalonia for at least the past 20 years. During this past decade immigration has experienced an important boom rising from a mere 3% in 2000 to 13% in 2006. Although most recent immigrants come from Latin America

(34%) and have Spanish as their home language, a number of immigrants have come from other countries (15% from African countries, 9% from Asian countries and 5% from non-EU countries) and have vastly enriched the linguistic diversity of Catalonia. Participants were asked how long they have spoken Catalan. Most participants declared they have spoken Catalan either always or for longer than 4 years. Only 3,4% of the participants said they have spoken Catalan for less than one year. Thus, school age children and adolescents know and use Catalan (at school, at least) irrespective of their declared home language/s. Conversely, the presence of the Spanish language both in the social environment and in the media, makes it hard to find a Catalan speaker who is not bilingual in Catalan and Spanish.

3.2.2 Obtaining the corpus

Each participant carried out five different tasks. For the first, *Production of Vocabularies*, participants were asked to write down *"all the words they could remember"* for names of Food, Clothing, Leisure activities, Personality traits and Natural phenomena. The other four tasks asked the participants to produce four different kinds of texts – definition of words, explanation and recommendation of a movie, telling a joke –. The corpus was constituted with the productions obtained in the five tasks but in this paper we analyze only the vocabularies. Elicitation instructions were piloted at different school levels to warrant the most comprehensible wording. For each protocol participants had to fill out personal data such as name of the school, sex, school level, home language/s, and length of time they have spoken Catalan.

3.2.2.1 Procedure

Texts were gathered on paper because at the time of the study elementary school children were not familiar with word processing. Besides, teachers have recommended using paper to avoid confusion between orthographic or linguistic errors and typing errors. Except for preschool and first grade, the participants' teachers gathered the productions during regular Language classes. In the two youngest groups the children were seated in small groups (5 children) so that the

teacher helped them with technical aspects of writing. Teachers were trained in data gathering by the researchers. Mostly, the five tasks were completed in about 40 minutes.

3.2.2.2 Corpus transcription and digitalization

We developed a number of procedures to keep the original data but also to prepare it for further processing. Original productions for each task and subtask were introduced in a relational database (in MySQL) that enables us to trace information related to each element of the text by the independent variables (school, sex, school level, home language/s, and length of time they have spoken Catalan).

3.2.3 Criteria of Analysis

In order to appreciate lexical growth we define two levels of observation: *lexical forms* (or variants) and *lemmas*. *Lexical forms* are the written forms as they were produced by the participants. They may include orthographic errors (e.g., **encorós* for *rancorós* 'resentful'), creative forms of derivation (e.g., **antisimpatic* for *antipatic* 'unfriendly') multiword constructions (e.g., *anar a la piscina* 'going to the swimming-pool) or other languages (e.g., *arepa; shwarma*). All the lexical forms, except those which were illegible, were counted in order to evaluate the size of the lexicon.

A *Lemma* is the particular form chosen to represent the set of all the *lexical forms* with a similar meaning. The process of determining the *lemma* for a given word is called lemmatisation. Next, we explain the criteria for lemmatizing the lexical forms. Forms that did not coincide with the lemma were considered *variants.* The size of the lexicon was measured by the amount of *lexical forms*, the conceptual underpinning of the lexicon was evaluated by the amount and distribution of lemmas and its linguistic configuration was evaluated by an analysis of the *lexical forms*.

3.2.4 Lemmatization criteria

Like in any lexicographic study the goal of the lemmatization process was to define a canonical form that functions as a referent for a set of variants. Unlike in lexicographic studies, however, given the nature of the corpus, variants were not just inflected forms but might also be orthographic and graphic variants, paraphrases, expressions in other languages, and other kinds of variants that characterize the linguistic configuration of this corpus.

For example, the lemma *pa* 'bread' had the following variants associated:

pan Spanish for 'bread': other language
PA 'BREAD': graphic variant
Pa Bread': graphic variant
pain French for 'bread': other language
pà 'bread (with accent mark)' : orthographic variant
m'agrada el pa 'I like bread': periphrasis
vaig a buscar pa 'I am going to fetch some bread': periphrasis
el pa 'the bread': periphrasis
tinc pa 'I have bread': periphrasis

After an initial phase of problem detection, we define a number of general criteria as well as specific criteria concerning morphological characteristics, semantic attribution and lemmatization of multiword constructions (see Annex 1). Cases of doubt were solved by triangulation.

3.3 Results

Firstly, we present the results concerning the size of the lexicon and secondly we focus on its linguistic configuration as defined by the size of the units of production, the use of Catalan with either correct or deviant spelling, the use of words in other languages, hybrids or creative coinage.

3.3.1 Size of the lexicon

Participants produced a total of 242,404 *lexical forms* (tokens), 44,049 of which were different forms (types) lemmatised into 8,498 different lemmas. Table 2

shows the mean number of *lexical forms* produced in each semantic field by school level.

The mean number of *lexical forms* increases significantly for each semantic field from grade P5 to JH1 (for Food $F(10, 2{,}395)=261{,}769$, p =.000; for Clothing $F(10, 2{,}395)=141{,}499$, p =.000; for Leisure activities $F(10, 2{,}395)=102{,}890$, p =.000; for Personality traits $F(10, 2{,}395)=77{,}569$, p =.000 and for Natural phenomena $F(10, 2{,}395)=52{,}316$, p =.000). The younger group produces close to 5 forms in every field, except for Personality traits (the adjective category) for which they produce close to 3, whereas older groups produce more than 30.

Table 2. Mean number of lexical forms by school level and semantic field.*

School Level	Sample size	Mean by Semantic Field: Food	Clothing	Leisure act.	Traits of pers.	Natural phen.
	n	M SD	M SD	M SD	M SD	M SD
P5	132	4.53 (5.70)	4.40 (4.08)	4.74 (4.58)	2.67 (2.60)	5.11 (4,71)
ES1	225	19.32 (8.16)	13.72 (7.30)	13.61 (8.04)	10.18 (7.14)	10.11 (6.95)
ES2	102	19.37 (6.51)	13.28 (6.58)	11.67 (8.04)	9.01 (7.14)	9.48 (5.73)
ES3	219	24.95 (11.02)	12.95 (7.24)	12.53 (8.60)	6.76 (6.45)	6.84 (6.04)
ES4	137	32.10 (8.46)	20.12 (7.63)	19.22 (10.04)	11.35 (7.49)	15.39 (10.65)
ES5	300	33.16 (8.51)	21.66 (9.00)	23.38 (10.86)	14.19 (8.81)	14.12 (9.35)
ES6	164	32.84 (7.79)	22.26 (8.33)	26.35 (11.55)	15.49 (9.29)	13.22 (8.91)
JH1	307	37.28 (5.82)	27.89 (8.86)	30.33 (11.77)	21.02 (11.75)	19.96 (12.21)
JH2	279	33.91 (9.42)	24.93 (10.25)	25.23 (12.68)	18.31 (11.48)	17.19 (11.55)
JH3	329	35.50 (7.87)	25.73 (9.84)	24.56 (12.60)	17.70 (11.37)	15.13 (11.42)
JH4	202	36.20 (6.89)	27.14 (8.95)	25.80 (11.74)	20.86 (11.07)	20.32 (11.51)

* Note: P5= Preschool; ES1= 1st grade elementary school ; ES2=2nd grade elementary school, ES3= 3rd grade elementary school; ES4=4th grade elementary school; ES5= 5th grade elementary school ES6=6th grade elementary; school; JH1=1st year junior high school; JH2= 2nd year junior high school ;JH3=3rd year junior high school; JH4= 4th year junior high school.

There is a moment of pronounced growth at ES1 and two other not so marked but still pronounced moments of growth at ES4 and JH1. The observed spurts may be related to literacy attainments. It is during the first year of elementary school that children start gaining autonomy in reading and writing and access more written materials. Moreover, the mechanics of reading and writing are presumably overcome by the fourth year of elementary school and this might facilitate the use of specialized written texts for learning. This, in turn, might increase the availability of new lexical forms. Finally, the third spurt might be explained by the fact that the start of secondary school entails a more specialized organization of content. Secondary teachers are specialists that bring with them different forms of domain-specific discourse. After the third spurt, a decrease takes place in all five semantic fields from which only the Natural phenomena field shows recovery in the fourth year of Junior High school. This may be explained by the fact that the semantic-conceptual underpinning of the lexicon for the Natural Phenomena field is closely related to classroom content. Thus, while the growth of lemmas in the Clothing, Food and Leisure activities apparently has attained its stable state by the sixth year of elementary school, the growth of lemmas in the Natural phenomena field shows signals of further increase.

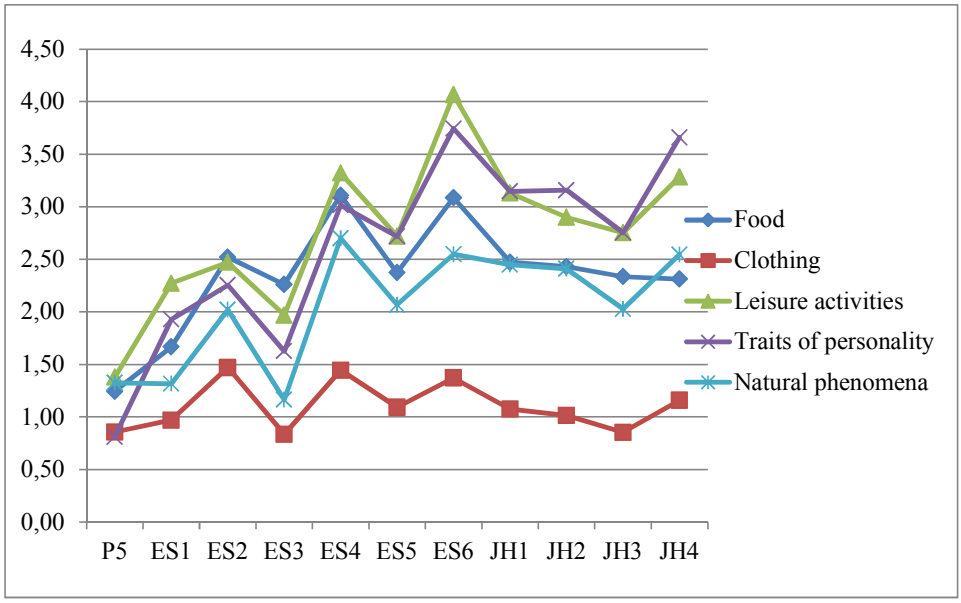
The most productive fields are Food, Leisure activities, and Clothing, in this order. The least productive field is Natural phenomena, followed by Personality traits. The syntactic category primed by the semantic fields does not explain these differences because, although two of the semantic fields that prime N as syntactic category (Food and Clothing) are the most productive, the field of Natural phenomena primes the same syntactic category and is the least productivenotwithstanding. This difference in productivity (in favor of Food and Clothing but against Natural phenomena) might be explained by the level of specific knowledge involved in the denomination of natural phenomena. While naming of Food and Clothing are part of environmental vocabulary from very early on, tokens about Temporal phenomena – except for the most common such as rain or wind – form part of more restricted or specialized contexts. As for the productivity of the Personality traits field, though it remains the second least

productive field it experiences the second most pronounced growth throughout schooling.

A univariate ANOVA on lexical forms across semantic fields showed that there is neither effect of the home language/s nor of the length of time participants have spoken Catalan on number of *lexical forms*. There is however an interaction both between home language/s and school level, and between length of time participants have spoken Catalan and school level. Although the Catalan group was the most productive and the ´other languages group' was the least productive at every school level, the distance between groups changes throughout schooling. It decreases from P5 to ES2, increases during the late years of elementary school and afterwards tends to decrease. As for the interaction with time, the participants that claimed to have spoken Catalan either always or for longer than four years were the most productive, but the difference between them and the participants that claimed to have spoken Catalan for a shorter time changes with schooling and becomes more evident within the last years of compulsory school.

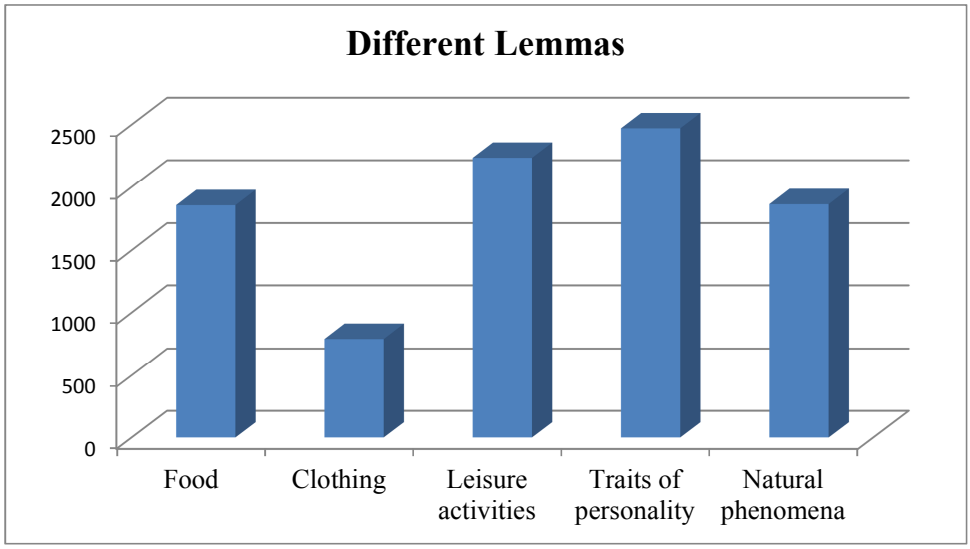
Our next analysis focuses on the lemmas, that is, the units that capture conceptual commonalities underlying sets of *lexical forms* or variants. Through this analysis we come closer to the conceptual characterization of the lexicon. We first look at the distribution of lemmas by school level and semantic fields and then we explore the relation between lemmas and variants. We have hypothesized an increase of the amount of lemmas with school level but we did not have any prediction concerning the conceptual diversity of the semantic field.

Figure 1. Mean proportion of different lemmas by school level and semantic field



As seen, except for the field of Clothing, there is an increase in the amount of different lemmas in every field throughout schooling. There are two moments of particularly pronounced growth. One occurs at the fourth grade of elementary school (ES4) and the other one at the sixth year of elementary school (ES6). After that, a decrease takes place in all five semantic fields from which only the Personality traits and Natural phenomena fields show recovery in the fourth year of Junior High school. Thus, the picture (situation) we are witnessing here looks similar to the one depicted by the growth of lexical forms. The first moment of growth (at (ES4)) coincides with the second moment of growth for lexical forms, the second one is produced a year earlier than the third discovered moment of growth for lexical forms. Except for the fields of Natural phenomena (as was the case with the growth of lexical forms) and Personality traits (in consonance (accordance) with the results showed by adults in a similar study) there is no further increase. It seems that the semantic-conceptual underpinning of the observed fields has attained its stable state.

Figure 2. Different lemmas by semantic field.



The field of Personality yields the highest amount (number) of different lemmas whereas the field of Clothing obtains the poorest one. As for the other three fields, Leisure activities follows Personality traits and is subsequently followed by Natural phenomena and Food. This entails a greater conceptual diversity for the Personality traits than for the other observed semantic fields. The order of productivity of lemmas has to do with the different conceptual basis of semantic fields. Personality traits are conceived as more differentiable entities than items of clothing. This rational is further supported by looking at the relation between lemmas and variants and verifying the extent to which semantic fields trigger different mechanisms for creating variants.

A look at the relation between lemmas and variants in each semantic field completes this picture. The Clothing field is at the higher end and obtains the highest ratio (63, 49) followed by the field of Food (38, 80) while the Personality traits field is at the lower end with a much lower ratio (14, 16). Between the two ends are the field of Natural phenomena (18, 20), close to the lower end, and the field of Leisure activities (22, 92) For those fields with higher ratio (Clothing and

Food) the diversity lays in the *lexical forms*. For Clothing, there are, for example, *samarreta de màniga curta* 'short sleeve shirts', *samarreta de màniga llarga* 'long sleeve shirts', *samarreta de màniga tres quarts* 'three-quarters sleeve shirts' and for Food there are, for example *arròs amb tomàquet* 'rice with tomate sauce', *arròs amb gambes* 'rice with shrimps.' They produce multiple variations of a relatively limited number of entities. In contrast, for those fields with lower ratios (Natural phenomena and Personality traits) the diversity lays in the conceptual underpinning. In these fields more than producing multiple variations of a limited number of entities they produce comparatively more entities. For instance, *derivats del petroli* 'oil derivatives.' This difference became evident during the lemmatization process. Confronted with creation of new lemmas for variants of Clothing or Food, we had to resort to triangulation in cases of blurred references (see lemmatization criteria). This was not necessary for Natural phenomena or Personality traits for which the reference to distinct entities was clear.

There is no influence of the home language/s on this distribution. Irrespective of the home language/s variable, Clothing presents the smaller amount of different lemmas, followed by Food, whereas Personality traits and Natural phenomena are the two richest fields in terms of lemmas. Irrespective of language group, the semantic fields have a similar conceptual basis, the linguistic influence of the home language/s appears in the relation between lexical forms (Table 3).

Table 3 Mean number and ratio of different lemmas by language spoken at home

Language	Sample size	Different Lemmas	Ratio variants/lemma
Catalan	n=513	8.41	3.35
Spanish	n=906	5.61	4.17
Catalan & Spanish	n=792	6.94	3.77
Other languages	n=185	11.59	2.89

The other language/s group contributed the highest mean of different lemmas. They probably contributed lemmas in other languages. They were followed by the Catalan group that most probably contributed Catalan lemmas. But, the other language/s group had a lower ratio of variants per lemma than the Catalan group. This might be related to the diversity of languages involved in the other language/s group. The Catalan/Spanish group, in contrast, had a lower number of lemmas, but a higher ratio of variants/lemma than the other two groups. The Spanish group had the lowest number of lemmas and the highest ratio of variant/lemmas. This means that both the Catalan and the other language/s groups have a larger semantic-conceptual underpinning than the Spanish and the Catalan/Spanish groups. However, these two groups express similar meaning with a large number of alternative *lexical forms*.

3.3.2 Linguistic configuration of the lexicon

The size of the production units along with five other dimensions were considered for characterizing the linguistic configuration of the lexicon-in-use. First, we present the results related to the size of the production units in the whole corpus by school level and semantic field. Afterwards, the results concerning the use of Catalan words and multiword constructions with (1) correct or (2) deviant spelling; words and multiword constructions in (3) other languages, including trademark calques; (4) hybrids and (5) presence of creative coinage. Fifteen percent of preschoolers' productions and about 3% of productions at other school levels were left unclassified due to incomprehensibility or illegibility and excluded from further analysis. The remaining production – a total of 239,029 lexical forms – were coded first for the size of the production unit and afterwards for the five above-mentioned dimensions.

3.3.2.1 Size of the production units

We distinguish between single words and multiword constructions. By the latter we mean every construction beyond single words, ranging from simple 'det +

Noun' constructions up to full sentences, irrespective of whether they contain spelling mistakes, are hybrids, show creative coinage or are written in other languages. The distinction between words and multiword constructions was established according to orthography. If the speaker/writer wrote without separating words that according to orthography should be written separately (e.g., **ancasa* for *en casa* 'at home'), we counted this (**ancasa*) as a multiword construction. And, conversely if the speaker/writer produced unconventional intraword spacing (**la sanya* for *lasanya* 'lasagna') we counted this (**la sanya*) as a word. That is, we considered orthographic words in the system and not in the speaker/writer's way of spelling. Table 4 presents the mean number of multiword constructions in each semantic field by school level.

Table 4. Mean number of multiword constructions by school level and semantic field.*

		Mean by Semantic Field				
School Level	Sample size	Food	Clothing	Leisure act.	Traits of pers.	Natural phen.
	n	M SD	M SD	M SD	M SD	M SD
P5	132	.80 (.34)	.34 (.93)	1.39 (2.91)	.25 (.63)	.43 (1,26)
ES1	225	1.03 (1.80)	.89 (1.47)	3.12 (3.49)	.81 (1.82)	.77 (1.51)
ES2	102	1.09 (1.71)	1.17 (1.66)	3.19 (3.32)	.15 (.45)	.54 (1.03)
ES3	219	1.50 (2.38)	.89 (1.50)	2.36 (3.35)	6.76 (6.45)	6.84 (6.04)
ES4	137	2.17 (2.15)	1.93 (2.88)	4.66 (5.37)	.66 (1.91)	.36 (.70)
ES5	300	3.08 (3.61)	2.17 (2.55)	5.91 (6.00)	.61 (1.50)	.60 (1.28)
ES6	164	2.80 (3.20)	2.55 (2.85)	7.57 (6.58)	.82 (2.06)	.76 (1.52)
JH1	307	3.53 (3.82)	3.16 (2.87)	9.70 (7.68)	.64 (1.34)	1.17 (2.12)
JH2	279	2.91 (3.10)	2.39 (2.82)	8.56 (7.99)	.75 (1.66)	1.20 (1.97)
JH3	329	2.93 (3.62)	2.11 (2.69)	6.83 (6.63)	.39 (1.41)	.96 (1.64)
JH4	202	2.47 (2.67)	1.72 (2.22)	7.50 (6.83)	.39 (1.41)	.84 (1.42)

* Note: P5= Preschool; ES1= 1st grade elementary school ; ES2=2nd grade elementary school, ES3= 3rd grade elementary school; ES4=4th grade elementary school; ES5= 5th grade elementary school ES6=6th grade elementary; school; JH1=1st year junior high school; JH2= 2nd year junior high school ;JH3=3rd year junior high school; JH4= 4th year junior high school.

For every semantic field there was a significant increase in the use of multiword constructions during elementary school and up to the first year of junior high school, which tended to decrease afterwards (for Food $F(10, 2{,}395)=24{,}741$, p =.000; for Clothing $F(10, 2{,}395)=25{,}705$, p =.000; for Leisure activities $F(10, 2{,}395)=42{,}315$, p =.000; for Personality traits $F(10, 2{,}395)=4{,}537$, p =.000 and for Natural phenomena $F(10, 2{,}395)=8{,}956$, p =.000). We should say that, in general developmental terms, there is a tendency to use fewer multiword constructions with school level.

The use of multiword constructions is both a developmental matter and a linguistic resource sensitive to the constraints of each semantic field. In line with prediction 4, the developmental tendency is to reduce the use of full expressions that somehow describe real life activities, actual items of clothing or dishes and to increase the use of isolated words – conventional lexical items of each semantic field. The structure of these full expressions, i.e., the kind of multiword constructions, however, differs by semantic field. Within Leisure activities, schemas such as [light V + N/ infinitive] (e.g., *anar a basket* 'to go to basketball'='to go play basketball') were preferred over bare infinitives or nouns. In such schema, it is the noun or the infinitive that actually holds the semantic weight while the light verb carries the meaning of actually performing the activity instead of simply stating the activity and that might be the reason for speakers/ writers' preference for these types of construction.

For Food and Clothing, although to a lesser extent, speakers/writers tended to use [N + Adj / N + PrepP] (e.g., *samarreta màniga curta* 'short sleeve t-shirt' / *arròs amb tunyina* 'rice with tuna') which are closer to actual pieces of clothing or dishes than to lexical categories. Here too it is the noun that carries the semantic weight while the adjective or prepositional phrase adds specificity to the noun. Multiword constructions appear as a strategy for producing a higher amount of

lexical forms; they function as pattern of productivity. Once a particular schema is set forth it triggers the production of new forms that follow a similar schema (e.g, *anar al cine* → *anar a la platja* →*anar al camping* 'to go to the movies' → 'to go to the beach' → 'to go camping').

The other two fields deserve considerations of another sort. In the field of Natural phenomena, production is tightly related to knowledge of specific vocabulary. Multiword construction typical of this domain reflects specialized vocabulary (e.g., *roques sedimantàries* 'sedimentary rocks'; *escalfament global* 'global warming'). In the field of Personality traits, in contrast, the kind of multiword constructions reflect the search for a desirable description of a state of affairs for which the subject lacks a suitable word (e.g., *que sempre arregla els problemes* 'that he always solves the problems' for *apanyat* 'resourceful'; *da bagadas arpota malament* 'sometimes he behaves badly' for *indisciplinat* 'undisciplinate'). The two fields differ developmentally and while multiword constructions of the kind just shown for the Personality traits field tend to diminish with schooling, multiword constructions of the kind shown for the Natural phenomena field tend to increase along with specialized knowledge.

Except for Food, we have found a significant effect of the home language on participants' production of multiword constructions in every semantic field (for Clothing $F(3, 2{,}281)=9{,}277$, p =.000; for Leisure activities $F(3, 2{,}281)=3{,}478$, p =.000; for Personality traits $F(3, 2{,}281)=5{,}191$, p =.000 and for Natural phenomena $F(3, 2{,}281)=2{,}787$, p =.000). The Catalan group produced the highest mean of multiword constructions in the field of Clothing (M=2, 26) and Leisure activities (M=6,65) whereas the Catalan/Spanish group produced the highest mean of multiword constructions in the field of Food (M=2, 57). In contrast, the other language/s group produced the lowest mean of this kind of construction in every semantic field. Tukey Post-hoc analyses show that the Catalan and the other language/s groups differ significantly from the other groups in every field except Food. Food items are treated as names of dishes across languages; a tendency perhaps influenced by all subjects' (Catalan and the other language groups) daily exposure to school menus written in Catalan.

We have also found a significant effect of the length of time that participants have spoken Catalan on their production of multiword constructions. Participants that have spoken Catalan for longer than four years produced the highest mean of multiword constructions in the field of Food whereas speakers that have always spoken Catalan were the most productive in the other fields.

3.3.3 Other linguistic dimensions

The linguistic configuration of the corpus has been characterized according to the following five dimensions.

1. Catalan words and multiword constructions: words and multiword constructions included in the Dictionary of the Catalan Studies Institute (DIEC), the Dictionary of the Enciclopèdia Catalana (DEC), the Diccionari Català-Valencia-Balear by Alcover-Moll and the TERMCAT terminological data-base, that exhibited correct spelling.

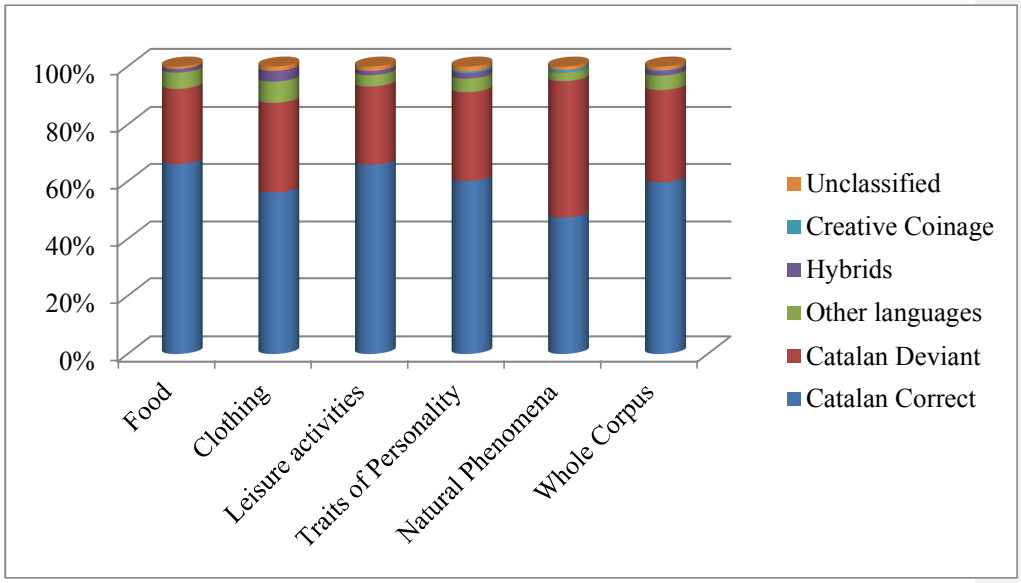
2. Catalan words and multiwords with deviant spelling: words and multiword that appear in the dictionaries but exhibited deviant spelling. This dimension includes deviant spelling due to commonly-used mispronunciations (e.g., **enciamada* for *ensaïmada* 'pastry cake', **atmelles* for *ametlles* 'almonds').

3. Other languages: words and multiword constructions not written in Catalan (e.g., *calcetín* Spanish for *mitjó* 'sock', *play station*, *schwarma*). National and international common trademarks used to name objects or actions of reference (e.g., *Cacaolat*, *Coca-Cola*) are included here. Also, calques, that is, Catalan translations for Spanish expressions that keep the exact word-to-word correspondence while not necessarily keeping the global semantic correspondence with the original expression (e.g., *arc-iris* after Spanish *arcoiris* 'rainbow' for the Catalan *arc de Sant Martí* 'rainbow')

4. Hybrids: Spanish forms, in most cases, that are partially or completely wrapped in a morphological or phonological, or both, Catalan-like form (e.g., *jamó* (from the Spanish *jamón* 'ham') for Catalan *pernil* 'ham'; *suadora* 'hoody' (a mixing between the Spanish *sudadera* 'hoody' and the Catalan *dessuadora* 'hoody')).

5. Creative coinage: forms that do not belong to the language normative repertoire although being constructed by the language morphological rules. (e.g., **insimpatic* for *antipatic* 'unfriendly' from *in+simpàtic*[ii]). Figure 3. presents the linguistic configuration of the whole corpus by semantic field according to the five dimensions. For the picture to be complete, it includes the percentage of unclassified forms.

Figure 3. Linguistic Configuration of the Corpus



In the whole corpus, 80% of the forms are Catalan either with correct (60 %) or with deviant spelling (20%). For the other 20%, the majority of forms are in other languages and only 2% are hybrids or display a creative coinage. A similar distribution is obtained in the fields of Food and Leisure activities. Three other fields differ to some extent. The field of Natural phenomena obtained a lower percentage of Catalan forms with correct spelling (48%) and the highest rate of Catalan forms with deviant spelling (47%) probably due to usage of specialized vocabulary that is learned at school but commonly used or read through (for)

short periods of time. The field of Clothing, comparatively, obtained the highest percentage of forms in other languages and hybrids. The name of clothing items is either preserved in the original language without translation to the local language –many fashion terms appear in the original language in the social enviroment as well– or converted to Catalan-like language through a hybrid. It thus appears as the most permeable semantic field. Finally, creative coinage appears almost exclusively in the field of Personality traits. It looks as though facing the need to express a quality for which they do not have the precise term, children resort to their morphological knowledge to coin the required word. So we found, for example a series of four Personality traits with the regular suffix –iu *aprens**iu*** → *comprens**iu*** → *express**iu*** → 'apprehensive' →'comprehensive' → 'expressive' folowed by a lexical creation by means of suffixation to the root *lluita* 'fight' →**lluitat**iu*** 'fighter'.

School level has a main effect on all the dimensions (for correct Catalan (F(10, 2,281)=23,760, p =.000); for deviant Catalan (F(10, 2,281)=7,431, p =.000); for other languages (F(10, 2,281)=8,019, p =.000); for hybrids (F(10, 2,281)=3,497, p =.000); and for creative coinage (F(10, 2,281)= 3,494, p =.000). Table 5 displays the breakdown of *lexical forms* per school level according to the five dimensions.

The use of Catalan forms with correct spelling increases steadily throughout school. And, with a much lower number of occurrences, there is also an increase with school level in the use of expressions in other languages and hybrids. Even lower is the mean number of creative coinage that also increases throughout elementary school and decreases afterwards. Similarly, the use of Catalan forms with deviant spelling tends to show an initial increase followed by slight ups and downs and finally decrease after the first year of junior high school. This means that the enlargement of the size of lexicon results from an increase of correct Catalan forms, forms in other language and hybrids that somehow compensate for the decrease in deviant forms of Catalan and creative coinage.

Table 5. Linguistic Configuration of the corpus by school level.

School Level	Sample size	Mean by Semantic Field: Cat. correct		Cat. deviant		Other lang.		Hybrid.		Creative Coin.	
	n	M	SD	M	SD	M	SD	M	SD	M	SD
P5	132	5.81	(13.57)	9.68	(7.78)	1.10	(1.27)	.24	(.46)	.02	(.14)
ES1	225	24.62	(15.41)	35.33	(17.00)	3.83	(3.37)	1.24	(1.38)	.11	(.36)
ES2	102	30.95	(14.56)	27.54	(10.66)	2.24	(1.95)	1.16	(1.27)	.01	(.09)
ES3	219	33.95	(18.24)	24.19	(13.04)	3.84	(4.28)	.19	(1.50)	.05	(.24)
ES4	137	57.04	(22.86)	31.40	(14.49)	6.23	(6.09)	2.06	(1.72)	.18	(.46)
ES5	300	66.34	(27.98)	31.52	(14.85)	4.95	(4.18)	2.12	(2.08)	.18	(.51)
ES6	164	69.37	(25.40)	30.02	(14.72)	7.10	(5.67)	2.25	(1.81)	.20	(.45)
JH1	307	90.37	(32.31)	34.72	(15.02)	7.21	(5.85)	2.56	(1.89)	.19	(.48)
JH2	279	82.00	(38.51)	26.90	(13.25)	6.89	(7.52)	2.35	(2.07)	.16	(.39)
JH3	329	83.62	(35.84)	25.97	(11.38)	6.12	(4.78)	2.01	(1.73)	.10	(.30)
JH4	202	97.02	(32.84)	25.15	(11.71)	5.32	(3.63)	2.11	(1.67)	.11	(.35)

Note: P5= Preschool; ES1= 1st grade elementary school ; ES2=2nd grade elementary school, ES3= 3rd grade elementary school; ES4=4th grade elementary school; ES5= 5th grade elementary school ES6=6th grade elementary; school; JH1=1st year junior high school; JH2= 2nd year junior high school ;JH3=3rd year junior high school; JH4= 4th year junior high school.

There is not a significant effect of the home language/s on any of the observed dimensions but there is an interaction between school level and home language/s on use of correct Catalan forms. At every school level, the Catalan group produces more correct occurrences than any other language group, however, with school level, particularly after the fifth year of elementary school, the differences between the Catalan and the Catalan/Spanish groups become

smaller while the difference between these two groups and the Spanish group increases. The length of time participants have spoken Catalan has a main effect only on the number of correct Catalan forms (F (4, 2,281)=13,134, p =.000).

3.4 Discussion

The construction of the mental lexicon is not based on the mere associative pairing of sounds (or written strings) and meaning. The ability to form and extend a lexicon of words or word-like signals is a main component of the human conceptual-intentional system (Tincoff and Hauser, 2006) and it is, probably, inseparable from other aspects of linguistic development. Linguist and psycholinguistic approaches may differ as to the separability between lexical and syntactic knowledge but they all agree on the centrality of this domain in language development (Anglin, 1993; Borer, 2005; Bybee, 2007; Dockrell & Messer, 2004), and linguistic literacy (Ravid & Tolchinsky, 2002). We have tracked changes in breadth and linguistic configuration of the written lexicon of Catalan students from childhood to adolescence. Four are the main findings of the study: firstly, we have found a notable increase in the vocabulary throughout compulsory school both behaviorally and conceptually but this increment is neither linear nor open-ended; secondly, there is not a single mechanism for lexical production across fields, rather each semantic field triggers particular mechanisms for creating variants and filling lexical gaps; thirdly, units of different size and complexity were elicited by a task that asked for production of isolated words, and lastly, the written lexicon is rather impermeable to multilingual interference. Following, we elaborate on each of these findings and discuss some psycholinguistic and educational implications.

Firstly, the Catalan students that participated in the study progress from producing about 5 lexical forms at age 5 to producing more than 30 as a mean at age 12. No doubt, the expansion of children's linguistic interactions together with the crucial influence of writing and written language facilitates the provision of tokens. This behavioral progress has a conceptual support shown in the concomitant increase of lexical forms and different lemmas. Children are not

gaining only in expressive means; they are abstracting lexical categories and progressing in conceptual integration.

We have detected three moments of pronounced increase, by the first year of elementary school, again at the fourth year of elementary school and then at the first year of junior high school (Table 2). After the second year of junior high school there is a sort of stagnation in this growth in four semantic fields (Food, Clothing, Leisure activities and Personality traits) where they have attained their stable state. The first moment of pronounced growth appears across fields and is most probably due to an improvement in children's writing skills. For the latter moments Food, Clothing, Leisure activities and Personality traits behave similarly and differ from Natural phenomena. The former four show the maximal number of lexical forms at the first year of junior high school and then decrease; whereas the last one has its peak of growth delayed and continues to grow until the end of high school. In urban environments of the kind that characterizes our sample, children have attained the stable state of the lexicon of Food, Clothing, Leisure activities and Personality traits at age 12 apparently from informal /social linguist interaction but the pattern of growth of the other fields may have different reasons. The pattern of Natural Phenomena is most probably related to the moment at which teaching of this subject starts at school, the growth of this field is nurtured by specialized knowledge and takes more time to attain its stable state. This development illustrates the influence of content on lexical productivity. Acquaintance with new or more specialized contents, at the beginning of junior high school, turns into higher productivity. Despite the results shown up until the fourth year of Junior High school, the particular development of the field of Personality traits might be related to its semantic content. The nuances of mood, states of mind or ways of behavior that define Personality traits might be more accessible and verbalizable to adolescents than to children. Perhaps also the syntactic category involved in the expression of Personality traits (mainly adjectives) is related to the pattern of growth of this domain. Thus, the same pattern of increase in lexical form might be explained in one case by increasing accessibility to specialized knowledge while in the other case by accessibility to emotional/personal content. The comparison highlights possible sources of lexical richness: technical, intellectual and emotional sources. These findings are partially

supported by a study we have carried out with a group of adults using the same task and instructions of the present study. In all the semantic fields, the mean number of lexical forms was very similar to -sometimes even lower than- those attained by junior high school children (students) in the present study (Tolchinsky & Marti, 2007). There was no further increase in the size of vocabulary during adulthood except for Personality traits that showed a high level of productivity, more mature people have more adjectives for qualifying mood, states of mind and ways of behavior.

We have found a similar division between semantic fields for level of productivity. Food appeared as the most productive field –the one containing more lexical forms–, followed immediately by Clothing and Leisure activities whereas Personality traits and Natural phenomena are the least productive fields. The predicted effect of syntactic category cannot possibly explain these differences for the three N-priming fields yield both the two highest and the one lowest amount of lexical forms. We think that the same reasons that explain the differential pattern of growth are at the basis of the differential productivity.

Our second finding relates to the field-dependent specificity of mechanisms for word generation. In order to generate more variants informants resort more to other languages and to multiword expression in the fields of Food, Clothing and Leisure activities than in the other two. The use of foreign language appeared more frequently for naming dishes, pieces of clothing, sports and games than for denoting Personality traits or Natural phenomena. In contrast, the field of Personality traits yielded more varying lemmas than any other field and a greater amount of creative coinage. When participants had some content to convey they found ways to express it. Although this phenomenon was not very salient in quantitative terms it illustrates clearly how speakers/writers find means of expression even when lacking the conventional term. In the field of Natural phenomena, use of multiword construction seems to relate to specialized knowledge, as for the linguistic configuration of the field it yielded the highest proportion of Catalan lexical forms, whether deviant or correct, and the lowest proportion of hybrids, foreign language and creative coinage. Its being closely tied

to school content learning would seem to provide less opportunity for deviant creation other than a high ratio of deviant spelling.

The third finding concerns the use of units of different size and complexity. In line with Bybee's (2007) suggestion, we have found units of different size – isolated words, NPs and full sentences – suddenly appearing in response to a divergent naming task. Participants used multiword constructions, even when they were explicitly asked to say single words. It shows that lexical elements of different syntactic complexity coexist and function as units of production sensitive to communicative constraints. In this case the production units were sensitive to the constraints of modality – production of written forms – and semantic specificity of each field. The patterns of use of multiword constructions have two facets: On the one hand they can be viewed as a means of producing more terms; on the other hand, they can be viewed as transitional forms toward the use of words as category labels. Both facets hint at the intimate relation between syntax and lexical items.

The possibility to resort to some sort of multiword construction notably improves the production of new forms. The use of schemas of the form [movement verb + N (place) or [light verb + N (activity) in the Leisure activities field and [(Adj+) N + (Adj/SP)] in the Food and Clothing fields allowed participants to produce extended lists even in spite of a rather limited number of lemmas (as in the case of the Food and Clothing fields, see Figure 2). These schemas are functioning as gears of productivity. We have many examples of informants resorting to applying the same schema to produce most of the terms for Leisure activities, Food or Clothing. The use of these productive schemas is deeply similar to the process of creative coinage used mainly in the field of Personality traits. In the use of productive schemas a syntactic strategy is mobilized while for creative coinage the strategy is basically morphological. However, in both strategies pivot grammatical elements (e.g., light verbs in schemas, prefixes in creative coinage) are combined with open–ended content elements (e.g., nouns in schemas, adjectives in creative coinage) so as to produce novel terms. Both strategies have a triggering effect: the production of a syntactic or morphological schema sets the pattern for the production of the next one

having a sort of "internal priming effect" (Bybee, 2007). In the dynamics of on-line lexical production syntactic and morphological schemas are functioning as mobilizing patterns in a comparable sense that schemas are functioning in early language development (Tomasello, 2003).

We contend that these multiword constructions are not only means for the generation of variants but also transitional to the abstraction of lexical categories. Words are category labels, they refer to multiple entities; when we use a word to denote an object we ignore certain properties of the object (Anglin, 1970; Brown, 1958) When writing for example, *samarreta* 'shirt' we ignore its color, size, or length of sleeves; its domain or reference is very wide. By adding features to the word *samarreta de màniga curta* 'short sleeve shirts,' *samarreta de màniga tres quarts* 'three-quarters sleeve shirts,' we reduce its domain of reference, and at the same time its generalization and conceptual integration. Multiword constructions of the form *vaig a buscar pa* 'I am going to fetch some bread' are syntactically more complex than the isolate word *pa* 'bread' they are paraphrasing but much narrower in their domain of reference and also closer to tokens of use. In this sense, we interpret the reduction in the use of multiword constructions as a gain in categorization. Our informants do not progress from producing isolated words in the early years of schooling to producing complex constructions in the later years. Rather, they produce a diversity of processing units, activated by the semantic constraints of each field, and from this diversity might emerge eventually some lexical items in the form of isolated words. Although the use of full descriptions and schemas tended to diminish with age, it was far from disappearing. In line with Bybee's (2007) ideas, the lexicon conceived in terms of lexical items or pieces of vocabulary that could be used for filling our request in all the semantic fields seems to emerge from a singling-out process. Initially, they are represented as part of larger units. Producing lexical forms in the form of single words would be more of an outcome of lexical development than a starting point of it.

Finally, we have found less effect of multilingualism on the configuration of the corpus than expected. In spite of the fact that for many informants Catalan was mainly the language of school, in spite of the strong immigration and the fact that

schools are multilingual environments and adolescence develop their idiolects, the written lexicon elicited in a school context shows a low level of permeability. There is a clear increase in the use of Catalan correct forms, and concurrently a decrease in the number of deviant forms. Exposure to written texts did cause a reduction of the orthographic variants used in naming entities at the different semantic fields. Moreover, the presence of other languages and hybrids was not very notable in the whole corpus and did not show significant differences with school level. Although due to their moderate presence in the corpus we are treating the use of other languages and the production of hybrids as similar processes, we must no forget that there are clear-cut differences between the two. Hybrids result from an interaction between languages at a morphological or phonological level; they are not part of any language. Foreign language forms, instead, are part and parcel of other languages and might be eventually incorporate into Catalan. They are still absent from our reference dictionaries but they seldom refer to technology, sports, fashion-related items or activities that are vividly, though orthographically unstable, present in media discourse and in everyday discourse in consequence. We believe that two characteristics of the elicitation procedure may explain the apparent lack of permeability of the corpus. Given that Catalan is the language of school, the fact that the task was undertaken at school with the usual teacher and in writing might have constrained the use of non-Catalan forms that are probably used out of school in normal spoken interaction. In order to fully appreciate the permeability of Catalan –or of any other language– it might be necessary to construct an equivalent corpus in the spoken modality.

The existence of this corpus, which is of public access (http://clic.ub.edu/cesca), enables psycholinguists and educationists to get an authentic and updated picture of the state of the Catalan language not just in terms of frequency of use or words and constructions in five semantic fields but also in the range of field and developmental variants. It is possible to access the multiple expressions children and adolescents use for entities, qualities and activities and also to approach the widths and developmental trends of such vocabularies, in terms of spelling but also in terms of language and accuracy throughout schooling. This kind of knowledge would be helpful in designing

strategies for explicit instruction of vocabulary, often neglected in classrooms despite its proven usefulness. Also, a better understanding of authentic lexical development by different language groups in the school years should be of considerable help in better addressing the needs of the children that do not have Catalan as their first language. Certainly, there are many features of words that were not addressed in the study and are crucial for completing our understanding of the lexicon. Future research should explore the different relations that bind words together. We have observed some examples of lexical chains (e.g., writing a quality and it's opposite) that hint at lexical links but more studies are needed to deepen our understanding of the written lexicon by exploring systematically semantic relations –e.g., synonymy, hyponymy–. Because the same informants produced, apart from vocabularies, five different types of texts that are part of the present corpus, we hope to address this aspect of words by analyzing word definitions, use of lexical forms in different contexts and the development of polysemy.

Another important point to be explored concerns the relation between spoken and written vocabulary. We are not assuming any primacy of the spoken over the written lexicon. Once written language becomes part of children's linguistic experience it is both a source of learning and an environment of use of similar caliber to spoken language. However, each modality imposes its own constraints; some words are acquired and used more in the written than in the spoken modality, and other in the spoken one. Hybrids and foreign expressions will be probably more frequently used in the spoken than in the written modality. In contrast, the written modality, along with the fact that the task favored triggering of production patterns, will favor creative coinage over the spoken modality.

Finally, although the present study provides evidence as to the difficulty of separating syntactic from lexical knowledge in the configuration of the lexicon, it is necessary to address the relation between these two aspects of linguistic knowledge across discourse genres to gain a clear picture of their interaction in the use of language in different communicative circumstances.

CHAPTER 4

Growth of text-embedded lexicon in Catalan: from childhood to adolescence

Anna Llaurado, and Liliana Tolchinsky

University of Barcelona

In press in *First Language*

Abstract: Lexical development is a key facet of later language development. To characterize the linguistic knowledge of school age children, performance in the written modality must also be considered. This study tracks the growth of written text-embedded lexicon in Catalan-speaking children and adolescents. Participants (N = 2,161), aged from 5 to 16 years produced 6 different texts: a film explanation, a film recommendation, a joke telling, and definitions of a noun, a verb and an adjective. The resultant corpus of 11,332 texts was analyzed using four distributional measures of lexical development: word length, lexical density, use of adjectives and nominalizations. Heylighen's F-measure of level of text formality was also computed. Word length, use of adjectives and nominalizations were powerful indicators of lexical development. Text type and home language had an effect on these measures. Lexical density showed no clear developmental change, and did not vary by type of text. Heylighen's F-measureI was a weaker developmental indicator. Educational implications are discussed.

Key words: lexicon; lexical growth; Catalan; later lexical development; written language; school age language; adolescents' language

4.1 Introduction

Children acquire most of the linguistic forms and constructions of their language very early on. However, they achieve proficiency and flexibility in the use of these forms in a wide range of communicative settings only after a long process of development, both in the spoken and written modalities (Berman, 2004). School-age language exhibits an extended repertoire of linguistic items, categories, and constructions as well as increasingly more efficient and explicit ways of representing the language and thinking about it (Berman, 2004, 2008; Berman & Ravid, 2008; Nippold, 1998; Ravid & Berman, 2008; Ravid & Berman, 2010). Later language development has gained increasing attention from linguists and psycholinguists (Nippold, 2002; Tolchinsky, 2004). The current paper focuses on the development of the written lexicon used by Catalan school children and adolescents – from 5 to 16 years of age– when producing different types of texts.

Lexical development is a key facet in later language development (Anglin, 1993; Nippold, 1998; Ravid, 2004). Moreover, given the strong relationship between lexical command and grammatical development in the preschool years (Bates & Goodman, 1997) and from primary school to high school (Berman, 2006), the study of lexical development is critical for shedding light on language development beyond vocabulary acquisition.

Throughout schooling, the retrieval of words becomes faster and more accurate (Dockrell & Messer, 2004); lexical-conceptual diversity grows attuned to the characteristics of different semantic fields (Tolchinsky, Martí, & Llauradó, 2010).The use of derivational morphology expands, playing an important role in lexical enrichment (Anglin, 1993; Ravid & Schiff, 2006) and in the lexicon syntax interface (Friedman & Novogrotsky, 2004; Ravid & Saban, 2008, Scott, 2004). A literate lexicon is built up through which abstract concepts and advanced figurative meanings are accessed, and complex verbal reasoning is enhanced (Peskin & Olson, 2004).

The lexical domain shows, in a unique manner, the ways in which context and cognition interact, as well as the changes in such interaction with development (Dockrell et al., 2004). Vocabulary development has been strongly related to cognitive development (as measured by intelligence tests; see Anderson &

Freebody, 1981). Research on lexical development is also relevant for a number of educational reasons. Vocabulary knowledge predicts academic success (Cunningham & Stanovich, 1997; Leong & Ho, 2008) and explains individual variance in reading comprehension (Laufer & Nation, 1999; Leong & Ho, 2008). The frequency of use of nouns and verbs plays an important role in reading speed (Holmes, Stowe, & Cupples, 1989). Children with reading difficulties usually exhibit poorer command of vocabulary than their more skilled peers. Moreover, educational interventions on lexical aspects of language lead to progress in reading comprehension (Nation, Snowling, & Clarke, 2007).

The linguistic knowledge of school age children and adolescents can hardly be characterized without taking into account their performance in the written modality (Ravid & Tolchinsky 2002; Tolchinsky 2004). Increased exposure to and practice with the written modality influence major aspects of later language development. The speaker/writer moves from command of the writing system as a notational system to mastering the written language as a discourse style. The appearance of, for instance, low frequency syntactic structures (Jisa, 2004; Scott, 2004) and longer sentences in the written texts of school children (Nippold, 2002) promotes, in turn, greater flexibility in the choice between formal versus informal language registers (Jisa, 2004, Tolchinsky, 2004). Writing becomes the necessary platform without which the remarkable changes that occur at the lexical, morphosyntactic and discursive levels, all of which are key to the successful attainment of literacy, could hardly take place (Berman & Ravid, 2008; Cameron, Hunt, & Linton, 1988). That is why we focus on the development of the lexicon in the written modality.

In a previous study (Tolchinsky et al., 2010), we examined the production of written vocabularies in five semantic fields. Our findings revealed lexical developmental growth in both size and conceptual underpinning. Nevertheless, from the language-usage perspective adopted here (Bybee, 2007; Goldberg, 2005), linguistic forms must not be considered as abstract, isolated elements but rather in relation to how people use them in different kinds of texts (Berman, 2006; Berman & Verhoeven, 2002;) and under the constraint of differing communicative circumstances, goals and audience. In other words, the genre-specific features that

characterize language use have an impact on the selection of expressive devices and grammatical constructions in non-expert text production (Berman, 2005, 2007; Berman & Nir-Sagiv, 2007; Nir-Sagiv, Bar-Ilan, & Berman, 2008; Ravid, 2006, 2010; Tolchinsky, 2004).

Throughout schooling, speakers/writers are meant to move from involved though scarcely informative and rather informal conversational productions on to detached, more accurate highly formal academic like texts (Snow & Uccelli, 2009). Control over the level of text formality, a notion that can be associated with text register, involves the ability to adjust one's use of linguistic forms in a variety of ways so as to suit the circumstances of their use. It is, therefore, an important feature of communicative competence (Jisa, 2004). Developmental changes embrace both local features/elements concerning lexical and syntactic choices and global features such as a text's level of formality (Biber, 1995, 2007; Heylighen & Dewaele, 1999;). In the current study, we focus on text-embedded lexical development as it is deployed in four different genres: narrative, argumentation, colloquial and definition. We examine lexical development by means of four distributional measures: word length, lexical density, presence of adjectives and nominalizations, considered to be suitable for the investigation of lexical development in a diversity of languages in addition to measuring the level of text formality.

Word length measured by the number of syllables per word is widely used in corpus linguistics research as a way of gauging lexical complexity (Riedemann, 1996; Wimmer & Altmann, 1996). It has been shown to be developmentally sensitive to literacy levels when measured by the number of letters per word in written language (Malvern Richards, Chipere, & Durán, 2004). Longer words reflect both the advanced use of sophisticated, precise, low frequency terms (Biber, 1995) and an increased command of structurally complex derivatives (Anglin, 1993), a finding that was corroborated for a morphologically complex language such as Catalan (Cordero, 2002).

As content words convey the bulk of semantic content and propositional information, *lexical density* - a high proportion of content words relative to the total number of words- is considered to be a good indicator of textual richness and

informativeness (Halliday, 1985, Malvern et al., 2004, Nir-Sagiv et al., 2008). Throughout schooling, the increasingly abstract and academic nature of school-based texts entails densely informative linguistic constructions with a complex hierarchical, and varied syntactic architecture supported by rich lexical density (Berman & Ravid, 2008; Ravid, 2010; Ravid & Berman, 2009). Lexical density has been shown to be an indicator of school-age children's language development (Malvern et al., 2004; Strömqvist, Nordqvist, & Wengelin, 2004) and serves to distinguish between narrative and non-narrative usage (Nir-Sagiv et al., 2008; Ravid, 2004b). The usefulness of lexical density as a measure of lexical development remains debatable, however, and other developmental studies have not observed genre effects (Johanson, 2009), or differences by age between school age children and adults (Jisa, 2010; Johanson, 2009). In another vein, Hyltenstam (1988) points out that lexical density may be not the best measure for lexical development in written productions in a second language (L2) as one can obtain high density score with a small vocabulary. Thus, due to the scarcity of previous research on lexical development in languages typologically similar to Catalan, an additional benefit of this study will be to examine the suitability of the tested measures for cross-linguistic comparisons.

We avoided the more eschewed measure of lexical diversity, in addition to lexical density, and instead we used a syntactic category to characterize later language development. For English, Russian and Hebrew, the use of *adjectives* in written texts increases markedly as children move upwards in the school system to higher grades (Bar-Illan & Berman, 2007, Caselli, Bates, Casadio, Fenson, Fenson, Sanderl, & Weir, 1995; Ravid, 2010). While children are aware of the informative value of adjectives in relation to nouns from early on, a full array of adjectival categories is far from present in 6-year-olds (Blodgett & Cooper, 1987). An improved command of school-based, nominally denser texts would thus entail a rich adjectival texture grounded in complex nominal syntactic structures. Hence, the size and makeup of the adjective category can be taken to constitute a yardstick for language 'richness' (Ravid, 2010). Such an increase has been suggested to coincide with the consolidation of an 'advanced,' high-register, literate lexicon and its cognitive correlates (Dockrell et al., 2004) and has been shown to be indicative of language development in school age populations (Ravid,

Levine, & Avivi-Ben Zvi, 2003).

Finally, specific, abstract concepts are progressively encoded by *nominalizations,* which can assist the writer in maintaining an impersonal tone and a detached stance and help to strengthen textual cohesion. Nominalizations require prior integrated knowledge in the domains of morphology, syntax and discourse, hence implying sophisticated morphologically complex lexical uses and, to some degree, constitute a feature of accomplished academic writing (Baratta, 2010; Ravid, 1998;Tyler & Nagy, 1989).

Due to increased experience with the written language and improved literacy levels, we predict school grade to be associated with all five measures. We expect that participants will gain in the ability to adjust to text-genre specific features and predict an effect of text type on the tested measures. In particular, we expect definition and explanation –both frequently practiced in academic settings— to be lexically denser, and to show a higher proportion of nominalizations, therefore yielding a higher word length average. In contrast, we expect more oral-like types of text such as joke telling and recommendation of a film to show less density, and to contain few morphologically complex words. The effect of the participants' linguistic background on text-embedded lexical development will also be examined. To elaborate, in Catalonia, Catalan and Spanish are both official languages. Since all children use Catalan at school and Spanish is massively present both in the media and in social settings, one is unlikely to find a monolingual child or adolescent in either of the two languages. Rather, some degree of bilingualism, though unbalanced (Schlyter, 1993), is the norm. 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. The linguistic background of Catalan school children is highly heterogeneous and the designation L1, L2 and so on, does not correspond exactly to the ecological situations in which languages are acquired by children and adolescents nowadays. Our sample includes multilingual children some of whose home language coincides with the school language and others for whom this is not the case.

Indeed, both multilingual and monolingual learners face the same problem of mapping form and function to produce meaningful utterances based upon their language experiences (Ellis 2002; Lieven & Tomasello 2008). Certainly, phenomena like code switching and code mixing are restricted to multilingual speakers (Myers-Scotton, 1993, Poplack, 1987) but they are considered to be signs of a particular kind of linguistic competence rather than indications of a lack of proficiency **(**Gollan & Ferrera, 2009; Zentella, 1997). Multilingual development has not been proven to be detrimental to language development (Bialystok & Feng, 2010). However, when new languages are learned in formal contexts the amount of exposure and multiple motivational and individual factors may lead to important differences in the performance of multilingual learners (Gersten & Baker, 2000). It is reasonable to assume that, the lexical uses of children and adolescents with differing home languages learning Catalan mostly at school and living in a multilingual environment will differ from their peers who, in spite of living in the same multilingual environment, have a home language that coincides with the school language. We therefore expect texts produced by participants who speak Catalan at home to be semantically richer, morphologically more complex and better adjusted to patterns of text formality than texts produced by children who do not speak Catalan at home, and whose use of Catalan is mostly restricted to school-based interactions.

4.2 Methodology

4.2.1 Participants

A cohort of 2,161 children and adolescents took part in this study, aged from 5 to 16 years of age and distributed by school grade. At the time of the study they were attending 32 schools in the Catalan education system.

A sociolinguistic questionnaire was used to gather information on the participants' sex, age, school grade, home language or languages, as well as how long they had been familiar with the Catalan language. Five different groups were established according to the participants' self-declared home language: Catalan only (C); Both Catalan and Spanish (CS); Spanish only (S); any language except

Catalan or Spanish but familiarity with Catalan for more than 4 years (O>4), and any language except Catalan or Spanish at home and familiarity with Catalan for less than 4 years (O<4). It must be noted that the group of participants who speak neither Catalan nor Spanish at home is highly heterogeneous and includes speakers of Romance, Germanic, Slavic, Semitic, Austronesian and Sinotibetan languages. Table 1 shows the distribution of the participants by school grade and home language or languages.

Table 1. Distribution of participants by school grade and home language.

Home language	School grade										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Total
No response	81	12	28	4	20	6	7	11	12	12	193
Any language - Catalan >4 years	23	9	32	12	23	10	26	25	23	11	194
Any language – Catalan <4 years	21	10	26	18	55	19	38	29	35	23	274
Spanish only	21	12	42	17	51	40	80	84	130	68	545
Both Catalan & Spanish	18	21	34	32	99	43	99	62	75	57	540
Catalan only	52	36	48	43	46	30	43	51	38	28	415
Total	216	100	210	126	294	148	293	262	313	199	2161

Our sample reflects the current linguistic situation in Catalonia. Thus, 32% of the participants identified Catalan as their sole home language, 20% identified Spanish as their home language and a further 28% identified both Catalan and Spanish as their home languages. The home language of 16% of the participants was a language other than Catalan or Spanish. The remaining 4% of the sample did not provide a response regarding their home language and were not included in the analysis.

4.2.2 Tasks

Participants were instructed to produce six different types of texts: a film explanation, representing the narrative genre, following the instruction "*think of a film or TV series that you like and tell us about it*", a film recommendation, accounting for the argumentative genre, with the instruction "*think of a film or TV series that you like and recommend it to a friend*", the telling of a joke, accounting for the colloquial genre, following the instruction "*think of a joke or funny story that you know and tell it*" and the definitions of three words (a noun, a verb, and an adjective) "*give a definition of*".

4.2.3 Text analysis

Tokens, types and lemmas (see below for definitions) were counted in order to explore overall text-embedded lexical growth, lexical diversity and conceptual underpinning in the texts. However, all analyses were performed at the token level (rather than type) since we were interested in actual usage patterns (Bybee, 2007; Goldberg, 1995; Ravid, 2010).

4.2.3.1 Criteria for lexical characterization

Four distributional measures were applied to different dimensions of vocabulary: (1) word length, (2) lexical density, (3) use of nominalizations and (4) use of adjectives.

(1) Word length was measured by number of letters per word.

(2) Lexical density was measured as the proportion of words included under the grammatical categories Noun, Verb and Adjective in relation to the total number of words in the text.

(3) Nominalization refers to the process by which a noun is obtained from a verb (1) or an adjective (2). It can also be the outcome of a process of zero derivation by which a stem can be realized as a noun without involving any affixation (3)[1].

1. [satisfe(r)]v 'to satisfy'→ [satisf + acció] 'satisfaction'
2. [brut] 'dirty'→ [brut + ícia] 'dirtiness'.
3. [cost] stem [cost]n 'cost'.

The measurement used was the proportion of nominalizations relative to the total number of words in the text.

(4) Adjectives normally follow the noun in Catalan and are marked by gender and number, in agreement with the noun. They can be grouped into two classes according to their morphological complexity. The first class contains adjectives expressed by a root plus inflected gender and number (if necessary) (1). The second class is formed by a root plus one (or more than one) suffixed or/and prefixed morphemes plus inflected gender and number (if necessary). Participles are included in this group as, in Catalan, the characteristics of the participle morpheme are more derivational-like than inflectional-like (Mascaro, 1986) (2).

(1) [calb (root) + e (fem. gender) + s (plural number)]
(2) [nation (root) + al (suffixed morpheme) + (fem. gender) + s (plural number)]

The measurement used was the proportion of adjectives relative to the total number of words in the text.

Level of text formality

An index of formality was computed using Heylighen's F-score (Heylighen et al., 1999):

F = (noun frequency + adjective frequency + preposition frequency + articles frequency– pronoun frequency – verb frequency – adverb frequency – interjection frequency + 100)/2.

Heylighen bases his measure on the frequencies of different word classes in a corpus. In his account, a high frequency of nouns, adjectives, prepositions and articles characterizes detached, accurate, highly formal texts whereas a high frequency of pronouns, verbs, adverbs and interjections are more like involved informal texts. Therefore, the higher the F value, the higher the level of text formality.

4.2.4 Procedure

Children performed the task in class groups. Texts were written by hand – in order to avoid possible graphic, spelling and textual deviations due to a lack of text processing skills. Both completion of the sociolinguistic questionnaire and text writing took place in the participants' regular classrooms at the request of their usual Catalan language teachers. The teachers received training in text elicitation. The task did not last more than one class session. Sociolinguistic questionnaires were completed by the participants before they engaged in the text-writing task. The task was carried out as part of their everyday school activities. A total of 11, 332 texts was generated.

4.2.5 Text preparation

Three levels of linguistic units were established: lexical forms or tokens, that is the form as produced by participants; types, subsuming all the occurrences of a particular token; and lemmas, the canonical form of the word, that is, the form that represents all the word inflections (e.g., tense, number, gender), graphical and orthographical variants of the word. Graphical and orthographical variants were included given that the corpus was made up of non-normative texts.

The original version was written by hand and three new versions were produced from the original one: 1) an original version which reproduced the texts as they were written by participants without correcting spelling mistakes, 2) a normalized version which standardized orthography for conventional separation

of words in written Catalan, and 3) a labeled version in which words had been lemmatized and morphologically labeled (for an extended characterization of the corpus see Llaurado, Marti & Tolchinsky, 2011)

4.3 Results

This section consists of four parts. Firstly, we provide a general description of the corpus in quantitative terms regarding the number of texts by type of text. Secondly, we present quantitative results about the linguistic units in the corpus – tokens, types, lemmas and syntactic categories. Thirdly, we approach the lexical configuration of the texts by means of the measures that were computed on these linguistic units (word length, lexical density, use of nominalizations and adjectives). Fourthly, we show the results of applying an index of text formality.

A series of two ways ANOVAs school grade (10) by home language (5) with repeated measures on type of text (6) were performed on the distribution of linguistic units (tokens, types and lemmas), the measures for characterizing the lexical composition of the texts and the level of text formality in order to determine the effect of school grade and home language as well as possible interactions on these dependent variables. The eta squared value (η2) is used to report the effect size of both the main effect and the interactions. The size effects of relevant pairwise comparisons are reported using Cohen's *d*. An alpha level of .05 was used for all statistical tests. When the assumption of sphericity was found to be violated, degrees of freedom were corrected using Greenhouse_Geisser estimates.

4.3.1 General description of the corpus

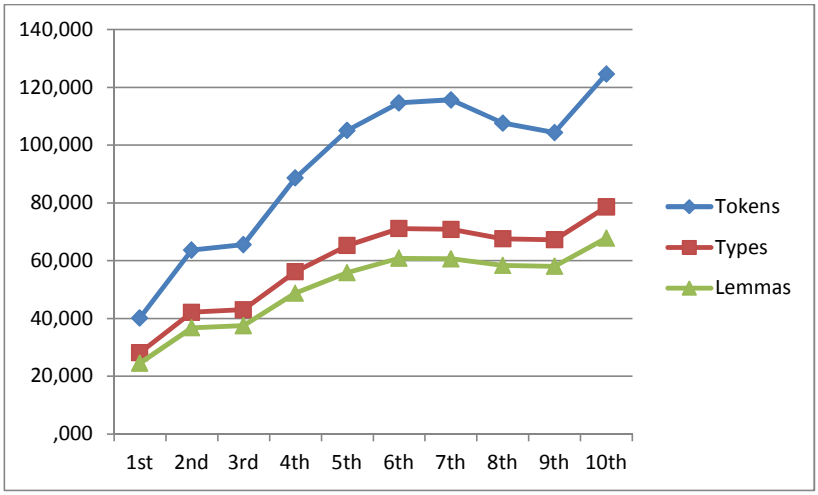
There were 1,830 definitions of a noun; 1,820 definitions of a verb; 1,829 definitions of an adjective; 2,037 film explanations, 1,955 film recommendations and 1,861 joke telling texts. Of the 2,161 participants, only 1,385 participants produced all six required texts.

4.3.1.1 Linguistic units

The 11,332 texts yielded 207,028 tokens, of which 131,263 were types and were lemmatized into 113,160 different lemmas.

The growth of tokens, types and lemmas followed similar developmental patterns. School grade had a significant impact on the growth of each linguistic unit F (9, 2161) = 28.129, p < .001, η2 = .13, F (9, 2161) = 39.057, p < .001, $\eta 2$ = .14 , and F (9, 2161) = 42.966, p < .001, η2 = .18, for tokens, types and lemmas, respectively. Bonferroni post-hoc comparisons revealed significant developmental gains between 1st and 3rd grade (d = 0.72 ; 0.74; 0.78 for tokens, types and lemmas, respectively), and 3rd and 5th (d = 0.85; 0.91; 0.93 for tokens, types and lemmas, respectively). An additional significant difference between 7th and 9th grade reflected a decrease in the number of units (d = 0.30; 0.11; 0.11 for tokens, types and lemmas, respectively). Overall growth of tokens, types and lemmas between 6th and 10th grade proved to be moderately significant for both types (p = .039) and lemmas (p =.021) but not significant for tokens. Thus, the mean number of tokens, types and lemmas grows consistently up to 6th grade (11;6 mean group age) and then stagnates throughout secondary education, showing recovery only by 10th grade. By the end of compulsory schooling, growth of the conceptual underpinning of texts continues in the absence of a significant increase in text length.

Figure 1. Plotted means of tokens, types and lemmas by school grade



It must be noted in Figure 1 that, while the total number of tokens equals the sum of tokens in each text, the total number of types and lemmas does not because repeated units are excluded from the count.

Type of text also had a significant impact on the increase in linguistic units $F(5, 2161) = 462.044$, $p < .001$, $\eta 2 = .18$, $F(5, 2161) = 549.027$, $p < .001$, $\eta 2 = .21$ and $F(5, 2161) = 549.514$, $p < .001$, $\eta 2 = .21$ for tokens, types and lemmas, respectively. For the three linguistic units, pairwise comparisons revealed significant contrasts between each type of definition and explanation of film and joke telling ($d > 0.82$), definitions and recommendation of a film ($d > 0.32$), and recommendations and both explanations and joke telling ($d > 0.58$).

Figure 2. Mean number of tokens by school grade and text type

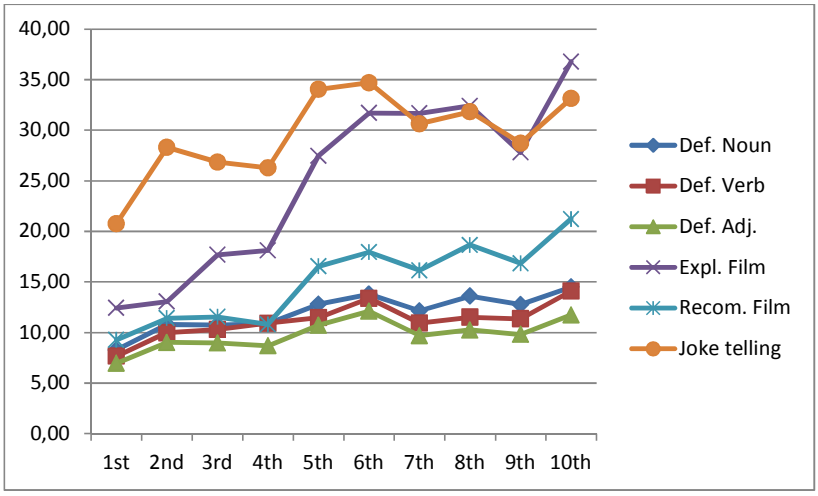


Figure 2 shows that all types of text experience two marked increases, one between 4th and 6th grades, and a second one in 10th grade. Joke telling has an additional surge between 1st and 2nd grade and is the wordiest text up to 7th grade, then it is overcome by explanation of a film, the type of text that experiences the most pronounced increase in the number of tokens. Definitions experience rather mild overall growth throughout schooling. Their most pronounced surge happens between 4th and 6th grades, recommendations behave very much like definitions up to 4th grade and then experience a mild increase. This pattern of growth yielded

a significant interaction between the type of text and school grade $F(45, 2161) = 6.955$, $p < .001$, $\eta 2 = .04$, $F(45, 2161) = 8.603$, $p < .001$, $\eta 2 = .04$, $F(45, 2161) = 5.529$, $p < .001$, $\eta 2 = .07$ for tokens, types and lemmas, respectively.

Finally, home language had a significant impact on the three linguistic units $F(4, 2161) = 13.870$, $p < .001$, $\eta 2 = .03$, $F(4, 2161) = 15.299$, $p < .001$, $\eta 2 = .04$, $F(4, 2161) = 15.939$, $p < .001$, $\eta 2 = .04$ for tokens, types and lemmas, respectively. Bonferroni post-hoc analyses revealed significant contrasts between speakers of other languages with less than four years of experience of Catalan (O<4) participants ($M = 70.20$, $SD = 46.50$; $M = 45.50$, $SD = 24.47$; $M = 39.46$, $SD = 19.87$, for tokens, types and lemmas respectively)) and both their Catalan speaking (C) peers ($M = 110.19$, $SD = 67.14$; $M = 67.99$, $SD = 34.03$; $M = 58.11$, $SD = 27.71$ for tokens, types and lemmas, respectively) ($d = 0.70$) and Catalan and Spanish (CS) peers ($M = 101.78$, $SD = 59.05$; $M = 64.28$, $SD = 30.28$; $M = 55.45$, $SD = 24.21$ for tokens, types and lemmas, respectively) ($d = 0.60$), and between the Spanish speaking (S) participants ($M = 89.12$ $SD = 55.96$; $M = 57.35$ $SD = 29.61$; $M = 49.62$ $SD = 24.08$), and both the Catalan speaking (C) ($d = 0.34$) and the Catalan and Spanish (CS) ($d = 0.22$) groups. The Catalan speaking (C) and the Catalan and Spanish (CS) participants consistently produced longer, more diverse texts than all the other groups. At the bottom end, the speakers of other languages with less than four years of Catalan (O<4) consistently produced the shortest least diverse texts.

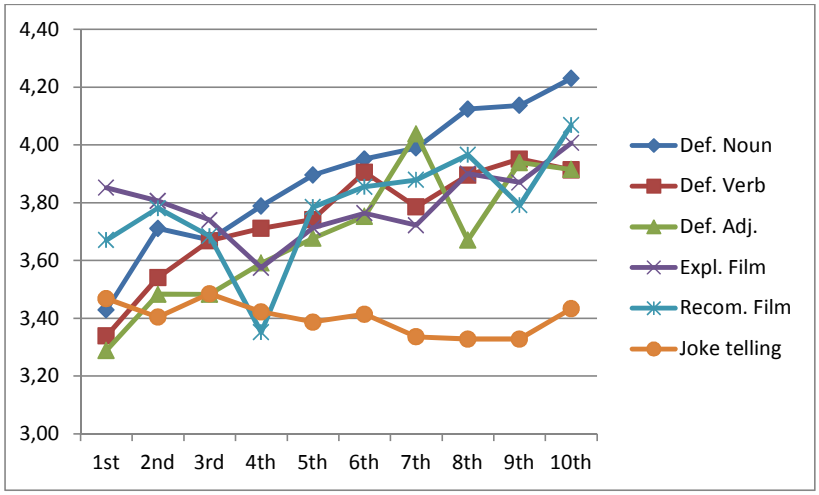
4.3.2 Lexical characterization of texts

4.3.2.1 Word length.

Word length increased significantly with school grade $F(9, 2161) = 32.082$, $p < .001$, $\eta 2 = .15$. Bonferroni post-hoc analyses revealed significant developmental gains between 1st ($M = 4.22$, $SD = 0.46$) and 3rd grade ($M = 4.35$, $SD = 0.34$) ($d = 0.34$) 3rd and 6th grades ($M = 4.51$, SD =0.31) ($d = 0.71$), and 6th and 10th grades ($M = 4.73$, $SD = 0.34$) ($d = 0.68$,). Furthermore, we found significant differences by type of text $F(5, 2161) = 100.632$, $p<.001$, $\eta 2 = .06$. Pairwise comparisons revealed significant contrasts between joke telling ($M = 3.34$, $SD = 0.25$) and all other types of text ($d > 0.54$), and also between definition of a noun ($M = 3.92$, $SD = 0.29$) and definition of an adjective ($M = 3.67$. $SD = 0.27$) ($d = 0.92$).

An interaction was found between school grade and type of text $F(45, 2161) = 6.244$, $p < .001$, $\eta 2 = .06$.

Figure 3. Mean word length by school grade and text type.



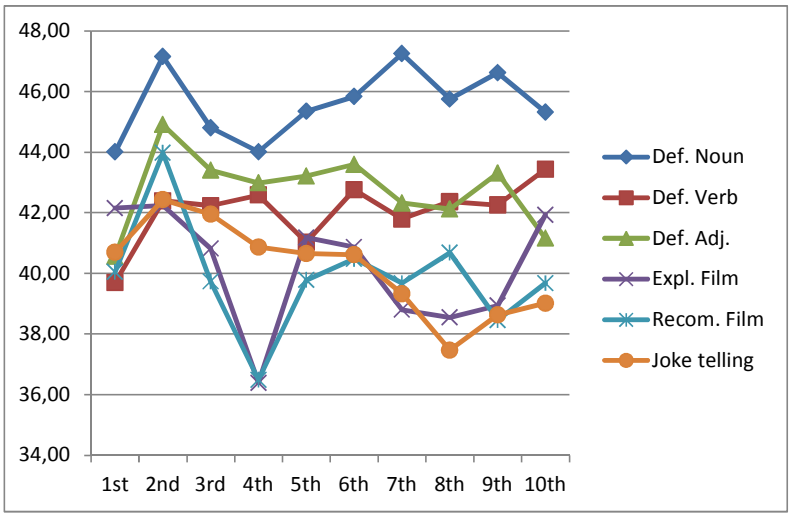
Thus, from early school grades onwards, word length in definitions involves the use of long, low frequency, sophisticated (morphologically) complex words such as nominalized verbs, e. g., *llegir: ensenyament de comprensió* 'to read: teaching of comprehension'. In contrast, word length decreases between 1st and 4th grades in explanations and recommendations of films possibly because they start off as almost a mere series of nouns with a lack of function words (prepositions, referential pronouns, etc.) which as a rule are shorter than content words, e. g., *Los dibuxos (de) spider-man (i de) super- nan (i de) piratas del caibe* 'the cartoons (of) spider-man (and of) super-man (and of) pirates of the caribean' (words within brackets missing in the original). The omission of a written representation of such particles decreases sharply with age as children become better acquainted with uses of grammar, cohesive devices and other mechanisms necessary for text construction.

We also found a significant impact of home language on word length $F(4, 2161) = 3.976$, $p < .001$, $\eta 2 = .02$. C participants scored highest for word length ($M = 3.77$, $SD = 0.14$). They were followed by CS speakers ($M = 3.73$, $SD = 0.13$). Next, came the S speakers ($M = 3.71$, $SD = 0.12$). Finally the O>4 participants yielded the same mean word length as their O<4 peers ($M = 3.68$, $SD = 0.27, 0.42$). However, only the contrast between C and O<4 participants was significant ($d = 0.32$).

4.3.2.2 Lexical density

Lexical density was significantly impacted by school grade $F(9, 2161) = 4.303$, $p < .001$, $\eta 2 = .05$. Bonferroni post-hoc analyses showed significant differences between 1st and 3rd grade ($d = 0.6$), and 3rd and 5th grade ($d = 0.6$). Type of text also had a significant impact on lexical density, $F(5, 2161) = 67.442$, $p < .001$, $\eta 2 = .04$. Pairwise comparisons revealed significant contrasts between definitions of a noun and explanations, recommendations and joke telling ($d > 0.80$), between definitions of adjectives and explanations and recommendations and joke telling ($d > 0.54$), and between definitions of verbs and explanations and recommendations and joke telling ($d > 0.40$). The interaction between school grade and type of text was significant, $F(45, 2161) = 2.619$, $p < .001$, $\eta 2 = .01$. Definitions were lexically denser than the other three types of text throughout schooling. Surprisingly, the findings suggest that young children produce denser texts than their older peers. In line with the findings for word length, there is evidence of a lack of full command of grammar by younger children, e.g., *Quan (en**)** goku va matar al bubu i (se'n**)** va anar a un altre joc* 'When (the) Goku killed (to the) Bubu and (himselh there) went to another place' (words within brackets missing in the original). This fact may partly explain this otherwise unexpected result. Another possible interpretation is that this measure is not particularly sensitive text-embedded lexical development in the present context.

Figure 4. Mean lexical density by school grade and text type.



Lexical density was significantly affected by home language, $F(4, 2161) = 3.138$, $p = .008$, $\eta 2 = .02$. Unexpectedly, O<4 Participants produced the densest texts ($M = 42{,}92$, $SD = 0.31$). They were followed by C and CS participants, both groups producing equally dense texts ($M = 42{,}24$, $SD = 0.20$. Next, came O>4 speakers ($M = 41.89$, $SD = 0.26$). Finally, S participants yielded the least dense texts ($M = 41.82$, $SD = 0.22$). However, these contrasts were not significant.

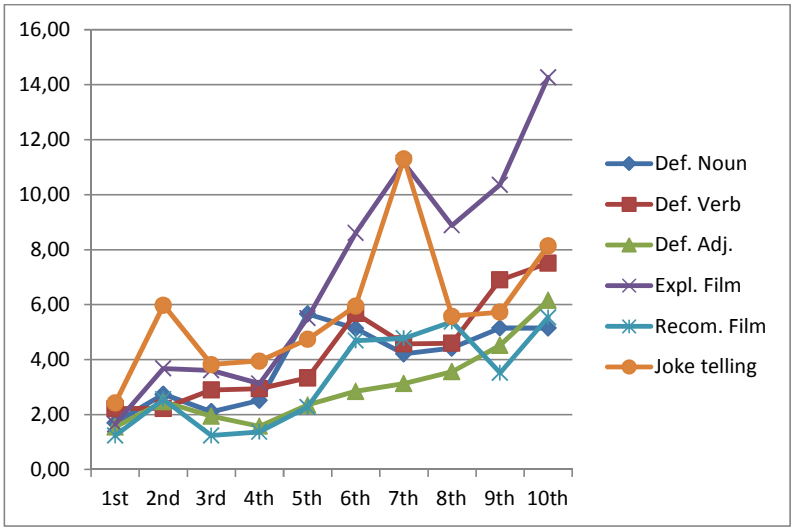
4.3.2.3 Nominalizations

Participants produced a significantly higher proportion of nominalizations as they progressed through school grades $F(9, 2161) = 33.933$, $p < .001$, $\eta 2 = .10$. Post-hoc analyses revealed significant differences between 1st ($M = 1.81$, $SD = 1.01$) and 5th grade ($M = 3.76$, $SD = 2.62$) ($d = 1.04$), between 5th and 7th grade ($M = 6.06$, $SD = 3.31$) ($d = 0.78$), and between 7th, and 10th grade ($M = 7.55$, $SD = 3.52$) ($d = 0.44$).

The use of nominalizations was also significantly affected by type of text $F(5, 2161) = 57.397$, $p < .001$, $\eta 2 = .04$, and pairwise comparisons showed significant contrasts between definitions of adjectives and all other types of text ($d > 0.30$) , between explanations and all the other types of text ($d > 0.47$) , and

between joke telling and all other types of text (d > 0.46). Explanation of a film showed the highest mean number of nominalizations (M =7.00, SD=3.00) and definition of an adjective showed the lowest mean number (M =2.78, SD =1.50).

Figure 5. Mean proportion of nominalizations by school grade and text type.



Type of text had an additional effect: it was determinant in the distribution of complex versus non-complex nominalizations. Thus, definitions provided a fertile context for rich complex nominalizations, e.g., *pau* 'peace': *amistat, confiança* 'friendship, trustfulness'; in contrast, recommendations of a film/TV series promoted the use of a rather basic, colloquial vocabulary, and also allowed for the presence of imported anglicisms, e.g., (...*) L'advertiria també que no es deixi "menjar el coco" perquè en aquest món hi ha molt de "marketing*" '(...) I would also warn him not to let himself be fooled because this field is full of marketing'.

Table 2. Ratio of complex/non-complex nominalizations by type of text.

	Type of Text			
	Definitions	Expl. Film	Recom. Film	Joke Telling
Ratio Complex/Non-complex nominalizations	2.42	1.58	1.22	1.03

Taken together, these findings suggest that definition tasks foster the greatest use of morphologically complex vocabulary. It appears that the less school-based the type of text is the less likely participants are to use sophisticated vocabulary.

A significant interaction between school grade and type of text was also found $F(45, 2161) = 2.581$, $p < .001$, $\eta 2 = .02$. The use of nominalizations increases in every type of text. Joke telling starts off as the type of text containing the highest number of nominalizations but, in 5th grade, it is overtaken by film explanations, which from then on surpasses all other types of texts. Film recommendations and the definition of verbs obtain the poorest results up to 4th grade, but by 10th grade it is the definition of nouns that yields the lowest mean use of nominalizations.

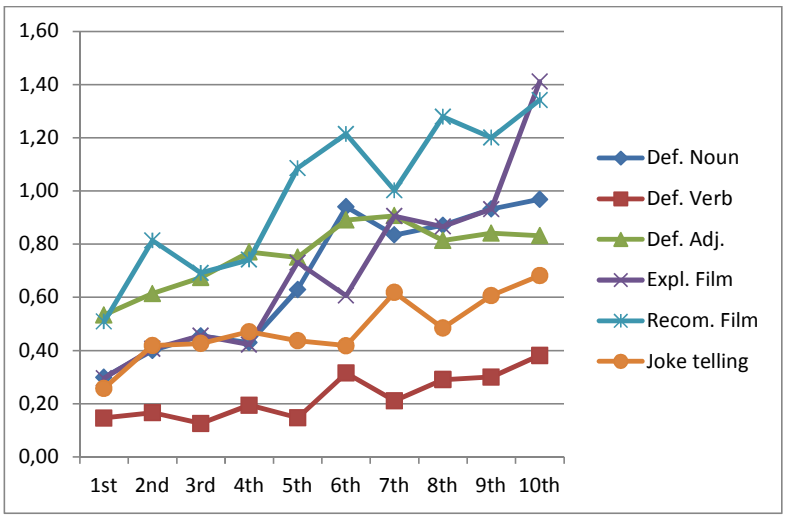
We found a marginally significant effect of home language on the use of nominalizations $F(4, 2161) = 8.678$, $p = .048$, $\eta 2 = .06$. Post-hoc Bonferoni analyses revealed significant contrasts between the O<4 (M =3.98, SD =3.10) and both C (M =5.57, SD =1.90) (d = 0.63) and SC (M =4.65, SD =2.00) (d = 0.26)). We found an additional significant contrast between the S group (M =3.80, SD =2.10) and their C peers (d = 0.88).

4.3.2.4 Adjectives

The relative proportion of syntactic categories varied within each school grade. Only adjectives (and conjunctions) showed a steady increase.Adjectives

grew from 3% of the total number of tokens produced by participants in 1st grade to 9% in 10th grade.

Figure 6. Mean proportion of adjectives by school grade and type of text.



We found a significant effect of school grade, $F(9, 2161) = 28.899$, $p < .001$, $\eta 2 = .11$. The use of adjectives increases significantly throughout schooling. Bonferroni post-hoc analyses revealed significant differences between 1st (M =3.37, SD =3.2) and 3rd grade (M = 4.71, SD = 2.9) (d = 0.44), 3rd and 5th grade (M = 6.12, SD =2.8) (d = 0.49), 5th and 7th grade (M =7.76, SD =3.4) (d = 0.53), and 7th and 10th grade (M =9.32, SD =3.6) (d = 0.45). We observed a significant effect of type of text, $F(5, 2161) = 126.863$, $p < .001$, $\eta 2 = .11$. Pairwise comparisons revealed significant contrasts between definitions of a verb and all other types of text (d > 0.82), between joke telling and all other types of text (d >0.54) and between recommendations and all the other types of text. (d > 0.35) . In sum, four types of text: recommendation of a film (M = 9.97, SD = 2.7), definition of adjectives (M = 7.52, SD = 2.2), film explanations (M = 7.05, SD = 3.0), and definition of nouns (M = 6.66, SD = 2.4), favour the use of adjectives more than joke telling (M = 4.92, SD = 2.4), and the definition of verbs (M = 2.30, SD = 1.3).

Type of text has an additional effect: it was related to the distribution of complex versus non-complex adjectives. For instance, definitions provided participants with the opportunity to produce sophisticated complex adjectives, e.g.,

fastigós 'nasty': *esser viu despreciable* 'despicable living being'. In contrast, joke telling tended to foster the use of more basic kind of adjectives, e.g., *Un acudit **verd** i molt **ràpid**: una granota en moto* 'a **quick green** joke: a frog on a bike'.

Table 3. Ratio of complex/non-complex adjectives by type of text.

	Type of Text			
	Definitions	Expl. Film	Recom. Film	Joke Telling
Ratio				
Complex/Non-complex nominalizations	0.72	0.58	0.50	0.24

Thus, the two types of texts most frequently practiced at school - definitions (here taken together) and explanations - appear to provide the context for the use of morphologically complex adjectives. In contrast, less school-based, more oral like texts tend to contain non-complex, i.e., more basic kinds of adjectives.

A significant interaction was observed between school grade and type of text, $F(45, 2161) = 3.493$; $p < .001$, $\eta 2 = .06$. Specifically, three types of text: definition of a noun, film recommendation and film explanation undergo a clear boost in the mean number of adjectives in 4th grade (film explanation also shows a second burst in 9th grade). Definition of a verb and joke telling fall behind the other types of text but also evidence an overall increase in adjective use. In contrast, data on the definition of adjectives indicates a more consistent pattern of use throughout the age range studied here.

Home language had a significant effect on the use of adjectives, $F(4, 2161) = 8.519$, $p < .001$, $\eta 2 = .02$. Post-hoc Bonferoni analyses revealed a significant contrast, between S speakers ($M = 5.8$, $SD = 2.0$) and both C ($M = 7.5$, $SD = 2.0$) and

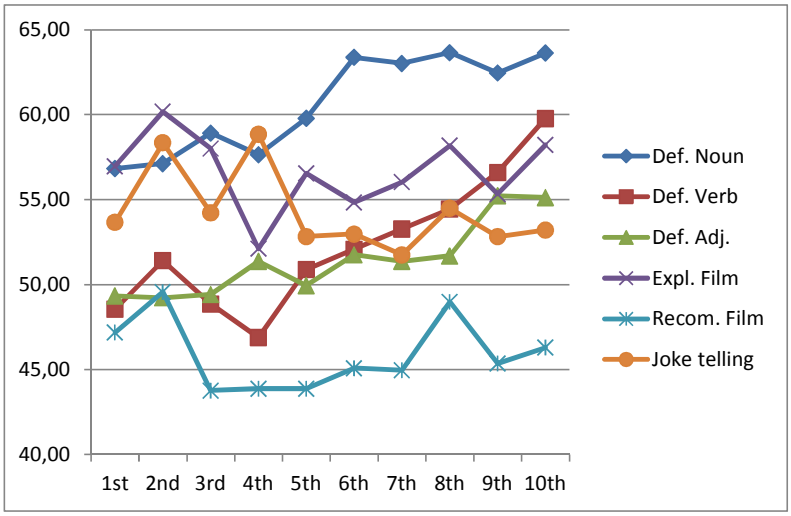
CS (*M* = 6.6, *SD* = 2.1) (d > 0.40). O<4 (*M* = 0.56, *SD* = 0.03) speakers produced less adjectives than any of the other groups, though this difference was not significant.

In sum, word length, the use of nominalizations and the use of adjectives showed a developmental increase in all types of texts. In contrast, lexical density showed neither a clear developmental pattern nor consistent distributional particularities by type of text. In other words, measures determined by the proportion of content words over the total number of words (lexical density) do not seem to account for developmental changes nor genre differentiation. In contrast, measures specifically related to the characteristics of the lexical pieces (length, the use of nominalizations and syntactic category) were better suited to characterize genre-specific developmental patterns.

4.3.3 Level of text formality

Results show significant differences by school grade, $F(9, 2161)= 6.992$, $p < .001$, $\eta 2 = .03$ and type of text $F(5, 2161) = 163.162$, $p < .001$, $\eta 2 = .07$. Post-hoc analysis revealed significant differences between recommendation of a film and all the other types of text (d >.40), while definitions of a noun contrasted with every other text except explanations (d > .60).

Figure 7. Mean *F* –level of text formality, by school grade and type of text



We also found a significant interaction between school grade and type of text $F(45, 2161) = 3.594$, $p < .001$, $\eta 2 = .01$. Thus, we observed a developmental increase in text formality for definitions (of nouns, verbs, and adjectives). However, the formality of film explanations and recommendations and joke telling appear to decrease with age. This decrease was, however, far from linear and was characterized by marked ups and downs. From 5th grade on, the definition of nouns and film explanation showed higher levels of formality than other types of text. The recommendation of a film consistently yielded lower levels of text formality. Home language had no significant effect on the level of formality, $F(4, 2161) = 1.988$, $p = .077$.

4.4 Discussion

We have tracked changes in the breadth and token/characterization of the written lexicon of Catalan students from different home language background, from childhood to adolescence. The study offers four main findings. Firstly, text-embedded lexicon increases notably throughout compulsory schooling, both behaviorally and conceptually. Secondly, text-embedded lexical development can be assessed by a number of measures of lexical usage. Word length, lexical density, and the use of nominalizations and adjectives have been used as criteria to discriminate text-embedded lexical usage both developmentally and by genre (Johanson, 2009, Nir-Sagiv 2005, Ravid 2004a, 2004b, 2010). Our results support the suitability of word length, the use of nominalizations and the use of adjectives for assessing developmental lexical usage. As for lexical density, our results are less clear. We suggest that both the characteristics of the tasks and the typological characteristics of Catalan may underlie the lack of sensitivity of this measure. Thirdly, Heylighen's *F*-measure of the level of text formality is more reliable for assessing genre differences than for assessing developmental changes. Lastly, text-embedded lexical usage is somewhat sensitive to participant's' home language and familiarity with Catalan. In the following, we elaborate on each of these findings and discuss a number of the linguistic and educational implications.

Regarding lexical increase, 10th graders produced three times as many tokens as their 1st grade peers. The observed differences also affect the growth of types and lemmas. However, while the size of the lexicon (measured in tokens) did

not differ significantly from 6th grade on, by age 16 their lexicon had a greater diversity and a deeper conceptual underpinning as shown by the fact that the mean number of types and lemmas continued to increase. However, further research is needed in order to examine whether such patterns continue beyond compulsory schooling.

School children's text-embedded lexical usage is sensitive to communicative purposes and circumstances. With age and increased experience with the written language, participants produce texts that are both more informative and better adjusted to genre specific features, affecting text length and lexical quality. In terms of length, joke telling is the first type of text to experience a marked increase in the number of tokens (by 2nd grade). This is most likely due to the combination of an initial command of writing and the reproductive, rather than productive, nature of this type of text. Next comes the narrative genre, as indexed by the explanation of a film, which shows two marked bursts, one between 4th and 6th grades and a second one at 10th grade when it becomes the wordiest type of text. The fact that narrative is the most intensively practiced genre at school may partly account for this finding. The argumentative genre, represented here by the recommendation of a film evidences a text length burst at 4th grade, once command over the notational aspects of writing is presumably achieved and children write for increasingly different purposes. Also, age 9 (5th grade) was found to be a turning point in light of the proposal that 'explanatory discourse' (Blum-Kulka, 2010) allows for background, shared information and sources of knowledge (Goetz & Shatz, 1999). Finally, the data on definitions suggests a burst in the number of tokens around 4th and 5th grades –overlapping with instruction in formal definitions at school. Although definitions, whether of nouns, verbs or adjectives, are systematically the least wordy texts, there are, nonetheless, relevant differences between them. The definition of nouns is the wordiest type of definition followed by the definition of verbs and the definition of adjectives. This may be a consequence of the fact that the definition of nouns is progressively attuned to the canonical definitional pattern whereas the definition of verbs is resolved in most cases by simpler constructions of the sort *córrer*: *que va molt depressa* 'to run: that he/she goes very fast' and the definition of adjectives by the even more concise method of using a synonym, *bo*: *bondadós* 'good: gentle'.

With respect to the application of the four distributional measures, our results indicate that word length, nominalization use and adjective use, are the most powerful measures of text-embedded lexical-usage in Catalan. Developmental changes in word length showed different patterns by type of text. For instance, it increases steadily throughout schooling in definitions, a task that even very young children take as a markedly school-based task requiring specific text structure and high level vocabulary. In contrast, it decreases between 1st and 4th grades in film explanation and recommendation, both tasks allowing for a more relaxed tone involving colloquial language and showing a relatively frequent omission of the written representation of mandatory function words. Later on, the use of increasingly sophisticated vocabulary and of more accurate morphology reverses this tendency causing word length to increase. As for joke telling, it consistently yields the shortest mean word length, possibly due to its oral-like lack of morphosyntactic complexity, and an abundance of interjections and other attention prompts. Word length in this type of text experiences no remarkable developmental changes most likely due to its reproductive character.

Next, we focus on the use of nominalizations in the texts. This was affected by both school grade and, to a lesser extent, type of text. A marked increase starts at 4th grade and continues up to 6th or 7th grade. Then, except for definitions of verbs, it decreases in all types of text during secondary school but gains new momentum by 10th grade, especially for explanations of films. This result does not support our initial predictions. We had hypothesized that the more academic-like the texts (i.e., definition) the higher the number of nominalizations. Whilst the pattern used for definitions of nouns (super-ordinate + relative clause) appeared to foster the use of nominalizations, the pattern used for either definitions of verbs (that + simple clause) or adjectives (synonym) did not. However, a qualitative view that distinguishes between the use of morphologically complex versus non-complex nominalizations would suggest few differences between definitions and film explanations on the one hand and film recommendation and joke telling on the other hand. In other words, our two academic-like texts favoured the use of morphological complexity more than the other two types of texts of a more oral-like nature which accommodate more colloquial and less complex lexical choices.

Distribution of the text-embedded lexicon by grammatical category revealed that the verb category was the most used at each school grade while the adjective was the least used (interjections are dismissed here, as their percentage of use is hardly significant). In line with other studies showing the relevant growth of adjectives in later language development, adjectives (and prepositions) were the only syntactic categories that underwent a sustained increase by school grade in their percentages of use relative to other syntactic categories. Type of text also affected the use of adjectives, though not as markedly as school grade. This particularly evident from 4th grade on, when four types of text: recommendation of a film, explanation of a film, definition of nouns, and definition of adjectives, experienced a marked growth in the use of adjectives. Thus, when recommending a film, participants move from using oral-like formulae such as inviting a friend to watch the film together, to qualifying the recommended film, elaborating on plot description and their personal impressions upon watching it.

Similarly, film explanation grows from an action-based narration in the lower levels to including more in-depth elaboration of both characters and events. The definition of nouns also fosters the use of adjectives, as children gain command of the canonical definitional structure. Finally, when defining adjectives, children naturally provide synonymous adjectives from very early school grades. Definition of verbs and joke telling account for a less pronounced developmental increase in adjective use. In line with nominalizations, type of text had an impact on the distribution of adjectives by morphological complexity. Thus, the two types of texts most marked as academic-like: definitions and film explanations, concentrate higher ratios of morphologically complex adjectives. In contrast, less school-based types of text such as film recommendations and joke telling foster the use of basic morphologically simple adjectives. This characterization by morphological complexity leaves out semantic aspects (Boleda 2006; Ravid & Avidor 1998) that may well contribute interesting information in the processes studied here. and thus should be taken into account in future research.

The computation of lexical density yielded no clear developmental pattern, though we did find some distributional differences by type of text. Unlike other studies in which lexical density showed critical developmental differences (Nir-

Sagiv et al., 2008) or at least a tendency to increase with age, we found no clear increase between 1st and 10th grade. Importantly, our sample did not include adults, the subgroup that most markedly yields an age effect for lexical density (Johanson, 2009). Also, lexical density has been tested as a measure of lexical development in studies focusing on different languages (English mostly, but also Hebrew, French and Swedish) and considerations that it could be language dependent have been suggested before (Johanson, 2009). Consistent with this argument, in Catalan, attainment of a full command of required mandatory use of function words is an important goal of school-based tasks. Consequently, short texts still showing faltering language uses written by young children might be denser than longer, more proficient texts written by their older peers.

A somewhat clearer picture arises regarding the distribution of lexical density by type of text. In line with some previous research (Nir-Sagiv et al., 2008) and in contrast to other studies that suggest limited effects of type of text on lexical density (Johanson, 2009), we found that throughout schooling definitions obtain higher scores of lexical density than film explanations and recommendations and joke telling. In other words, when producing markedly academic-like texts (definitions), participants tend to use more informative, denser language than when (re)producing more oral-like texts such as joke telling and, to a certain extent, film recommendations and film explanations. Thus, lexical density does not appear to be a reliable developmental indicator of texts written in Catalan from childhood to adolescence. Further research focusing on the use of syntactic structures in our corpus may shed light on developmental change in the use of function words that may overcome the shortcomings of this measure.

Thirdly, we found level of text formality to be more clearly affected by type of text than by age. Only definitions, and particularly definitions of nouns, showed a signs of developmental pattern. Definition is a clear exponent of academic-like text and the development of definitional skills is well documented in the literature. It is worth noting that, despite the similarity in the magnitude of increase for the three types of definitions, the definition of nouns starts out with a higher F-score than the other two types of definitions and remains the highest. This should not surprise us, since the definition of nouns, that is, of referential entities, is

commonplace in children's interactions and also we would argue is the most practiced in school. Level of text formality for film explanation and recommendation and joke telling showed no developmental trend. It exhibited only a small overall increase in film explanations and actually decreased in film recommendations and joke telling.

A plausible interpretation of the above findings may be that, unlike definitions, both recommendation and joke telling have an oral-like, informal nature. While this would explain why these two types of text scored lower on the level of text formality, it does not necessarily imply an overall decrease. However, while Heylighen and Dewaele (1999) argue that the high frequency of pronouns reduces the level of the text formality, as we have pointed out above, the appropriate deployment of a full range of pronouns is characteristic of text formality in Catalan. Importantly, it has been argued elsewhere (Teddiman, 2009) that while the *F*-score works extremely well in genre discrimination, – and indeed our results attest a greater effect of text type– it fares less well at accommodating internal make-up differences of the lexical categories upon which it is based. In future research, the corpus driven oriented exploration of the intra category frequency distribution of the use of different pronouns, for instance, may produce more enlightening results. Furthermore, we will pursue the issue of the level of text formality by adding syntactic markers. The inseparability of lexicon and syntax has been established for word production tasks (Tolchinsky et al., 2010). Pursuing this issue by addressing it in text-embedded lexical usage is therefore of relevance in order to obtain a clearer picture regarding the ways lexicon and syntax interact in the use of language in different communicative contexts.

Finally, our results attest to some effect of home language on every text-embedded lexical measure tested here. Participants who identified Catalan either as their sole home language or as their shared home language along with Spanish scored systematically higher than all other groups, implying that the extended opportunity for using the language (at least orally) for a wide variety of communicative purposes and circumstances affects performance in text writing. In contrast, participants who spoke another language or languages at home but not Catalan or Spanish and who have known Catalan for less than 4 years consistently

obtained the poorest results. The exception was lexical density. This may have been due to the fact that these individuals produced the densest texts which we would argue was at least partially due to their lack of full command of the language as evidenced by their erratic use of function words. These findings also suggest that lexical density may not be a suitable measure of text-embedded lexical development when studying both young and far-from-native-like speakers/writers. Interestingly, participants who speak another language other than Catalan and Spanish at home but who have known Catalan for more than four years score considerably better and, in fact, slightly overtake participants who only speak Spanish at home. This would support the view that multilingualism does not harm language development or interfere with academic performance.

4.6 Implications

Our findings have a number of linguistic and educational implications. Firstly, whereas word length, the use of nominalizations and the use of adjectives gain validity as measures of text-embedded lexical usage, there is some doubt as to the reliability of lexical density. Further research on the lexical density of written texts in typologically distinct languages is needed in order to establish the true power of this measure to cross-linguistically assess written texts. Secondly, by including data on non-narratives, this study makes a contribution to developmental studies which, with some notable exceptions (Nippold, 1998; Scinto 1986, Scott & Windsor 2000), have traditionally been more focused on the narrative/expository division. Here, we have included joke telling and film recommendation, two types of texts that remain underexplored so far. As we expected, joke telling yields a less marked developmental pattern than its counterparts. Notably, the number of tokens and word length (both strong developmental measures) show an increase with age but do so at a lower rate than in other types of texts. As for film recommendations, they serve here as "explanation" in the sense used by Blum-Kulka (2010), because they afford an opportunity for the justification of one's arguments. Our results suggest that the recommendation of films is a genre that particularly fosters the use of adjectives, more so than the explanation of films. Thus, we expand the widely held conception that narratives are the most natural setting for the occurrence of adjectives to

include argumentative genres. It is also interesting to note that writing down the recommendation of a film appears to allow for a wide range of register distinctions, from very spoken-like texts, e. g., *he!! as de veure vendela!! no tu pots perdre!! ho fan al dilluns i el dimart!! ho fan a les 10 i 10 o a les 10 i 1/2.* 'Hey!! You must watch it!! You can't miss it!! It's on Mondays and Tuesdays!! It's on at 10:10 or 10:30' to other, far more distant, informative, written-like language, e.g., *Es una sèrie molt entringuda i divertida on la barreja d'humor, drama, amor acció, sarcasme i ironia són constants i molt bé combinats.* 'It is a very entertaining, funny TV movie where humor, drama, love, action, sarcasm, and irony are constantly mixed up and greatly balanced'. The notable effect of type of text on most of the measures tested here provides evidence of the importance of assessment of lexical development in text-embedded contexts.

As for the educational implications, our findings highlight the relevance of providing extended opportunities for practice in a wide variety of genres as part of the school curricula. To the best of our knowledge, this is not always the case in Catalan schools especially before 4th grade, since children are kept focused on the notational aspects of writing. The fact that we have found the 4th grade to be a clear turning point regarding participants' ability to adjust to genre-specific requirements reopens the debate as to the importance of giving young children ample opportunities to gain command of the writing system by producing authentically motivated pieces of text.

The existence of this corpus, which is publicly accessible at http://clic.ub.edu/cesca, enables psycholinguists and educationalists to obtain an updated picture of the state of the Catalan language as it is used in writing by children attending compulsory education from childhood to adolescence.

CHAPTER 5

Developing a written lexicon in a multilingual environment

Anna Llaurado and Liliana Tolchinsky

Universitat de Barcelona

In press in

Grommes, P. & Hu, (eds.): *Plurilingual education: Policies, Practice, language development*. Hamburg Studies in Linguistic Diversity (HSLD) Amsterdam, NL: John Benjamin

Abstract: Children educated in Catalonia are growing in a multilingual environment. Catalan is their school language but not necessarily their home or social language. Our goal was to track the presence of such multilingual input in the written lexicon of 2,436 schoolers throughout compulsory schooling. Participants were asked to write down as many names as they remembered of five semantic fields and to produce 6 types of text. The two corpora were tapped for the presence of non-Catalan and hybrid constructions. Unexpectedly, these accounted for only 3% of the total number of lexical forms in the corpora. The imperviousness of the corpora to multilingual influence is discussed in terms of the constraints placed by the written modality and by the school-situated conditions of task production.

Key words: multilingualism, lexicon, written language

5.1 Introduction

During the school age period children's linguistic interactions undertake important diversification. The family environment is enriched with the introduction of new interlocutors –peers and adults beyond the family– and children enter the 'world on paper' (Olson, 2004) increasing their experience with the written language both as a notational system and as a discourse style. The diversification of interlocutors, communicative circumstances and modalities brings with it different registers and styles, and have a crucial impact on children's linguistic development. The present study focuses on one central component of linguistic development, the lexical component. Vocabulary development has been related to cognitive development and vocabulary knowledge predicts academic success (Cunningham & Stanovich, 1997; Leong & Ho, 2008). In all, the study of lexical development is critical for throwing light on language knowledge, beyond vocabulary acquisition (Bates & Goodman 1997). We examined the growth of Catalan written lexicon throughout compulsory schooling, a growth that takes place in a multilingual environment.

Since 1983 Catalan has been the language of instruction in every Catalonian school. However, the Catalan language holds a status of co-officiality with the Spanish language across the four provinces in north-eastern Spain. Intensive contact between Catalan and Spanish (or other languages spoken in the Catalan territory) favors frequent code switching and mixing (Perera et al, 1999). We aimed at taping how interactions between languages would be reflected in the written language of school age children. More specifically, we looked for the presence of non-Catalan forms in texts written by children and adolescents attending compulsory schooling. Two indicators of such interaction were used: presence on non-Catalan forms in children's written productions and presence of hybrids, i.e., forms in which language mixing occurs within the word, at the morphophonological level. In what follow we put forward the specific goals, methods, and findings of our study and we discuss these findings from an educational perspective. Before that, a brief reference to the current sociolinguistic

situation in Catalonia and some background on the so called process of normalization of the Catalan language is in place.

5.1.1 Sociolinguistic Background

Catalan has been the language used in the current Catalonia since as far back as the medieval days. Such use, however, has been endangered at two different points in history, the first back in the 18th century, the second, more recently, throughout Franco's dictatorship (1939-75), in both instances due to political efforts targeting substitution of Catalan by Spanish.

Shortly after Franco's death, a law for linguistic normalization was passed in 1983, aiming at establishing policies that would counterbalance damages caused to the use of Catalan language by the recently overruled regime, on the one hand, and the arrival of important numbers of Spanish speaking workers during the 50's and 60's, by the other hand. In this frame, the 1983 law, decided for establishing a program of linguistic immersion that was to be applied throughout compulsory school both in state and semi-state schools. According to this program, Catalan was Catalonia's own language '*llengua propia*', and to know it was a right for everyone living in Catalonia. Therefore, and in order to eradicate social division between Catalan born citizens and immigrants, Catalan was to be the vehicular language in schools and all instruction was to be provided in that language. Data supports success of this program in extending familiarity with the language. Thus comparison between years 1986 (total population: 5.856.433) and 2009 (total population: 7.049.900) yields positive results in the use of Catalan: the percentage of people capable of understanding Catalan increased from 90% to 93%, the percentage of people capable of speaking the language increased from 64% to 76%, the percentage of people capable of reading in Catalan increased from 61% to 73% and the percentage of people capable of writing in Catalan increased from 32% to 56% (Idescat).

From the 80's and to the beginning of the 21st century, rate of immigrants had come down to a stable 3%. However, this was about to see a new dramatic

burst: in 2004 the rate of immigrants had risen to a 9.5 % and by 2008 immigrants were a 15 % of the total population in Catalonia. This second wave of incomings was far more diverse than the previous regarding linguistic and cultural origin of the immigrants. Thus, 29% of the new immigrants come from a variety of south and Central America countries, 26% come from Africa (mostly Morocco, 20%), 12% come from Asia (China and Pakistan), 23% from different countries from the EU-27 and 10% from elsewhere.

The current sociolinguistic situation has an important impact on educational policies. Unlike with the previous wave of Spanish speaking immigrants, nowadays teachers in Catalan schools are not t familiar neither with the home language nor with other cultural practices of their students. Also, some of the newcomers are mere passersby and therefore feel not committed with the educational (and linguistic) demands posed by both the school and the welcoming society. A variety of resources has been set up in order to shelter children arrival into a new school. More so since, not infrequently, the school becomes the most real meeting point between the receptive culture and the newly arrived. Welcoming school classrooms 'aules d'acollida' have been created with the purpose to teach Catalan to immigrants within the school context. Budgets have been allocated to create a network of out of school activities conducted in Catalan addressed to both local and immigrant school age children, specially in areas where Catalan is not the preferred language for social interaction and therefore children have little opportunity for extended use of the Catalan language out of the school.

In the past three years arrival of immigrants in Catalonia has slowed down sharply although it continues to experience slight increase and has risen from a 15% of the total population in 2008 to a 16% in 2011. With the financial crisis, however, budget allocated to address this group has been cut significantly.

In sum, in addition to a long standing interaction between Catalan and Spanish, Catalan being more under the influence of Spanish than vice-versa due to its official status in overall Spain, it is not uncommon in nowadays Catalan schools, to find children who use Catalan in school-based tasks and interactions, Spanish in peer exchanges and other public purposes, and one (or more than one) other

language at home. In such scenario more often than not the linguistic productions of a multilingual speaker show linguistic mixing, evidencing interaction between languages. Although this mixing of codes is seen negatively by some authors, as showing lack of competence (Payrato, 1996), others see it as inevitable, and even further an expressive resource that serves communicative goals (August & Hakuta, 1997; Banks, 1993).

Code switching has been widely investigated in multilingual settings. Thus, a variety of functions such as structured play, games, and other activities, negotiating meanings and rights, and asserting their shifting identities and allegiances in the context of spontaneous speech has been researched (Auer 1984, 1998; Garrett 1999; Myers-Scotton, 1995; Paugh 2001; Rampton 1995, 1998;). Although study of code switching is most habitually not situated in a classroom, it is considered a natural occurrence, which can support academic achievement, cognitive development, and multilingualism (August & Hakuta, 1997; Banks, 1993; Krashen, 1996). Despite all the attention received, much of it has focus on oral uses of the language in informal communicative context. Our approach is different, as instead we will identify presence of multilingual interaction in written productions and in a formal context such as a classroom.

5.2 Goals of the study

This study is part of a larger project that aims at exploring developmental patterns of lexical growth in written Catalan throughout compulsory schooling. The first goal of the study is to track the development of the written lexicon from the age of 5 to 16 years old. The diversification of the children's linguistic circumstances during this period together with their school related increased experience with the written language should have an impact on vocabulary growth, most particularly on written vocabulary growth.

In this frame, we assessed lexical development by computing the number of *lexical forms* –i.e., expressions as written by the subject– in two different tasks: one of vocabulary production and the other one of text production. Since a diversity of

lexical forms may be underpinned by the same concept, all the lexical forms were lemmatized. This made possible a separate exploration of the participants' diversity of lexical forms on the one hand and their conceptual vocabulary (measured by lemmas) on the other hand.

As a consequence of the increasing command of written language an increasing ability to adjust to genre-specific features is to be expected. Thus, we examine the use of lexicon in a variety of semantic fields and types of text and predicted that lexical development may differ for different genres and semantic fields.

All the above notwithstanding, it is reasonable to assume that in the context of multilingual environment the enlargement of the size of the lexicon in school age children would result from an increase of Catalan lexical forms, but also from use of lexical forms in other language/s as well as other non-Catalan forms. In the present study, we aim precisely at tracking the presence of such multilingual forms in children's written lexicon throughout compulsory schooling. We use two different measures of language mixing: presence of forms in languages other than Catalan and of hybrid forms. We counted as presence of Spanish (or other language) forms each case where a word in a language other than Catalan was provided (1):

(1) **zapato** 'Spanish for shoe' instead of **sabata** 'Catalan for shoe'

Substitution of a Catalan form by its Spanish (or other language) counterpart may be due to a number of reasons from expressive preferences to lack of knowledge of the required form. Differently, hybrid uses result from a combination of elements belonging to different languages. It can result from straight mixing between the Catalan and the Spanish for one same word (2) or it can be the outcome of a word formation process where a Spanish stem and a Catalan root (or vice-versa) have been mixed (3):

(2) **relampec** from the Spanish **relámpago** 'lightning' + the Catalan **llampec** 'lightning'

(3) **perezos** from the Spanish **perezoso** 'lazy' + the Catalan suffix '–ós'

Thus, it can only be applied to words and not to multiword constructions as can be the case with presence of forms in other languages. Unlike most previous research, which focused on natural speech in informal communicative situations, we explored the influence of multilingual input within a school setting and on written performance. Interaction between languages is considered a main feature of multilingual competence rather than an indication of lack of competence. However, since both presence of words in other languages and hybrid forms represent depart from the norm, their use is considered to denote low level of competence in Catalan (Payrato, 1996). Unfortunately, a majority of teachers share this very normative view and fight hard against their presence in school writing practices.

In this frame we expected presence of words in other languages and hybrids to decrease with school level and this decrease to differ by semantic field and type of text. Some semantic fields are more academic-like in nature than others. For instance, many of the lexical forms belonging to the semantic field of natural phenomena are acquired through science lessons and the reading of textbooks whereas the vocabulary associated to the semantic field of food or clothing are part and parcel of children's daily input in and outside school. Similarly, definitions are a more school-based practice than telling a joke. Therefore, we predicted that the more academic-like the context the lower the presence of non-Catalan forms.

5.3 Method

A sample of 2,436 children and adolescents from 5 to 16 years having a diversity of home languages took part in the study. There were two different tasks. First, participants were asked to write down "as many names as they could remember" of five different semantic fields: food, clothing, leisure activities, personality traits and natural phenomena. After 275 children abandoned the study due to a diversity of reasons, the other 2,161 were also asked to produce 6 different texts: a film explanation, a film recommendation, a joke telling, a definition of a noun, of a verb and of an adjective. Both the vocabularies and texts were gathered on paper because at the time of the study elementary school children were not familiar with word processing. Besides, teachers have

recommended using paper to avoid confusion between orthographic or linguistic errors and typing errors. Completion of the writing tasks took place in the participants' habitual classrooms at the request of their habitual Catalan language teachers who had received training regarding text elicitation. Although there was no time-limit, the task did not last more than one class session.

Additionally, participants filled out a sociolinguistic questionnaire including information on their sex, age, school level and home language or languages. Sociolinguistic questionnaires were always answered before the vocabularies and text writing tasks.

Four different groups were established according to the participants' self-declared home language: (1) Catalan only; (2) both Catalan and Spanish; (3) Spanish only; (4) any language except Catalan or Spanish.

5.4 Some General Features of the Corpora

The corpus of vocabularies included 242,404 lexical forms that were lemmatized into 8,498 different lemmas and the corpus of texts yielded 207,028 lexical forms that were lemmatized into 113,160 different lemmas.

A mirror version was created reproducing with total exactitude the lexical forms as written by participants in both tasks. No spelling corrections were introduced. Due to the nature of the participants, the corpus obviously contained many Catalan forms but it also a great variety of graphic variants, orthographic errors, creative forms of derivation, creative forms of hybridization, other languages, multiword constructions and segmentation errors.

A second 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 standardized only with regard to aspects concerning the conventional separation of graphic words in orthography.

The two corpora (vocabularies and texts) were tapped for the presence of words and constructions in languages different from Catalan and of hybrid forms, that is, forms that combine morphemes from two or more languages.

5.5 Lexical Growth through Compulsory Schooling

Whether tapped by the isolate vocabulary or by the text production task, both lexical forms and lemmas were found to increase markedly throughout compulsory schooling. In particular, 4th and 7th grades turned out to be the two moments when the lexicon experienced most robust bursts. Also, we found a clear impact of both semantic field and type of text on lexical growth. For instance, expression of clothing items produces a high ratio of lemmas over lexical forms, that is, the field shows relatively low semantic-conceptual underpinning (low number of lemmas) but it is expressed with a multiplicity of equivalent lexical forms. In contrast, the semantic field of traits of personality displays a comparatively low ratio of lemmas over lexical forms, that is, a high amount of lemmas is expressed by means of few variants each. Another instance of this effect, the more academic-based the semantic field, i.e., natural phenomenon, and type of text, i.e., definition, the more important the effect of school level. School level affects also spelling uses, both within the word and at the word segmentation level. Thus, the use of Catalan lexical forms correctly segmented and with normative spelling increases steadily throughout school and, accordingly, the use of Catalan deviantly spelt lexical forms tends to decrease, most pronouncedly after 7th grade. This increase was found irrespectively of the participants' home language. There is, however, an interaction of school level and home language/s on the use of correctly segmented and spelled words. Although at every school level, the participants who speak mostly Catalan at home produced more correct occurrences than any other group of participants, the differences between groups diminished with school level, particularly after the fifth year of elementary school.

5.6 Presence of Multilingual Input

In addition to spelling, we were interested in exploring patterns of use of non-Catalan forms and in tapping possible developmental aspects throughout schooling as well as effects of semantic field and type of text.

A group of 314 texts out of the total of 11,882 texts in the corpus were written in straightforward Spanish. However, 162 out of the total 314 belong in the explanation of a joke; therefore the language the joke is habitually told may have motivated the language choice. Definition, the most academic type of texts, concentrated the lowest percentage of texts written in Spanish.

Participants who declared to be primarily Spanish speakers obtained the highest record for Spanish written texts in almost all types of text (2% for definition of a noun; 2% for definition of a verb; 1% for definition of an adjective: 4% for explanation of a film; 2% for recommendation of a film) except for the explanation of a joke where the highest percentage of texts written in Spanish (9%) was obtained by the group who had a neither Catalan nor Spanish as their home language. Participants who had declared themselves as Catalan speakers primarily obtained the lowest percentage of texts written in Spanish for all types of text (1% for definition of a noun, and less than 1% for definition of a verb; definition of an adjective; explanation of a film and recommendation of a film, in contrast there were 4% of texts written in Spanish for explanation of a joke). In all, the number of texts written in Spanish represents less than 3% of the total number of texts.

Non-Catalan forms were also used into the Catalan texts. They were introduced as a Spanish quotation (direct speech) using whole sentences or brief passages (4):

> (4) “Eren tres nens que només deien: "**nosotros tres nosotros tres", " en bicicleta - en bicicleta”, "por el dinero- por el dinero".** ‘There were three children who were saying only: (in Catalan) “the three of us, the three of us”, by bicycle, by bicycle”, for the money, for the money”. (in Spanish)

or as an isolated Spanish word (5) or using a word in another language (6) proposing an hybrid form (7) maybe to fill a lexical gap:

> (5) "…i al seu amic li va inpresionat al fil del dolent i le volie mata i va **quitarse** la mascara i va casi plora…" (*) '…and his friend was impressed to the bad guy's son and he wanted to kill him and he **took** the mask **off** and almost cried…'
> (6) "… quan vag (*) preparar una festa de **jalowin**" '…when I prepared a **Halloween** party'
> (7) "…Esuna ingecsio (*) que serveix per evitar **enfermats**…" '…It is and injection that serves for avoiding **illnesses**'

Despite the fact that classrooms are highly multilingual environments, non-Catalan forms accounted for 4% (2% of Spanish forms and 2% of hybrids) of the total number of lexical forms in the corpora only. School level had a moderate effect on simultaneous use of terms from different languages. Presence of non-Catalan forms is light in 1st and 2nd grade (M = 0.25, 0.28, respectively), probably due to constraints placed by the learning-to-write process. It increases visibly y 3rd grade (M = 0.41) and holds steady on. Only in 9th and 10th grade it shows a tendency to recede (M = 0.23, 0.24 respectively).

Semantic field and type of text had an impact on the use of non-Catalan forms. In the vocabulary task, distribution and type of non-Catalan forms was related to semantic fields. Thus, clothing items and leisure activities were the two semantic fields that presented a greater presence of forms in other languages. English words such as top, legging or shorts are common use in the clothing semantic fields just as *ballar hip-hop* 'dancing hip-hop', *practicar break-dance*, 'practicing break-dance' *anar en mountain-bike* 'riding on mountain-bike' are not at all rare in the leisure activities field. At the other end, natural phenomena yielded the lowest presence of forms in other languages while fostering use of more specialized terms learnt, in many instances, through school-based activities or through reading of specialized textbooks. To illustrate, terms such as *sisme* 'seism', *sedimentació* 'sedimentation', *fossa oceànica* 'oceanic trench', appear in the natural phenomena semantic field are a by-product of school related knowledge. Hybridization was most present in the food semantic field, most likely as a

consequence of children to referring to products they consume at home for which they do not have the Catalan term. See for example (8) and (9):

(8) **piment** 'from the Spanish **pim(iento)** ?green pepper + Catalan suffix *ment*'

(9) **ou frit** from the Spanish word **fri(to)** 'fried' + the Catalan suffix *it*.

The particularities of the process of lemmatization we have applied to the vocabularies distinguish between idiosyncratic mixing of languages, use of other languages in cases when the equivalent forms are available in Catalan and use of lexical items that fulfill lexical gaps in the Catalan language which are mostly specialized terms coined in foreign languages and that are in the process of becoming incorporated to the Catalan dictionary as such. The description pointing at the finding that the overall use of non-Catalan forms remain stably low in each semantic field concerns the first two cases but does not include the third possibility. See for example (10), (11) and (12):

(10) **ballar hip hop** (Catalan + English) 'to dance hip hop'
lemmatized: **ballar** (Catalan) 'to dance'

(11) **anar en mountainbike** (Catalan + English) 'to go mountainbiking'
lemmatized: **anar en bicicleta** (Catalan) 'to ride on bicycle'

(12) **practicar breakdance** (Catalan + English)'to practice breakdance'
lemmatized: **practicar un ball** (Catalan) 'to practice a dance'.

As regarding text-embedded use of words in other languages, word definition presented the smaller number of such forms and jokes the largest (M= 1.29, 0.07, respectively). A likely explanation to this would be that jokes are highly associated with orality and therefore not much restricted in terms of code switching. In contrast, definition is a rather school based task that imposes that performance should abide by the norm. As for use of hybrids, they appeared least in word definition (M = 0.03). However, they experienced similar distribution

concerning jokes, explanation of a film and recommendation of a film (M = 0.87, 0.90, 0.96, respectively).

Catalan monolinguals produced the least number of both Spanish and hybrid forms whereas the participants who speak neither Catalan nor Spanish at home produced the most forms in other languages. The bilingual Catalan-Spanish participants produced most hybrid forms, showing morphophonological manipulation skills that are beyond non-native ability.

5.7 Discussion

We have found far less presence of multilingual input in the written productions of Catalan schoolers than expected. In spite of the fact that for a majority of these participants Catalan is mainly the language of school, in spite of the strong immigration and the fact that schools are multilingual environments, the written lexicon elicited in a classroom context and by a language school teacher shows a low level of permeability to multilingualism.

We did find developmental changes in the spelling patterns but neither in the presence of words in languages other than Catalan or hybrids. We have witness an increase in the use of Catalan correct forms with school level —with the concurrent decrease in the number of deviant forms— with school level, across semantic fields and types of text. This improvement is likely to be caused by the exposure to written texts and practice with literacy activities. Neither the presence of non- Catalan forms nor hybrids show relevant differences with school level, rather their incidence was notably low in the two corpora. Indeed, there are clear-cut differences between these two indicators of interaction between languages. Hybrids result from an interaction at a morphological or phonological level and entail a capability to manipulate words at this level but this process produces non-words. Hybrids usually come to fill lexical gaps and may uncover low lexical competence in the language. The use of foreign or Spanish forms, instead, implies the incorporation of words or constructions that are part and parcel of other languages and might be eventually incorporated into Catalan. They

are still absent from Catalan reference dictionaries but they often refer to technology, sports, fashion-related items or activities that are vividly, though orthographically unstable, present in media discourse and in everyday discourse in consequence.

As said, however, the presence of both hybrids and forms in other languages was negligible in our corpus. We believe that two inter-related characteristics of the elicitation procedure might explain this apparent lack of permeability of the corpus: the fact that we collected written productions and the fact that they were collected by participants' usual Catalan teacher. The written modality enables better monitoring and lexical selection so that the use of non-Catalan forms, in particular Spanish forms, which is not highly appreciated in written works produced at school, becomes restricted. The specific constraints of the written modality and the fact that writing is perceived as more formal than speech may explain the scarcity of non-Catalan forms. This finding confirms the prevalence of Catalan forms in the written modality that was reported by previous studies comparing the written and the spoken modality (Perera et al, 1999). Nevertheless, more research is needed to determine the precise differences between the two modalities, in particular the extent to which this apparent control over the presence of non-Catalan forms may affect lexical richness and fluency.

Another characteristic of the elicitation procedure that we deem related to the low permeability of the corpus to multilingual input is that the gathering was undertaken at school with the usual teacher. It is evident that the requirement was interpreted by the informants as a school task rather than as a communicative activity. The informants produced what it was expected from them in a Catalan class. In this sense, it would be useful to analyze written productions that are fulfilling an authentic communicative function rather than fulfilling a school requirement. In this way we would be able to separate the uses of language that respond to the constraints of modality to those that relate to speakers' perception of the task.

5.8 Implications for the study of multilingualism

This work has some relevant implications both for language assessment and for language didactics. As for assessment, the study highlights that in order to characterize the state of knowledge of a language it is necessary to take into account multiple dimensions. Our study shows that, at a similar age, speakers display a wide variety of forms for naming clothing but are far less fluent for denominating traits of personality. Thus, in order to assess the level of vocabulary richness we must considered vocabulary use in different semantic fields. In the same line, our study shows that speakers of similar age are able to use higher register forms for defining words and more colloquial uses for telling a joke. Moreover, telling a joke involves the (re)production of a text whereas definitions entails the use of metalinguistic knowledge. Thus, for the assessment of vocabulary in text-embedded contexts, different genres should be contemplated because each genre promotes access to differing kinds of language use and reflection on language.

In addition to semantic content and genre, the characterization of linguistic knowledge should also consider the **modality** of production. Oral and written production abide by different constraints. Writing lightens the online pressure imposed by production of speech and furnishes the writer with more time and editing facilities. These conditions of production may allow the emergence of not yet automated constructions in writing, although not yet in oral discourse. Including written performance is of relevance for any characterization of linguistic knowledge because it enables to tap more elaborated forms of language use. The written modality plays therefore a double function: in the development of language it becomes the necessary platform without which the remarkable changes that occur at the lexical, morphosyntactic and discursive levels could hardly take place (Berman & Ravid, 2008) and in the use of language it facilitates the deployment of a different level of linguistic competence.

Finally, the context of the task is another dimension to be taken into account for assessment because it also places different constraints on the speaker/writer performance. Chatting, for instance although performed in writing imposes certain

requisites that are markedly different from those required by a writing task performed in a school environment under the surveillance of a language teacher. Writing in class with the language teacher is perceived even by very young participants as requiring some degree of formality. As seen, the context of production was considered in the present study as a main explicatory factor of the scarcity of non- Catalan forms in the discourse production of school children educated in multilingual environments.

The number of dimensions to be taken into account for characterizing the linguistic knowledge of multilingual speakers necessarily increases. Multilingual speakers do not use each of their languages independently from each other but, on the contrary, a diversity of interactions takes place between the languages learned and between the role these languages play in the learner's environment (Cenoz, 1997). The complexity involved in the assessment of the linguistic knowledge of multilingual speakers has turned this task into one of the pending topics in the study of multilingualism.

The didactic implications of the study are very much related to the multiple dimensions that enter in the characterization of linguistic knowledge. In a nutshell: school should provide children with a wide range of activities so as to mobilize the inherent variability of language use (Biber, 1995). Multilingual classrooms offer an unparallel occasion for enhancing diversity in use and depth of reflection. The presence of various languages and different levels of competence enables paraphrasing, translation, comparison of different forms fulfilling a similar communicative function, idioms, and idiosyncratic uses, multiplicity of grammatical constructions, variety of styles and rhetorical options. The idea is to take advantage from the extant diversity for developing linguistic awareness through comparison of linguistic expressions between and within languages. By multiplying the learner opportunities to address different audiences for different purposes in a diversity of circumstances (and in different languages) speaker/writers will not only gain more experience with language but also will increase their ability to reflect on language. It is this ongoing interaction between use of and reflection on language that causes gradual change in the speaker's linguistic representation that leads him to progressive command of a richer

repertoire deploying flexible and appropriate use of a vast range of discourse functions in differing contexts so as to turn native (multilingual) speakers into expert users of language.

CHAPTER 6

The development of syntactic complexity and its relation to lexical growth in the Catalan written language

Anna Llaurado and Liliana Tolchinsky

Universitat de Barcelona

Submitted to Journal of Child Language

Abstract: Using increasingly complex syntactic structures flexibly deployed for a wide range of communicative purposes is a key attainment of later language development. Although syntactic complexity is not restricted to the domain of written language, written texts are a favoured setting for complex uses of syntax. With age children's texts include longer clauses, partially due to use of extended more complex noun phrases. Also, children improve their ability to organize the flow of discourse in their texts and move from linearly chaining their statements to hierarchically packaging the information. Syntactic structures are clothed by lexical items: with schooling children gain experience with language and command over an increasingly sophisticated and more complex lexicon framed by denser and tighter syntactic structures. In this study we take a corpus-based approach and track the developmental pattern of syntax use as shown by Catalan schoolers in two different sites: the clause and the noun phrase, as they occur in three different types of text: an explanation of a film, a recommendation of a film and a joke telling. The texts were produced by Catalan school children attending 2nd (*M*=

7; 6), 6th (*M*= 11; 5) and 10th (*M*= 15; 6) grade, three turning points of lexical growth as revealed in previous corpus-based lexical characterization on the same texts. Our results show sustained but slow increase in use of complex syntax. 10th grade emerges as a cut-off point in syntactic complexity and, in particular, explanation of a film as the preferred type of text for complex structures. Measures of text-embedded lexical and syntactic usage correlate although these correlations do not yield, however, a concluding pattern.

Key words: complex syntax, complex noun phrases, developmental syntax, relationship between lexicon and syntax

6.1 Introduction

Later language development is characterized by increase in the ability to use one's linguistic repertoire flexibly for communicating a wide range of purposes (Ravid & Tolchinsky, 2002). Thus, although children as young as 5 are able to produce grammatically well-formed multi-clausal sentences (Diessel 2004), they still have a long way to go before becoming proficient language users, able to produce and comprehend mature complex linguistic productions deployed in a wide variety of genres, both in the spoken and written modality. From a developmental perspective we see the organization and re-organization of linguistic forms (Ravid & Berman 2008) as embedded in discourse, which provides children with "a developmental mechanism" (Hickmann, 2003, p. 335) for the acquisition of increasingly complex linguistic devices. This study aims at tracking developmental changes in text-embedded syntax usage by Catalan schoolers in three different school stages (2nd (7;6 years), 6th (11;5 years) and 10th (15;6 years) grade school).

Lexical items clothe syntactic structure, without words there is no way for the syntax to be realized. From a usage-based perspective, that is, considering development in real world language use, the claim that the syntax and the lexicon are two sharply distinct components of language, as has been proposed by generative linguistics, is hard to sustain. Constructions stored in memory include both syntactic phrases and clauses (Jackendoff, 2002). And corpus-based analyses

of real language reveal abundant evidence of cases in which specific lexical items go with and/or require certain grammatical structures. Thus grammar is seen as "built up from specific instances of use that marry lexical items with constructions" (Bybee, 2006: 21). With age lexical uses become increasingly sophisticated and morphologically complex, new lexical items expand their initial meaning to other polysemous, and more abstract new ones. A second aim of this study is to examine the relation between (selected features of) lexical command and syntactic complexity.

The ability to skilfully communicate for different goals has been related to the speaker's selection of genre-specific features in different communicative circumstances. Genres, understood as socially constructed language practices serving specific social purposes (Halliday & Hasan, 1985), may present differences in the micro-level aspects (linguistic features) as well as the macro-level characteristics (overall organizational principles and text structures) that each takes to express different ways of making meaning. Although young children can recognize and produce a variety of genres in oral language (Hudson & Shapiro, 1991; Purcell-Gates, 1988), such ability in understanding and producing written genres, however, is developed gradually and, not rarely, with difficulty (Snow & Uccelli, 2009), i.e., it has been shown that it takes children several years of schooling before they master written expository text, the most academic-like type of text (Berman & Verhoeven, 2002).

School-based engagement in literacy activities, provides the speaker (and now writer) with a unique way for analysing, reinterpreting and fine tuning his linguistic productions. Although use of complex syntax is not at all restricted to the written language (Biber, 1988), this modality alleviates some of the time pressure involved in online processing of spoken productions, therefore providing writers with more time to pack information into increasingly complex structures (Stromqvist, 1999). From a functional, discourse-oriented perspective, such deployment of complex syntax involves the growing ability to skilfully organize the flow of information, not only in the form of linear chaining, but as hierarchically packaged constructions (Verhoeven, Aparici, Cahana-Amitai, van Hell, Kriz & Viguie-Simon, 2002). Through schooling, increased command of the written language and higher levels of (meta)linguistic consciousness will affect both the

spoken and the written modalities, each influencing the other's development in late childhood and adolescence (Ravid & Tolchinsky, 2002, Jisa 2004, Tolchinsky, Aparici & Salas, 2012). In this study we will analyze how the participants change their syntactic uses in three different types of written texts: an explanation of a film, a recommendation of a film and a joke telling. Whereas spoken and written narratives have been extensively researched both form a linguistic and a psycholinguistic perspective (Berman & Slobin, 1994; Fey, Catts, Proctor-Williams, Tomblin, & Zhang, 2004; Longacre, 1996; Mackie & Dockrell, 2004), to date documentation on other genres is less abundant.

Later development of syntactic complexity has been evidenced, for instance, by use of longer clauses, increased use of subordination, decreased use of personal pronouns and more complex noun phrases (Berman, 2009; Myhill, 2009; Ravid & Levie, 2010; Scott 1988). Number of words per clause has been shown to increase by different operations (for example noun or verb phrase expansion) and it has been used as a means to discriminate developmental differences in syntactic uses. This measure reflects increasing density of packaging of more information inside a given syntactic unit (Saltzman & Reilly, 1999). In this line, syntactic structures associated with academic language allow the writer to compress several propositions into a single clause. To the extent that writers use such devices, such as nominalizations, attributive adjectives, and prepositional phrases, the text they produce is likely to have longer clauses. Therefore, growth of clause length is an expected outcome of school-based writing, especially as high school students engage in writing more argumentative or expository prose (Beers & Nagy, 2009; Ravid, 2005).

The very notion of syntactic complexity, notwithstanding the difficulties of defining such notion (see Szmrecsányi 2004; Cosme 2008), has been often linked to the developmental uses of subordination, considered a more complex syntactic phenomenon than juxtaposition and coordination. The domain of 'complex syntax' is typically associated with the traditional notion of complex sentences (Lyons, 1977), in linguistics (Bybee & Noonan, 2002; Cristofaro, 2003; van Valin, 2006), in language acquisition (Diessel, 2004; Lust, Foley, & Dye, 2008), and in pedagogically and clinically motivated research. It has been pointed out that use of subordinate clauses, relative clauses, and other syntactic devices such as complex noun

phrases, allows a writer to express more complex ideas (Coirier, Gaonac'h, & Passerault, 1996) and the possibility to do so more succinctly. Cause-and-effect relationships, expression of manner and conditionality, for example, may require the use of subordinate clauses (Olson & Astington, 1990). Adverbial clauses, used by the writer for a range or functions from providing new information, to organizing the information flow in the on going discourse (Chafe 1984; Givon 1990), are complex to process, and have been found to predominate in formal academic-like written texts compared to other genres/modality (Loban 1976; Scott & Stokes 1995). Research converge on a relationship between such use of increasingly complex syntactic structures an increased acquaintance with and skilfulness in school-related writing practices (Reilly, Zamora, & McGivern, 2005; Schleppegrell, 2004).

Though so far less researched, developmental changes in the use of noun phrase structure has been shown to be another important facet of syntactic acquisition from middle childhood to adolescence. In fact, it has been contended that syntactic development during the school years is concentrated at the phrase level rather than at the clause level (Loban, 1963). Even though this syntactic category is well established by age 3, a clear and consistent developmental increment has been found in noun phrase complexity measured in length in words, syntactic depth and number and nature of pre and post noun modifiers (Chafe & Danielewicz, 1987; Scott, 2004). Literate productions are the natural setting for developing high rates of syntagmatic density, heavy noun phrases in new grammatical roles (other than as post-verbal elements) containing (recursive) prepositional phrases (a category, for all the matters, whose rate of use has been suggested to be linked to the level of quality of a text (Loban, 1976)) and relative clause constructions, known to be a late developing usage in children's oral narratives and characteristic of advanced level writing in different languages (Berman, 1998; Loban, 1976; Scott, 1988). Expository texts composed by schoolchildren and adolescents have been found to contain more complex noun phrases compared to narratives (Ravid & Berman, 2010).

Studies on early language development, young and pre-school children, has shown that the emergence and elaboration of grammar are highly dependent on vocabulary size both in typical and atypical populations and in different languages:

Italian and English (Caselli, Casadio, & Bates, 1999; Fenson, Dale, Reznick, Thal, Bates, Hartung, Pethick, & Reilly, 1994; Marchman & Bates, 1994; Marchman & Thal, 2005) Irish (O'toole & fletcher, 2012) Finnish (Stolt, Haataja, Lapinleimu & Lehtonen, 2009) Icelandic (Thordardottir, Ellis & Evan, 2002) Slovenian (Marjanovic-Umek Feknoja-Peklaj & Podlesek, 2012).

The relation between early gains in syntactic development and preceding development of morphological richness in Catalan speakers did not yield, however, concluding evidence (Serrat, Sanz & Bel, 2004). Nevertheless, the significant correlations that emerged between lexical and syntactic usage for both English and Hebrew, two typologically different languages, in texts produced from middle childhood across adolescence yield support for a connection between command of the lexicon and grammatical development beyond early childhood (Berman, 2004).

The increasingly abstract and academic nature of school-based texts entails use of densely informative linguistic constructions, with high rates of occurrence of content-word, particularly nouns and adjectives, supporting a complex hierarchical, and varied syntactic architecture (Berman & Ravid, 2008; Ravid & Levie, 2010; Ravid & Berman, 2009). Literate uses of language make use of high register specialized words, whose meaning may have been metaphorically expanded over previously extant items, from various knowledge domains. The literate and the core lexicon differ in size but also in quality. Such advanced lexicon is enriched with items lying in the border between the major lexical categories and some of the function words such as adverbials, connectives and discourse markers that glue the linguistic forms in coherent and cohesive pieces of text.

6.2 The current study

We focus on the developmental pattern of syntax use as shown by Catalan schoolers in two different sites: the clause and the noun phrase, as they occur in three different types of text: an explanation of a film (accounting for the narrative genre), a recommendation of a film (accounting for the argumentative/persuasive genre), and a joke telling (accounting for a more conversational oral-like genre) produced by school children of different ages. Our texts are part of the corpus CesCa consisting of texts of different types written by Catalan school children and adolescents attending compulsory school (Llaurado, Marti & Tolchinsky, (2012).

Clause complexity will be measured by means of presence of adverbial subordinate clauses in the texts. In Catalan subordinate adverbials can be realized by either finite or non-finite clauses and they can be ante posed or postponed to the main cause. Adverbials are linked to the main clause by an adverb or a conjunction plus a finite verbal phrase (and verbal complements), or by a preposition (followed by an infinitive). Adverbials realized by either a gerund or a past participle do not take any subordinating particle.

Syntactic complexity at the noun phrase level will be measured by means of (recursive) presence of noun complement's prepositional phrases and/or relative clauses. In Catalan noun phrases range from a null pronoun realization (or zero anaphora) to an overt pronoun or a lexical expression. Lexical noun phrases are the ones with greatest potential for within- and between-subject variability. They range from as little as one word –in the case of a head-only noun phrase to remarkably long strings, including the noun phrase head plus several optional elements –determiners, modifiers, or others. In pro-drop languages, as is the case of Catalan, these options are available in subject position. Syntactic functions other than subject must choose between an (stressed or unstressed) overt pronoun and a lexical expression. Catalan has three main forms of noun complements: conventional adjectives, prepositional phrases, non-finite adjectival phrases and relative clauses. Typically, the relative clause is considered a type of noun modification structure which constitutes part of a complex noun phrase. The relativizer (*que*) is an invariant form not marked for gender, number or animacy, it is obligatory and follows the noun. Consistent post-nominal placement along with morphological simplicity contributes to the salience an accessibility of relative clauses and precocity of acquisition.

We expect an overall increase of syntactic complexity with school grade. However, we predict such development to present different patterns depending on linguistic site of occurrence and to be shaped by the writer's communicative purposes. We hypothesize a more marked increase in complexity at the noun phrase level compared with the clause complexity.

We expect narratives to emerge as the earliest favoured site for use of complex noun phrases where embedding relative clauses in the noun phrase would serve the function of presenting, and characterizing the characters/objects

present in the narratives. However, level of noun phrase complexity in recommendations is expected to level with narratives with school grade as a result of increased use of heavy subjects in recommendations. We expect noun phrases in joke-telling to experience a moderate developmental increase in complexity as a consequence of the reproductive nature of this type of text.

We expect complexity at the clause level to increase also although at a slower pace. Specifically syntactic subordination by the embedding of clauses into other clauses is expected to be increasingly favoured by narratives, in the first place, as a consequence of improved experience with the written language, and recommendations later on but not so much by jokes. Both narratives and recommendations would foster of adverbial clauses as verbal complement, they would serve to encode the reasons, conditions and/or the manner in which the narrated events happened and to argue reasons for and expected benefits of watching the film.

In previous research, the texts in the CesCa corpus were characterized for text length and text-embedded lexical usage by means of a range of distributional measures: word length, use of nominalizations, use of adjectives and level of text formality (measured by Heylighen's *F*-measure on the basis that more formal detached texts will be more noun-based against less formal involved texts which will have a more verbal nature). Overall, texts grew longer with age and richer in morphologically complex, longer words. With one lexical category –adjectives— showing particular growth from middle school on.

The explanation of a film, which accounts for a narrative and is habitually practiced at schools, showed more formal, sophisticated lexical uses than the other two types of text. Recommendation, which accounts for an argumentative text, showed to be a natural platform for use of adjectives and, with age, it lost some of the spoken-like involvement and became more detached and formal. The joke telling which reproduces some of the informal colloquial features of the spoken language, showed less developmental changes (Llaurado & Tolchinsky, 2012).

We found three peaks (in 2nd, 6th and 10th grade) in the development of lexical growth. The peak in 2nd grade was interpreted to reflect an increased command of the transcription abilities, this allowing children to write more, and more at ease. The peak in 6th came as the culmination on a sustained process of

lexical growth (in both quantity and quality through grade school. In contrast, growth in 10th grade occurred after a period of stagnation of lexical growth, between 7th and 9th grade with participants producing fewer tokens and lemmas between 7th and 9th grade, and showing a down turning tendency in their use of nominalizations and adjectives. In this study we will examine the relation between lexical command and syntactic complexity. We expect this relation to be stronger in 6th grade than in 2nd and 10th grade due to the differences in the lexical behaviour in the preceding grades

6.3 Method

In this study we follow a corpus-based approach. We draw on the three of the types of text compiled in the CesCa corpus.

6.3.1 Participants

A total of 180 participants took part in this study. At the moment of the study they were attending 32 different schools in Catalonia. They were from three different school grades . Specifically, 60 participants were in 2nd grade (group mean age = 7; 6), another 60 were in 6th grade (group mean age = 11; 5) and a third set of 60 participants was in 10th grade (group mean age = 15; 7) grade. The texts produced by these three groups had revealed marked lexical growth in a previous study.

6.3.2 Task and Procedure

Participants were instructed to produce three different types of texts: a film explanation, representing the narrative genre, following the instruction "*think of a film or TV series that you like and tell us about it*", a film recommendation, accounting for the argumentative genre, with the instruction "*think of a film or TV series that you like and recommend it to a friend*", and the telling of a joke, accounting for the colloquial genre, following the instruction "*think of a joke or funny story that you know and tell it*".

Children performed the task in their habitual class groups. Texts were written by hand – in order to avoid possible textual deviations due to a lack of text processing skills. Text writing took place in the participants' regular classrooms at

the request of their usual Catalan language teachers. The teachers received training in text elicitation. The task did not last more than one class session. The task was carried out as part of their everyday school activities. A total of 520 texts was generated.

6.3.3 Data processing

All texts were digitalized in a mirror version, that is, that exactly reproduces the hand written texts as were produced by the participants. In order to prepare the data for syntactic processing, texts were manually segmented into clauses in a second, stripped, version. A clause was defined as "a unified predicate describing a single situation (activity, state, or event)" following the work by Berman & Slobin (1994). The limits of the clauses were established on a case-by-case analysis. Punctuation was considered except in cases where it had been used non-conventionally. The segmentation does not show hierarchic relationships between the clauses. Both predicates finite verbs (1a) and predicated including non-finite verbs (1b) are considered separate clauses (misspellings have been corrected in the following excerpts).

(1a) *M'ha agradat molt la part* 'I have liked much the part '

on fan un ball 'where (they) make a dance'

perquè els nens marginats del carrer tinguin un lloc 'in order for the children marginalized from the street have a place' (= in order for the marginalized street children to have a place)

(1b) *per aprendre a ballar* 'for to learn to dance' (= where to learn dancing)

i sense estar sempre al carrer 'and not (they) are always in the streets'

fent tonteries 'doing nonsense'

i traficant droga. 'and trafficking (with) drugs.'

(Type of text: explanation of a film; grade: 6th grade; corrected spelling)

Clauses that contain a verb and coordinated clauses with verb gapping are considered as separate clauses (2).

(2) *guanyan un todo tereno* '(they) win a SUV'

i una maleta amb cinc cents euros 'and a case with five hundred euros'

(explanation of a film, 6th)

6.3.4 Criteria of analysis

In order to characterize the development of syntactic complexity for different communicative goals with school level we have examined:

(a) The number of clauses included in a text (MNCL)

(b) The mean length of clauses (MCL) measured as the total number of words in the text divided by the total number of clauses in the text.

(c) The syntactic complexity at two different sites: (a) noun phrases and (b) clauses.

(c1) Every noun phrase in the corpus was identified and coded for two types of nominal complements: propositional phrases and/or relative clauses. Noun phrases containing at least one propositional phrase or a relative clause was considered as a complex noun phrase. For each text we computed an *index of noun phrase complexity:*

I _cnp= total number of complex noun phrases/ total number of noun phrases

(c2) Every clause in the corpus was identified and coded for presence of verbal complements (adverbial subordinate clauses). For each text we computed an *index of clause complexity*

I_ccl= total number of adverbial clauses/total number of clauses.

In previous research, we had characterized the development of text-embedded lexical usage for different communicative goals with school grade (for a full review see Llaurado & Tolchinsky, 2012). In that study, for each text we had computed:

(d) Text length: total number of tokens produced

(e) Word length: measured by number of letters per word

(f) Use of nominalizations: measured as the proportion of nominalizations relative to the total number of words in the text

(g) Use of adjectives: measured as the proportion of adjectives relative to the total number of words in the text

(h) A index of text formality was computed using Heylighen's *F*-score (Heylighen et al., 1999):

F = (noun frequency + adjective frequency + preposition frequency + articles frequency– pronoun frequency – verb frequency – adverb frequency – interjection frequency + 100)/2.

6.4 Results

This section consists of four parts. Firstly, we present quantitative results about the mean number of clauses (MNCL) and mean clause length (MCL) by school grade and type of text. Secondly, we present quantitative results about the Index of complexity at the noun phrase level (I_cnp). Thirdly, we present quantitative results about the Index of complexity at the clause level (I_ccl). Fourthly, we present the correlation and regression analyses concerning the syntactic and lexical configuration of the texts.

A series of one way ANOVAs school grade (3) with repeated measures on type of text (3) were performed on the distribution of the measures for characterizing the syntactic complexity of the texts (MNCL, MCL, I_cnp and I_ccl) in order to determine the effect of school grade and home language as well as possible interactions on these dependent variables. The eta squared value (η2) is used to report the effect size of both the main effect and the interactions. The size effects of relevant pairwise comparisons are reported using Cohen's *d*. An alpha level of .05 was used for all statistical tests. When the assumption of sphericity was found to be violated, degrees of freedom were corrected using Greenhouse_Geisser estimates.

6.4.1 Text length in clauses and number of words per clause

The 540 texts contained 11,806 tokens grouped in 1,672 different clauses.

Table 1. Mean number of clauses (MNCL) and mean clause length (MCL) (SD) by school grade and by type of text.

Grade	Type of text	MNCL	SD	MCL	SD
2nd gr.	Joke telling	5,02	4,36	5,15	2,44
	Recommendation	2,15	1,36	4,25	1,4
	Film explanation	1,93	1,89	5,59	2,01
6th gr.	Joke telling	5,6	4,3	5,43	2,2
	Recommendation	3,28	2,09	4,78	2
	film explanation	4,48	4,65	6,66	3,57
10th gr.	Joke telling	4,7	3,05	5,28	2,07
	Recommendation	4,53	2,84	4,76	1,5
	Film explanation	6,17	3,83	6,38	1,6

With age, children produce significantly longer texts by means of increasing the number of clauses $F(2,180)=15,245$, $p<.001$ $\eta 2=.15$. The clauses increase in number of words even if moderately $F(2,180)=4,242$, $p=.016$ $\eta 2=.05$.

We also found a significant effect of type of text on both mean number of clauses $F(2,180)=14,252$, $p<.001$ $\eta 2=.08$ and mean clause length $F(2,180)=24,486$, $p<.001$ $\eta 2=.12$ but the interaction school grade by type of text had a significant impact only on number of clauses $F(2,180)=8,144$, $p<.001$ $\eta 2=.08$.

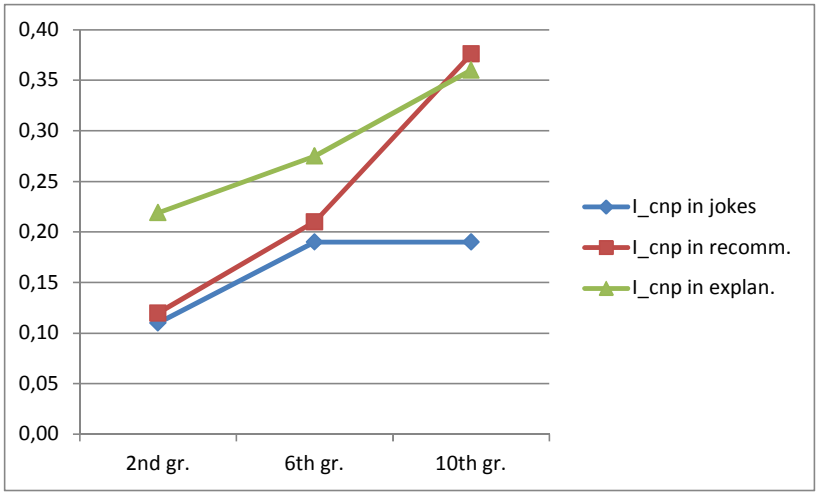
The effect of type of text on MCL changed with school grade. In 2nd grade MCL in recommendations was significantly lower ($p= .030$) than in both explanations and joke telling (these two were not significantly different). In 6th grade, MCL increases in recommendations but not in joke telling and the

differences between these two types of text lost significance. In explanations MCL increases too, and yields significantly longer MCL than in recommendations but not up to significantly longer clauses than in joke-telling. Finally in 10th grade, MCL in explanations is significantly longer than in both recommendations and joke telling.

6.4.2 Syntactic complexity at the noun phrase level

Figure 1 plots the developmental changes in the index of complexity in noun phrases for each type of text.

Figure 1. Plotted mean index of complex noun phrases by school grade and by type of text.



With schooling participants produced significantly more complex noun phrases $F(2,180)=8,278$, $p<.001$ $\eta 2=.12$. This development of complexity has a slow protracted nature as shown by post-hoc comparisons, which revealed significant differences only between 2nd, and 10th grade. The index of complex noun phrases was moderately affected by type of text also $F(2,180)=4,704$, $p=.010$ $\eta 2=.04$. Overall, the proportion of complex noun phrases was significantly higher when participants produced explanations than when they told a joke.

Noun phrase has come out as a powerful platform for school-age syntactic

development. This growth follows different patterns depending on the communicative goals. Thus, for school grade children, only explanations appear to elicit explicit expression of full noun phrases (underlined) referring to the involved characters, elements, time and locations (3) (misspellings have been corrected in the following excerpts).

(3) *Donç que hi ha un esquirol que és molt tonto, i que li* (pron. 3rd person sing.) *encanten les galetes*. 'Well that there is a squirel that is very dumb and that him (li) loves the cookies.'

(explanation, school grade: 6th grade, corrected spelling)

Instead in recommendations, the recommended film is very often referred to deictically by means of referential pronouns and content is concerned with qualifying the recommended object and much less with developing it (4).

(4) *Mira'l perquè és divertit* 'watch *it* because (it) is fun'

(recommendation, 2nd grade, corrected spelling)

This difference is maintained by 6th grade in spite an overall improvement in the quality of the text produced due to improved coherence and use of a more diverse range of adjectives (5).

(5) *Li diria que l'anés a veure que és molt divertida i emocionant* 'I would tell him to it go to see that (it) is very fun and exciting.

(recommendation, 6th grade)

Only by 10th grade, the participants show ability to overcome the more spoken like orientation of the recommendations, and to take a more detached stance, including explicit information about the commented film, conferring the text a more formal structure closer to written language (6).

(6) *Són sèries divertides i entretingudes que tracten la vida quotidiana de qualsevol persona, encara que surten coses que no, però fa riur*e.

'(They) are series fun and entertaining that talk about the daily life of anyone, although come up things that do not, but (it) makes laugh' (= They are fun entertaining series talking about anyone's daily life, although not everything, they make you laugh)

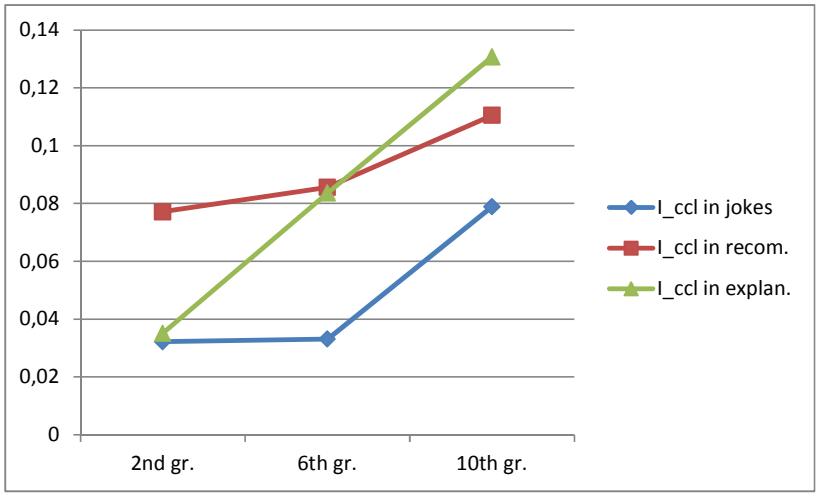
(recommendation, 10th grade)

6.4.3 Syntactic complexity at the clause level

Figure 2 plots the mean index of clause complexity in 2nd, 6th and 10th grade for the three types of text. It shows the significant effect of school grade F(2,180)6,891, p<.001 ŋ2=.07. The developmental process of getting to produce complex clauses in one's writing is a lengthy one, as shown by post-hoc contrasts, which revealed that clauses produced by 10th graders were only marginally more complex than 6th graders' (*p*= .047) but clearly more complex than 2nd graders' (*p*= .002).

Type of text had a moderate impact on clause complexity F(2,180)=4,306 , p=.014 ŋ2=.02. Post hoc comparison revealed that participants produced more complex clauses when writing recommendations and explanations than when writing jokes. These contrasts were only significant in 10th grade between explanations and joke telling (p= .003) and nearly significant between recommendations and joke telling (p= .052).

Figure 2. Plotted mean index of complex clauses by school grade and by type of text.



In sum, use of subordinate adverbials grows with age but continues to be very scarce in our corpus in every school grade. In 2nd and 6th grade, neither type of text favoured use of adverbials in a particular way. Only by 10th grade, explanations attracted more production of adverbials than both recommendations

and joke telling.

6.4.4 Relation between tasks: Correlations

The raw Pearson correlations across lexical measurements used in a previous study (vocabulary size, text length, word length, use of nominalizations, use of adjectives and level of text formality) and the syntactic measures tested in the present study are shown for each school grade and type of text separately in Tables 2,3 and 4 . For each grade level we found different patterns of correlations.

Table 2. Correlations among all lexical and syntactic experimental variables by type of text in 2nd grade.

		Vocab. Size	Text length	Word length	Nominalizations	Adjetives	Formality
2n							
	joke telling						
	NP complexity	-0,04	0,086	-0,234	-0,162	-0,105	-0,102
	Clause complexity	-0,025	-0,01	-0,114	-0,076	0,04	-,373**
	Clause length	0,2	-0,061	-0,033	-0,026	0,126	,414**
	Recommendation						
	NP complexity	-0,083	0,157	,350*	0,047	-0,02	,409**
	Clause complexity	-0,112	0,172	0,052	0,037	0,191	-0,163
	Clause length	-,351**	-,322*	-0,239	0,252	-0,041	,358**
	Explanation						
	NP complexity	-0,078	0,03	-0,109	0,037	-0,117	0,001
	Clause complexity	-0,025	0,11	0,223	0,127	-0,009	-,323*
	Clause length	-0,101	0,23	0,021	-0,094	-0,022	-0,04

In 2nd grade, level of lexical formality in the text was the lexical measure that showed a larger number of significant correlations with syntactic measures, particularly at the clause level, in the three types of text. The other lexical measures (text length, vocabulary size and word length) only correlated with syntactic measures in recommendation. Recommendation was the type of text showing a larger number of significant correlations between lexical and syntactic

measures. It was followed by joke telling and, finally, by explanation, which showed only one significant correlation between level of formality and clause complexity.

Table 3. Correlations among all lexical and syntactic experimental variables by type of text in 6th grade.

6th	Vocab. Size	Text length	Word length	Nominalizations	Adjectives	Formality
Joketelling						
NP complexity	,204	-,172	-,141	0,341**	-,232	-,029
Clause complexity	,026*	0,267*	,118	-,054	,068	,090
Clause length	,017	,003	0,358**	,076	,098	0,350**
Recommendation						
NP complexity	-,102	,112	-,094	0,301*	-,127	-,005
Clause complexity	,085	,061	,069	-,120	-,135	-,124
Clause length	-,248	,253	,010	0,358**	,044	0,655**
Explanation						
NP complexity	-,091	-,062	-,091	-,014	-,006	-,022
Clause complexity	-,217	,158	-,075	-,068	,081	,013
Clause length	,018	,066	,035	,041	0,254*	-,118

In 6th grade there is greater dispersion of the lexical measures that correlate with syntactic measures. Use of nominalizations and level of formality are the lexical variables that show more correlations with noun phrase complexity and clause length in both joke telling and recommendation. Joke telling is the type of text showing more correlations: vocabulary size, text length and word length correlate with clause complexity and clause length in this text. Explanation showed only one significant correlation between lexical (use of adjectives) and syntactic (clause length) variables.

Table 4. Correlations among all lexical and syntactic experimental variables by type of text in 10th grade.

10th	Vocab. Size	Text length	Word length	Nominalizations	Adjectives	Formality
joke telling						
NP complexity	,094	-,078	,132	,074	-,075	,157
Clause complexity	,094	,006	0,452**	,129	0,287*	,038
Clause length	-,023	,114	,101	-,056	,174	,253
recommendation						
NP complexity	,135	-,141	,118	,001	-,071	,099
Clause complexity	,131	-,070	-,036	-,036	,051	,104
Clause length	-,169	,216	0,279*	0,446**	0,284*	0,455**
Explanation						
NP complexity	,157	-,003	-,023	-,100	,067	-,035
Clause complexity	,007	-,057	,041	,036	-,083	-,214
Clause length	-,214	,111	,009	,027	,125	,170

Finally in 10th grade there is an overall decrease in the number of significant correlations across the board. Recommendation is the type of text showing more correlations between lexical (word length, use of nominalizations and adjectives and level of formality) and syntactic (clause length) variables. Word length and use of adjectives also correlated with syntactic (clause complexity) variables in joke telling. We did no find any correlation between lexical and syntactic variables in the explanations.

6.4.5 Relations between lexical command and syntactic complexity: regression analyses

Our next analyses focused on examining to what extent lexical command attained developmentally throughout compulsory schooling explained attainment of increased syntactic complexity in each school grade and type of text. We ran stepwise multiple regressions for each school grade using the lexical variables (text length, vocabulary size and vocabulary (a composite of word length, use of nominalizations, use of adjectives and level of text formality)) as the independent

variables and the syntactic variables as the dependent variables (Tables 6, 7 and 8 in annex 1). Table 5 presents an overview of the relations that emerged from this analysis. Asterisks indicate the significant relations.

Table 5. An overview of the relations between lexical and syntactic features

		NP complexity			Clause complexity			Clause length		
		2nd	6th	10th	2nd	6th	10th	2nd	6th	10th
joke telling										
	Text length					*				
	breadth							*		
	vocabulary		*				*	*	*	
Recommendations										
	Text length							*		
	breadth									
	vocabulary	*	*					*	*	*
Explanations										
	Text length									
	breadth			*						
	vocabulary								*	

Out of the three independent variables, vocabulary is the best predictor of syntactic complexity. In 2nd grade it predicted clause length in joke-telling $F(6,60)=4,320$, $p=.001$ (33%), and clause length $F(6,60)=5,732$, $p<.001$ (39%) and noun phrase complexity F(6,60)=2,325 , p=.046 in recommendations. In 6th grade it predicted noun phrase complexity in joke telling $F(6,180)=2,04$, $p=.046$ (19%) and recommendation $F(6,60)=2,281$, $p=.048$ (21%) and clause length in joke telling $F(6,60)=2,664$, $p=.025$ (23%), recommendation $F(6,60)=8,705$, $p<.001$ (50%) and explanation $F(6,60)=4,818$, $p<.001$ (35%). In 10th grade it predicted clause length in recommendation $F(6,60)=3,668$, $p=.004$ (29%). Text length and vocabulary size performed very poorly as predictors of syntactic complexity. Text length explained a (7%, p=.039) of the variance in clause complexity in joke-telling texts produced by 6th graders and a (9%, p= .012) of the variance in clause length in the recommendations written by 2nd graders

Out of the three dependent variables used in this study, clause length was the best explained one, in all grades but most particularly in 2nd and 6th grades. In contrast neither noun phrase complexity nor clause complexity was substantially explained by any of the independent variables used in this study.

In 6th grade, vocabulary predicted syntactic complexity at the NP and clause (length) levels in all three types of texts. The vocabulary-syntax relationship was less clear in 2nd grade and even weaker in 10th grade.

6.5 Discussion

Based on a corpus of written texts produced by Catalan school children and adolescents (CesCa), we have analysed the development of text-embedded syntactic complexity in two different sites: NP's and clauses, as produced in three different types of texts (an explanation of a film, a recommendations of such film and a joke telling) written in Catalan by 2nd ,(7 years) 6th (11 years) and 10th (15 years) graders. In previous research the corpus was characterized for text-embedded lexical uses. We deliberately separated the analysis of both domains of linguistic expression and in this study we have analyzed the relationship between the syntactic and the lexical uses in the texts. The study offers three main findings: first, Catalan school children and adolescents produce increasingly longer texts and more complex syntactic structures when facing the task of writing texts for different purposes. Development in both sites was more marked in late adolescence. Second, text-embedded uses of syntactic complexity interact with genre both at the noun phrase level and, more moderately, at the clause level. Explanation of a film favoured increase of complexity at both analyzed sites. Third, the results point to significant correlations between lexical and syntactic uses in Catalan, in a similar line with previous findings for English and Hebrew (Berman & Nir, 2009). In the following, we elaborate on each of these findings and discuss a number of psycholinguistic and educational implications.

First, our results show that, with age, children produce longer texts. This increase in text length is due both to the child's using more clauses and to his writing longer clauses. These two measures, however, follow different growth pathways: There is a sustained increase in number of clauses whereas increase in

clause length is located between 2nd and 6th grade but not between 6th and 10th grade. In fact, participants progressed from using simple short clauses in 2nd to producing longer (although not necessarily complex) clauses in 6th grade. In contrast, the differences between 6th and 10th grade do not result in a larger amount of words per clause but rather in a more complex organization of the words within the clause. In other words, a clause containing coordinated noun phrases could be as long as, but less complex than another clause containing a complex noun phrase. Although the number of elements is a suitable cue toward complexity, it not just a matter of number of elements. More information is needed for obtaining a more complete picture of the pathways towards complexity. This study is corpus-based, that is we have applied a number of measures that proved to be useful in other studies in order to establish level of complexity and relation between syntactic and lexical complexity. It would be worthwhile to complete such approach in future research by undertaking a corpus driven analysis, that is to examine the range (and characterization) of syntactic patterns underlying clauses of similar length.

Text length is affected by type of text, it increases in both recommendations and more so in explanations but not in joke telling, neither in number of clauses nor in clause length. This pattern is consistent with previous research showing high levels of productivity from early on most likely due to the fact that joke telling involves text reproduction rather than text production (Llaurado & Tolchinky, 2012). However, 2nd graders' joke telling texts were twice (or more) as long as their other types of text, and so were the clauses in this type of text. The fact that a joke is a brief piece of text, keeping many of spoken language features, may contribute to its being easy to remember (and reproduce) for a young child. Unlike other types of text that place higher cognitive demands on the child, writing jokes appears to serve as a platform for the child to expand on his writing. Further research is needed in order to confirm whether there is any relation between productivity in the written jokes and developmental aspects such as complex syntax or lexical growth in other more demanding types of text.

We analyzed the increase in complexity at the noun phrase level (measured in terms of prepositional phrases and/or relative clauses complementation) of

noun modification) as an indicator of developmental syntactic complexity. In line with previous research, in English and Hebrew (Ravid & Berman, 2010) our results show that use of increasingly complex noun phrase constructions grows with age in both recommendations and explanations, but more so in explanations and recommendations of a film, which, compared to jokes, emerge as the two favoured loci for use of complex sophisticated language. Moreover, and also consistent with previous research, our results show that growth in noun phrase complexity experiences a protracted, slow paced development. Thus, between 2nd and 6th graders the index of noun phrase complexity did not differ significantly. Only 10th graders produce significantly more complex noun phrases, than 2nd graders. Lack of complexity in the noun phrases produced by younger children may be due to lack of cognitive maturation as well as to limited world knowledge base. Thus, younger children's explanations habitually referred to a specific person or character just as their recommendations referred to a film or series denoted by its proper name or they denote very general entities or concepts. Therefore, they filled the corresponding noun phrases with the corresponding general nouns, perhaps accompanied by basic adjectives or coordinated with other similarly structured noun phrases. Only little by little, the characters or series became more precisely defined entities or concepts denoted by the (general) nouns but specified or narrowed down by means of prepositional phrases or relative clauses. In 2nd and 6th grade the three types of text do not yield differences in terms of noun phrase complexity. In 10th grade, in contrast, noun phrases in explanations and recommendations are more complex than noun phrases in joke telling.

In line with the development of syntactic complexity at the noun phrase locus, the development of syntactic complexity at the clause level (measured by the use of subordinate adverbials) also experiences a late gradual development. There is no significant increase between 2nd and 6th grade and a very marginal growth between 6th and 10th grade. The clauses produced by 10th graders, however, were significantly more complex than the clauses written by their 2nd grade peers. The text-embedded use of adverbial subordinates is driven by discourse pragmatics to indicate the temporal relationship between two or more events, or to express a condition for or a result of the realization of the main clause event, or to express a reason for the proposition expressed in the main clause. Therefore, it entails some

late emerging ability to use linguistic devices for the purpose of organizing the flow of information in a hierarchical structure. In writing, besides, it shows some level of text planning, a component of text writing more habitual at the upper grades.

Overall, increasing syntactic complexity is a feature of school-based language development. Such increase in complexity is driven by discourse, as a consequence of children's improved ability to fine tune their texts to their communicative needs according to genre specific requirements therefore. In this context, schoolers' use of complex syntax takes different pathways depending on the type of text children are writing. However, this is a lengthy process taking all compulsory school to develop.

The fact that 10th grade emerged as a cut-off point in syntactic architecture is in line with findings of other studies in which adolescence has shown to be a turning point for text-embedded language use in a variety of domains including syntax but also lexical density, diversity, and register in English, and Hebrew, and devices for downgrading agency in English, French, and Spanish. These findings underscore the close interconnection between linguistic and social and cognitive maturation. Development is critically revealed by local linguistic expression, with the lexicon and syntax going hand in hand in this connection.

In previous research, the text-embedded lexical uses in the same corpus have been documented (Llaurado & Tolchinsky, 2012). We found that text length, size and quality of vocabulary were developmentally diagnostic and discriminated by type of text (level of text formality (*F*-measure) was also found to be a good discriminator by type of text. The three grades we have picked for examining uses of complex syntax were found to be moments of marked lexical growth but they differ with respect to the development of vocabulary preceding them. We found significant correlations between measures of lexical and syntactic use in all grades, although more in 6th than in 2nd and 10th grade.

Level of text formality yielded the largest number of significant correlations followed by word length and nominalizations. Given that nominalizations commonly wrap semantic abstractness in Catalan, our present results would be consistent with (Berman & Nir, (2009) who found a significant correlation between semantic abstractness and complex syntax in both Hebrew. Lexical and

syntactic measures correlated in jokes and recommendations in all three grades and in explanations in 2nd and 6th grade. Lexical measures correlated with measures of syntactic complexity at the clause level in all three grades and at the noun phrase level in 2nd (in recommendations) and 6th (in jokes and recommendations) grade. And vocabulary measures had explanatory power at the clause level in all three grades and also at the noun phrase level in 6th grade. In fact, 6th grade, in which lexical growth peaks after sustained (lexical) growth in the three preceding grades yielded the largest number of correlations affecting more sites of syntactic complexity. In contrast, in 2nd grade, where the lexical burst could in fact be the result of improved command of transcription mechanisms, and in 10th grade after overall stagnation of text-embedded lexical growth, the number of correlations found was lower, and no correlations were found between lexical uses and syntactic complexity at the noun phrase level.

Notwithstanding the significant correlations between lexical and syntactic uses, our results show clear differences in the development of each linguistic component. Thus, while 6th grade turned out to be a point of major lexical growth, it is not until 10th grade that we find a cut-off point for the growth in the use of complex syntax. In our view this reinforces the idea that developmental language use must be examined taking into account the interaction between all the domains participating of this development. In future research we aim at exploring the connection between spelling abilities (which have been shown to rely on knowledge of different linguistic domains) and syntax.

Finally, we think this study provides additional support to the view contended by several researchers (Berman, 2008; Ravid & Zilberbuch, 2003, Tolchinsky & Rosado, 2005) that language use must be assessed in text-embedded contexts since they are the most appropriate setting for capturing the richness and diversity of the syntactic constructions deployed by (speaker) writer in their attempts to fulfil different communicative goals. Also we believe this study supports the idea that corpus linguistics provides a powerful tool for gathering and analyzing authentic language data focusing on the different aspects/domains of language involved in these productions.

CHAPTER 7

The developmental pattern of spelling in Catalan from 1st to 5th school grade

---

Anna Llaurado and Liliana Tolchinsky

Universitat de Barcelona

Submitted to Writing Systems Research

Abstract: Orthographies no only represent the phonology of a language but also aspects of morphology, syntax, and the lexicon. Learning to spell in a particular language involves understanding the relation between the graphic elements of an orthographic system and the levels of language it represents. The goal of this study was to track the developmental path to orthographic spelling in native speakers of Catalan. Typologically, Catalan is a synthetic inflectional language with a rich inflectional and derivational morphology that has a moderately transparent orthography. In most cases, strictly phonetic to written mapping renders incorrect spelling and spellers had to resort to morphology, word-contextual rules or to lexical knowledge to spell rightly. We analyse a corpus of written vocabularies from different semantic fields, that prime different lexical categories and word frequency, produced by 225 native speakers of Catalan from 1st through 5th school grade. The productions were characterized in terms of spelling (in)accuracy on the basis of whether phonographic, morphologic, word-contextual or lexical

knowledge was required to render the orthographically correct form. Results show that phonographic, morphological and orthographic errors decreased with school level relative to lexical errors. More errors occurred at the word stem than at the word affix level suggesting a role of morphological awareness in spelling. Some linguistic and educational implications of these findings are discussed.

Key words: spelling, later language development, morphology, type of misspelling

7.1 Introduction

Spelling is more than just an academic requirement. Learning to spell involves understanding the relation of the graphic elements to the different levels of language: phonology, morphology, syntax, and the lexicon. If a word contains a spelling mistake **baca* for *vaca* 'cow', one may manage to read it correctly. But in order to spell a word correctly one needs a complete orthographic representation of it; moreover one must grasp the nature of the particular orthography of the language in use. In some orthographic systems most graphic signs (letter or graphemes) have only one reading and one way of spelling irrespective of the word they are part of whereas in other orthographies one letter may have many different readings or one category of sounds many different spellings. Orthographies with a high degree of consistency (between letters and sounds) are considered shallower or more transparent whereas orthographies with a low degree of consistency are considered deeper or more opaque (Frost, 1992). Because it is all a matter of degree of consistency, orthographic systems lay on a continuum of transparency. At one extreme of the continuum we find orthographies with a high level of consistency such as Finnish (nearly 100% of the letters have only one reading and nearly 90% only one spelling) whereas at the other extreme orthographies such as French (75% of the letters have only one reading and 50% only one spelling). Catalan orthography, the one we are concerned with in this study, lays somewhere in the middle (70% (40% in the case of vowels) of the letters have only one reading and 76% only one spelling). It is less transparent than orthographies such as Finnish or Spanish but not as opaque as orthographies of French. The degree of transparency of an orthography posits

different problems for learning how to spell (Seymour, Aro & Erskine, 2003). To illustrate, strict assignation of letter to sound correspondences might render several orthographically inaccurate forms of the word menjar /mənʒa/ 'to eat' . In order to avoid the phonographically plausible <*manja> the child will need to use morphological knowledge of the infinitive suffix –ar /a/, as well as lexical knowledge of the stem –menj /mənʒ/. In other words, orthographies not only encode the phonological structure of words but also their morphological structure and their morphological relationship to other words (see: Baayen & Schreuder, 2003; Feldman, 1995), the deeper the orthography, the more morphological information it encodes. In addition to phonology and morphology the correct spelling of a word obeys a number of context-dependent rules that determine the legality or illegality of a particular string of letters and the use of a particular letter. And there are cases where the child needs to rely on lexical knowledge of the word form in order to spell that word correctly. Thus, for learning how to spell children need to orchestrate information from different levels of language.

Research has shown that word frequency also has an important impact on children's development of spelling competence (Alegria & Mousty, 1996, Leté, Peeremean & Fayol, 2008). Several studies have shown that repeated decoding of new words does generate orthographic representations of them (Cunningham, Perry, Stanovich & Share, 2002; de Jong & Share, 2007; Reitsma, 1983, 1983; Share & Shalev, 2004). In a different vein, the self-teaching hypothesis (Share, 1995, 1999, 2004) states that each successful decoding of a word increases the probability that its orthographic representation will be stored. Therefore, the primary via of exposure to words may also have an effect on spelling: an orthographic representation of words acquired primarily through reading could be stored faster than one of words primarily encountered through spoken interaction. In the framework of the self-teaching hypothesis, since success in word decoding is attained faster in a transparent orthography, storage of an orthographic representation of words must develop faster too (Carrillo, Alegria & Marin, in press).

In this paper we focus on the developmental pattern of learning to spell in Catalan. Specifically, we will analyze the role of phonological, orthographic, lexical and morphological information in native Catalan school children's spelling

decisions. With this aim, we will track children ranging from 5 to 11 years spelling isolate words from 5 different semantic fields.

In what follows we summarize the basic characteristics of the Catalan orthographic system and how it represents the Catalan phonology, and we will also refer to some of the basic principles governing the orthographic representation of the Catalan morphology. The considerations below refer to the Central variant of the Catalan language, the variant that serves as the standard and whose use is most widespread at school.

7.1.1 Selected features of the Catalan orthographic system

7.1.1.1 How the orthographic system represents Catalan phonology

The Catalan alphabet consists of a set of 27 letters (5 vowels and 22 consonants) which represent 33 phonemes (8 vowels and 25 consonants). As in most alphabetic orthographies the representation of vowels is far less consistent than the representation of consonants (Tolchinsky & Salas, 2012).

The 5 vowel letters [a] [e] [i] [o] [u], expanded to a set of 9 by the addition of diacritics: [à] [é] [è] [í] [ï][ó] [ò] [ú] [ü], serves to represent 8 vowel phonemes (Spanish consists of 5, French of 15).

In a stressed position, the phoneme to letter correspondences involve a many to one relationship:

(1) /a/ → [a] [à]
(2) /e/ → [e] [é]
(3) /ɛ/ → [e] [è]
(4) /i/ → [i] [í] [ï]
(5) /o/ → [o] [ó]
(6) /ɔ/ → [o] [ò]
(7) /i/ → [i] [í]
(8) /u/ → [u] [ú]

Decisions concerning the choice of the letter with which to represent a vocalic sound require using knowledge not only of phoneme-grapheme correspondences but also of the orthographic rules governing the use of diacritic marks. For instance, although both *reso* /'rezu/ 'I pray' and *resaré* /rezə're/ 'I will

pray' share the phoneme /e/, orthographic rules determine that *reso* does not need a diacritic whereas *resaré* does.

Only the phoneme /ə/ cannot appear in a stressed position and, together with /i/ and /u/, both of which can appear either in stressed or unstressed position, conforms the unstressed vocalic system. This reduced subsystem implies a many to one relationship from sound to letter:

(9) /ə/ → [a] [e]

(10) /u/ → [o] [u]

The choice of letter for the unstressed vocalic system requires lexical knowledge, that is rote knowledge of the orthographic form of the lexical item involved, in order to be able to render correct spelling.

As in other languages, i.e. Spanish, German, in addition to the 8 possible monothongs, there are 24 possible diphthongs in Catalan involving the combination of either a vowel and a semiconsonant or the combination of a semiconsonant (represented graphically by a vowel letter) and a vowel.

A set of 22 consonant letters serves to represent the 22 consonant, plus 2 semiconsonant, phonemes (Spanish consists of 17, French of 18 and English of 22): [b], [c] [ç] [d] [f] [g] [h] [j] [k] [l] [m] [n] [p] [q] [r] [s] [t] [v] [w] [x] [y] [z], and 10 digraphs [gu] [ig] [ix] [ll] [ny] [qu] [rr] [ss] [tg] [tj] [tx]. Fourteen consonants are mapped to only one letter whereas 11 consonant phonemes can be spelled by more than one letter or digraph, as follows:

(11) /s/ can be spelled [c] (cel)

(12) [s] (serra), [ç] (peça), and [ss] (passeig)

(13) /z/ can be spelled [s] (casa), [z] (pinzell)

(14) /k/ can be spelled [c] (casa)

(15) [q] (quadre), [qu] (quiso), and [g] (amarg)

(16) /b/ can be spelled [v] (votar) and [b] (botar, branca)

(17) /g/ can be spelled [g] (galta) and [gu] (guerra)

(18) /t/ can be spelled [t] (edat) and [d] (solitud)

(19) / ʒ / can be spelled [g](gel) and [j] (jugar)

(20) /ʃ/ can be spelled [ix] (calaix) and [x] (xocolata)

(21) /r/ can be spelled [r] (roda) and [rr] (carretera)

(22) /t□□/ can be spelled [tx] (txec) and [ig] (boig)

(23) /d□□/ can be spelled [tj] (platja) and [tg] (metge).

7.1.1.2 Role of context dependence rules

Spelling decisions must obey context dependent rules concerning either word position or letter combinations. For instance, the bigram [br], both initially and intraword, allows only [b]. Thus, (16) <*branca*> 'branch' not <*vranca> ; in (20), [ix] must be used to spell /ʃ/ both between vowels within the word and in word final position, e.g, *caixa* 'box', *calaix* 'drawer' whereas [x] must be used in word initial position *xocotala* 'chocolate' or after the consonant within the word *carxofa* 'artichoke'. A number of cases remain, however in which context dependence rules allow for more than one legal spelling, for instance, in *enciam* /ənsjam/ 'lettuce' /s/ could be spelled by both [s] or [c] and lexical knowledge is necessary in order to make the correct choice.

7.1.1.3 How the system relates to Catalan morphology

Learning to spell in Catalan implies the acquisition of many regularities but also dealing with a fair number of inconsistencies not only in the realms of phoneme to grapheme correspondences and orthographic rules but also in the way orthography represents morphology.

Catalan is a synthetic flexional language with a rich inflectional and derivational morphology. The orthographic rendering of inflectional and derivational suffixes is a very common source of spelling errors because of the inconsistency between their spoken and written expression. In particular, *the orthographic rendering of inflection in nouns (24), adjectives (25) and determinants (26), entails letter change with no phonological correlation,*

(24) *Cama* /kama/ 'leg –fem-sg' → *cames* /kamas/ 'legs –fem-pl'

(25) *Honesta* /unesta/ 'honest –fem-sg' → *honestes* /unestas/ 'honests –fem-pl'

(26) *La* /la/ 'the–fem-sg' → *les* /las/ 'the–fem-pl'

in verb infinitives (27) and gerunds (28) and in derivative suffixes for word formation (29 requires mapping of a written letter onto a phonological empty segment (a segment that is not pronounced)

(27) *jugar* /juga/ 'to play'

(28) *jugant* /jugan/ 'playing'

(29) *banya(dor)* /banado/ 'swim suit'

Some derivative suffixes present phonographic inconsistencies (30):

(30) *Recoman(able)* /rakumanapla/ 'recomendable'

In cases such as (30), it is worth noting that transcribing [recomanaple] would render a legal, though not normative, word form according to both the phoneme grapheme correspondence rules and the context dependency rules. Thus, both in cases of inflection and derivation, recognition of the morphological status of the segment involved is helpful for attaining conventional spelling without resorting to lexical, word per word, knowledge.

7.2 Goals and predictions

It has been shown that sublexical procedures of spelling are acquired at a faster rate in regular than in irregular and intermediate orthographies (Caravolas & Bruck, 1993; Wimmer & Landerl, 1997). Hence, the development of spelling in a transparent orthography such as Spanish takes less time than in a more opaque language such as English or French (Defior, 2005; Marin, Carrillo & Alegria, 1999). In addition, some studies have claimed that language-specific typology can also affect the rate and the pattern of the development of orthographic spelling (Ravid, 2001). Indeed in some cases, the morphological knowledge of the word, that is, the recognition of the word's morphemes, is useful as a path to conventional spelling, and far less costly than word by word storage in memory.

Against this frame, we therefore expect children to show the quickest progress in spelling phonographically consistent words. Also, given the richness of the Catalan derivational and inflectional morphology, we predict children to show the capacity to use morphological cues in their spelling of word affixes at an early stage. We also expect that Catalan speakers/writer to struggle with cases of phoneme grapheme inconsistency throughout grade school. However, we expect that inconsistencies requiring the use of context dependence rules to be solved earlier than spelling decisions based on purely lexical factors since lexical knowledge necessarily relies on sustained acquaintance with the printed language.

This study examines children's spelling of words from 5 different semantic fields (food, clothing, leisure activities, traits of personality and natural phenomena). Some fields, such as Food and Clothing, contain many high frequency

words children encounter primarily in oral interaction. Conversely, terms for Natural phenomena and, to some extent, traits of personality, are mostly encountered through print in textbooks. We will determine if children show different patterns when spelling high frequency words from everyday (spoken) language and low frequency terms from specialized (text-book) discourse. Since updated frequency dictionaries are not available in Catalan, we will consider a word's frequency such word's rate of occurrence within the corpus.

7.3 Method

7.3.1 Participants

This is a corpus based analysis of spelling development. The data consists of the written productions of 225 native Catalan speakers attending 1st to 5th school grade (5 to 11 years) in 32 different schools. These children were considered as native Catalan speakers on the basis of their declaring Catalan to bes their only home language, an information confirmed by their teachers.

7.3.2 Tasks and materials

Each participant was asked to write down vocabularies from five different semantic fields upon the following instruction *Escriu totes les paraules de menjars que puguis recordar* "write down as many food words as you can remember" as a prompt for producing the names of food items. This instruction was adapted in order to elicit names for Clothing, Leisure activities, Personality traits and Natural phenomena. Elicitation instructions were piloted at different school levels to guarantee the most comprehensible wording.

7.3.3 Procedure

Participants wrote by hand in order to avoid misspellings caused by lack of mastery of word processing software. The task was always carried out in the participants' habitual classroom as part of their everyday school activities and the instructions were given by their habitual language teacher.

7.3.3.1 Corpus transcription and digitalization

We developed a number of procedures to keep the original productions and

to prepare them for further processing. Original productions for each task were introduced in a relational database (in MySQL) that enables us to trace information related to each element of the text that serves to identify independent variables (school, sex, school level, home language/s, and length of time they have spoken Catalan).

7.3.3.2 Criteria of Analysis

All the words (tokens) written down, except those which were illegible, were counted and lemmatized, and assigned a grammatical category. As in any lexicographic study the goal of the lemmatization process was to define a canonical form that functions as a referent for a set of variants. However, given the nature of the corpus, variants were not just inflected forms but might also be orthographic and graphic variants. For example, the lemma *jersei* 'sweater' had the following variants associated:

(1) jerseis: plural variant

(2) JerSEI: graphic variant (graphic variants were not considered spelling mistakes here)

(3) xersell: orthographic variant.

7.3.4 Spelling error coding

Every error was manually coded for two different aspects: type of error and type of morpheme in which the error was produced. As types of errors we considered:

Phonographic errors

This category includes any error involving misuse of the sound to letter correspondences.

Specifically, we counted all instances of 1) omission of a sounded letter, e.g., **samarrta* for *samarret*a 't-shirt' 2) addition of a letter e.g., **amatble* for *amable* 'nice' 3) substitution of a letter by another one corresponding to a different sound, e.g., **baldilla* for *faldilla* 'skirt'4) inversion of letters either within or between syllables e.g., **alfuent* for *afluent* 'tributary'.

The number of phonographic errors is calculated as the proportion of

phonographic misspellings relative to the total number of words written by the participant.

Morphological errors

This category includes involve the non-use of morphological information. For instance, both in *atles* 'atles' and mapes 'maps' the final chunk –es represents /as/. However, only in mapes –es is it morphologically motivated as it corresponds to the plural form of *mapa* /mapa/ thus requiring the change from –a (singular) to –es (plural). While the child needs to access an orthographically correct representation of the word *atles* in order not to misspell it, he can use morphological knowledge to assist in his spelling of *mapes*.

The number of morphological errors is calculated as the proportion of morphological misspellings relative to the total number of morphologically affixed words written by the participant.

As for type of morpheme, for all the tokens that were inflected or derived words, we coded whether the error was produced in the word stem (****pasej****ar* for *passejar* 'to take a walk') or in the affix (**sabatas* for *sabates* 'shoes'; **amavle* for *amable* 'gentle)

The number of stem errors is calculated as the proportion of stem misspellings relative to the total number of words written by the participant. The number of affix errors is calculated as the proportion of affix misspellings relative to the total number of morphologically affixed words written by the participant.

Orthographic errors

This category includes errors involving the non-use of context -dependence rules. For instance, although /s/ can be transcribed as either [s] or [ss], the intraword context determines the choice of letter. Thus the use of [ss] in carabassó /kərəbəso/ 'zuchini' is determined by the intervocalic context for [s] represents /z/ in this context. The omission of diacritic marks, e.g., **cinturo* for *cinturó* 'belt', were considered as a non-use of orthographic rules.

The number of orthographic errors is calculated as the proportion of orthographic misspellings relative to the total number of words written by the participant.

Lexical errors

This category reflects the non-use of lexical knowledge as evidenced by the substitution of a vowel or a consonant by another phonologically legal but lexically (etymologically) inaccurate one. For instance, /b/ can be transcribed by either b, i.e., beure /bewra/ 'to drink' or v, i.e., veure /bewra/ 'to see'. Lexical knowledge is needed in order to render the correct orthographic representation.

The number of lexical errors is calculated as the proportion of orthographic misspellings relative to the total number of words written by the participant.

One token could contain more than one error. Each error was coded separately.

7.4 Results

This section consists of three parts. Firstly we provide a general description of the corpus in quantitative terms, specifically we present a breakdown by school grade and semantic field of the number of tokens, misspelled tokens and misspellings produced. Secondly we present a breakdown of spelling errors by types of error (phonological, morphologic, orthographic and lexical). Thirdly, we look at the distribution of misspellings in terms of whether they occur in the word stem or in a word affix.

A series of one way ANOVAs school grade (5) with repeated measures on type of error (4), and type of morpheme (stem or suffix) (2) were performed.

The eta squared value (η2) is used to report the effect size of both the main effect and the interactions. The size effects of relevant pairwise comparisons are reported using Cohen's *d*. An alpha level of .05 was used for all statistical tests. When the assumption of sphericity was found to be violated, degrees of freedom were corrected using Greenhouse_Geisser estimates.

7.4.1 General description of the corpus.

The 225 children produced a total of 21,210 tokens, of which 5,070 were types subsumed under 2,212 lemmas. A total of 7,347 words (tokens) contained one or more errors with the number of errors per word ranging from 1 to 6. The total number of errors was 10,253.

Table 1. Mean number of produced tokens (T)(SD), misspelled tokens (t)(SD) and misspellings (E)(SD) by school grade and by semantic field.

Grade **Semantic field**

	Clothing			Food			Leisure activities			Traits of personality			Natural phenomena		
	T	t	E	T	t	E	T	t	E	T	T	E	T	t	E
1st gr	17.1	10.3	14.6	21.0	10.8	16.9	21.1	11.8	18.5	13.7	6.1	8.6	13.4	8.6	13.9
SD	8.3	5.5	8.8	8.1	4.6	8.4	13.7	7.2	12.3	6.8	4.8	8.2	7.3	3.5	5.7
2ndgr	15.3	7.4	9.6	19.2	7.8	11.1	18.0	6.6	10.4	10.6	3.6	5.7	9.4	4.6	7.4
SD	6.8	3.9	5.4	5.8	3.1	5.6	12.4	4.2	7.3	4.9	2.9	4.8	6.5	1.7	2.9
3rdgr	14.8	5.8	7.1	22.3	7.5	9.9	13.5	4.6	6.4	7.4	3.1	4.5	7.8	3.8	5.2
SD	8.8	4.0	5.0	10.9	4.9	6.9	9.3	3.3	4.5	4.9	2.9	5.3	4.3	1.9	3.4
4thgr	21.7	6.2	7.1	31.5	8.8	9.9	28.7	6.5	6.4	16.8	4.5	4.5	12.6	4.1	5.2
SD	10.6	4.5	6.2	7.3	4.2	7.1	23.2	6.8	9.4	7.9	2.7	3.8	9.8	2.9	3.8
5thgr	25.4	6.9	8.2	37.8	9.7	12.2	35.3	6.1	7.7	19.1	4.4	5.6	15.1	4.3	5.3
SD	9.3	4.6	6.1	9.9	5.6	7.9	23.3	5.1	6.8	7.9	3.2	4.5	9.0	3.6	5.3

The mean number of tokens produced by each participant increased with school grade for the 5 semantic fields, the most pronounced growth occurring between 3rd and 5th grades, and most particularly so for Traits of personality (most words in this field were expressed by adjectives, a later developing grammatical category) (Tolchinsky, Marti & Llaurado, 2010). In contrast, the mean number of errors decreased throughout grade school, though more markedly between 1st and 3rd grades.

School grade had a significant impact on the number of errors made $F(4,225)= 52{,}838$ $\eta 2 = 0.49$ $p > .001$ and children improved their spelling accuracy throughout school grade. Bonferoni post-hoc analyses revealed significant differences between 1st and 2nd grade and between 2nd and 3rd grade, and

marginally significant differences between 3rd and 4th grade. The contrast between 4th and 5th grade was not significant.

We found a significant effect of semantic field on the number of errors made by children $F(4,225)=15,076$ $\eta 2=0.08$, $p>.001$. Thus, terms for Traits of personality accumulate the highest mean proportion of errors followed by Leisure activities, Clothing, Food and Natural phenomena, this being the semantic field accounting for the lowest mean proportion of errors. Bonferoni post-hoc analyses reveals significant differences only between the field of Traits of Personality and all the other semantic fields. These contrasts were significant throughout grade school ($0.28<d<2.26$).

The different semantic fields triggered words differing markedly in terms of frequency, a dimension with a relevant impact on spelling accuracy. For instance, Food contains many high frequency noun words common in everyday spoken conversation whereas Natural Phenomena contains also noun words but less frequent more written-like, subject specific words. In lack of updated sources of word frequency in Catalan, we calculated the frequency of each word within the corpus by semantic field in order to tap any possible effect of word frequency on the misspelling rate.

Table 2. Distribution of total number of tokens by semantic field, proportion of tokens with a frequency of occurrence over 10 and proportion of tokens with a frequency of occurrence equal to 1.

	Clothing	Food	Leisure activities	Traits of personality	Natural phenomena
Tokens	4531	6275	5462	2719	3144
Frequency >10	26%	22%	13%	10%	13%
Frequency =1	37%	40%	46%	53%	50%

Traits of personality concentrates the highest mean number of errors per child, shows the highest proportion of words occurring just once in the corpus and the

lowest proportion of words occurring more than 10 times in the corpus. It is worth noting that in this field the growth of tokens per child is triggered from 3rd grade on, which is later than in the other fields. Leisure activities shows a distribution of word frequencies in the corpus that is similar to Traits of personality and yields the second highest mean number of errors per child. The higher the proportion of words occurring over 10 times and the decrease of words with a frequency of occurrence of 1 the lower the mean number of errors per child. However, there is one exception: the Natural phenomena semantic field yielded the lowest mean number of errors per child in spite of showing the second highest proportion of words occurring just once in the corpus and the second lowest proportion of words with a frequency of occurrence over 10. We had hypothesized that this field would include school related terms acquired mostly through the reading of textbooks. It might be that this main via of exposure has some effect on children's spelling accuracy. However, tracking over the upper grades will be needed in order to fully confirm this hypothesis.

7.4.2 General developmental pattern of spelling.

The number of errors of all types decreased markedly between 1st and 5th grade, even though by 5th grade, 23% of the words produced were misspelled. Phonographic errors accounted for the lowest proportion of misspelling from 1st to 4th grade. Orthographic and lexical errors accounted for the largest proportion of misspellings. Lexical errors surpassed orthographic misspellings by far between 1st and 3rd grade, but they represented roughly the same proportion from 3rd to 5th grade, following a marked decrease of lexical errors. Morphological errors underwent a dramatic change from 1st to 2nd grade. Thus, while it was the most frequent type of error in 1st grade, it descended below lexical and orthographic errors in 2nd grade, and it is the least frequent type of error in 4th and 5th grade.

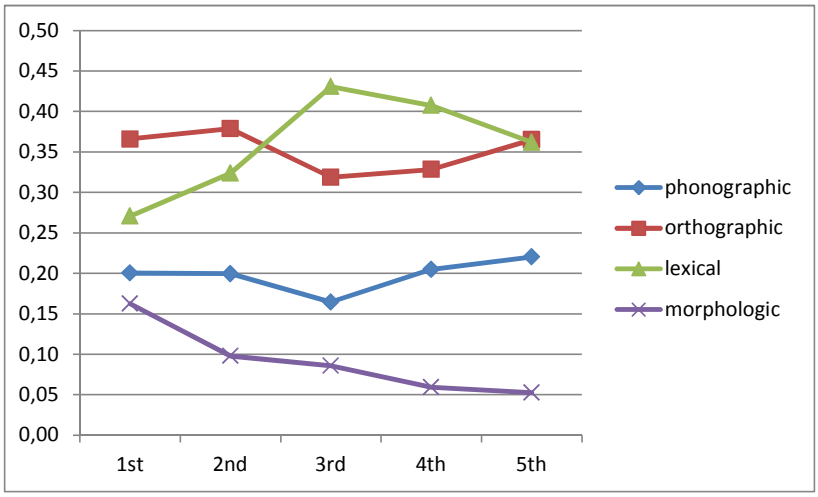
7.4.2.1 Developmental pattern of spelling by type of error

In order to track the developmental changes in the distributional pattern of errors of each type throughout schooling, we calculated the proportion of the number of errors of each type relative to the total number of errors made. An ANOVA with repeated measures by type of error showed a significant impact of

both school grade F(4 ,225)=18,4187 p< .001 η2=0.27 and type of error F(2,076,225)=1057,410 p< .001 η2=0.84 as well as a significant interaction between them F(8,304,225)=7871 p< .001 η2=0.13. Phonographic and orthographic errors showed a significant decrease between 1st and 3rd grade, and a slight not significant recovery thereafter. Morphological misspellings are the only type of error that showed a steady decrease steadily throughout grade school. Lexical errors showed a significant increase between 1st and 3rd grade and a not significant decrease thereafter.

In 1st grade we found significant differences between all types of misspellings. In 2nd, 3rd, 4th and 5th grades phonographic and morphological errors did not differed significantly from each other and both differed significantly from orthographic and lexical errors.

Figure 1. Mean proportion of errors by school grade and by type of error



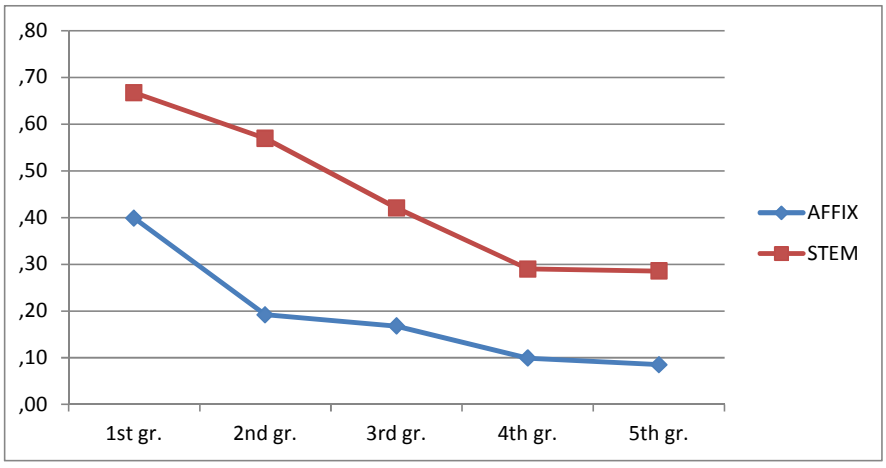
These results confirmed our predictions. We had expected a sustained decrease in phonographic and orthographic errors given that they are solvable by rule-based knowledge and that both phonographic correspondences and context dependence rules are a primary focus for school writing instruction and practice. This was the case and the proportion of both these types of errors decreased throughout grades.

We also expected a sustained decrease in morphological errors as an effect of the salient morphology in Catalan, which would compensate whatever lack of instruction might exist in classrooms on this type of analysis. This prediction was also confirmed by the results. Morphological errors represented and successively lower proportion of the total number of misspellings and this developmental trend yielded significant contrasts between 1st, 3rd and 5th grade. Finally, we had hypothesized that lexical errors would represent an increasingly higher proportion of the total errors made since lexical knowledge is accrued only on the basis of a long term acquaintance with written texts. The results confirm this prediction (a slight decrease between 3rd and 5th grade was not significant).

7.4.2.2 Developmental patterns of spelling different word morphemes.

A total of 9,349 of the words produced by children showed morphological affixes. The distribution of the 5,293 spelling errors occurring in these words was: 3,848 occurred in the word stem (spelling the stem relies on lexical knowledge) and 1,445 occurred in the affixed morphemes (spelling an affix relies on a morphological analysis of the word).

Figure 2. Mean proportion of misspellings by school grade and type of word morpheme



An ANOVA with repeated measures by type of morpheme on the subsample of tokens showing flexion or/and derivation marks revealed a significant impact of school grade $F(4,225)=56,980$ $p< .001$ $\eta 2=0.51$ with the number of errors decreasing more sharply in affixes than in stems. Bonferoni posthoc analyses showed significant contrasts between 1st and 3rd grade for stem misspelling and between 1st, 2nd and 4th grade for suffix misspelling.

The effect of type of morpheme was also significant $F(2,225)=591,117$ $p < .001$ $\eta 2=0.73$. The number of errors was significantly lower for affixes than for stems and the effect size of the contrast between the two types of morpheme was large ($d> 7.97$) for all school grades.

Two examples illustrate the different resources available to a child for solving apparently similar problems depending on whether he is spelling a stem or an affix. Apparently, a child spelling the inflected form /llantias/ llenties 'lentils-pl' faces two equivalent difficulties of phoneme-grapheme inconsistency /a/ spelled [e], both in the stem and the suffix. Similarly, a child spelling /iberna/ hivernar ''to hibernate' faces the task of representing a phonologically empty letter both in the stem, for letter [h] and in the suffix, for letter [r]. In order for children to either solve the phoneme grapheme inconsistency in the stem of *llenties* or to produce the phonologically empty [h] in the stem of *hivernar*, they need to rely on lexical knowledge of the words. However, the identification of the flectional suffix in *llenties* and in *hivernar*, provides the children with a helpful basis on which to produce the necessary graphic transcription.

7.5 Discussion

We have tracked the developmental path of spelling of native Catalan gradeschoolers from grades 1st to 5th in a written corpus of vocabularies. Particularly, we have examined the existence of possible differential learning/spelling patterns depending on whether phonological, orthographic, lexical or morphological analysis of the word can be applied. The study offers four

main findings: first, the number of misspellings decreases notably between 1st and 5th grade, though spelling competence is far from ceiling level by 5th grade. Second, results appear to suggest some degree of influence of frequency/grammatical category on spelling accuracy. Third, the developmental path of spelling differs depending on the type of linguistic knowledge the writer needs in order to produce a correct spelling. However, our results suggest that the developmental pattern of each type of spelling error is neither linear nor autonomous but rather affected to some degree by the developmental pattern of the other spelling errors at that same stage. Finally, and somewhat related to the former, spelling a word stem, being based on lexical knowledge, presents more difficulties, , than spelling a word affix, a task in which the child can rely on the morphological analysis of the word. In the following, we elaborate on each of these findings and discuss a number of psycholinguistic and educational implications.

Firstly, regarding the improvement in spelling accuracy, children in 1st grade made spelling errors in 55% of the words written whereas the proportion fell to 23% in 5th grade. Further research of the corpus will reveal whether improvement is sustained throughout compulsory schooling or whether there is stagnation in performance. Previous research using a larger sample of children and adolescents with diverse home languages has shown spelling accuracy to improve very moderately up to 3rd year of secondary school and then more markedly in the 4th (last year) of secondary school (Tolchinsky et al., 2010). In this study we are examining spelling abilities shown by a native Catalan sample of participants, that is children who extend their use of Catalan to out-of-school activities and to interactions with their families and friends. Therefore they have a better knowledge of the language in every language subsystem from phonology to syntax, and from the lexicon to pragmatics.

With regard to possible differences by semantic field on spelling accuracy, our results show that children made more spelling errors when writing terms for Traits of personality than when writing terms for Leisure activities, Food, Clothing and Natural phenomena. Two reasons explain this finding: frequency of use and the morphological category of the tokens used in each semantic field. The semantic field Natural phenomena contains the lowest number of tokens produced and yet yields the (significantly) highest number of spelling errors. This may be accounted

for by the distribution of word frequency in this semantic field, since it contains the lowest proportion of words with a high frequency of occurrence and the highest proportion of words with a low frequency of occurrence (relative to the corpus).

Moreover, the Traits of personality semantic field triggers the production of adjectives, a grammatical category that has been used as an indicator of later lexical development as shown by its moderate presence in the written productions of children ranging from 1st to 4th grade, and a more relevant occurrence from then on (Llaurado & Tolchinky 2012; Ravid & Levie, 2010). This later use of adjectives has shown increasing levels of morphological complexity. Such characteristics might contribute to explaining why spelling Traits of personality, i.e. (complex) adjectives, is more difficult for school children.

The relation between the proportion of errors and the distribution of word frequency is maintained for Leisure activities, Food and Clothing. However, neither size nor word frequency appear to have an effect on the error rate in the terms naming Natural phenomena. In spite of its small size, of a low proportion of high frequency words and of a high proportion of low frequency words it yields the lowest mean of number of misspellings. An explanation for this mismatch could lay in the fact that most of the terms for Natural phenomena are acquired through exposure to textbooks. This requires further research, however, since our expectation that this semantic field would include an important number of advanced, high register specialized terms was only partially fulfilled. Thus, young children produced rather common terms such as *neu* 'snow' or *roca* 'rock'. Only by 5th grade, did children start to include more sophisticated, morphologically complex terms such as *inundació* 'inundation' and *sisme submarí* 'sub aquatic seism'. Therefore, further research is needed to confirm the hypothesis of an effect of the semantic field in which word frequency is more rigorously controlled and upper school children are included.

Thirdly, the proportion of errors relative to the total number of tokens produced diminished grade by grade for all types of error. Future research expanding the analysis over the upper grades will confirm whether this pattern persists or tends to stagnation. Phonographic errors first and morphological misspellings from 2nd grade on are by far outnumbered by orthographic and lexical

errors. In other words, phonological and morphological analysis of the words to be written appear to play a more important role than analysis of the word context from very early on. While an emphasis on phonological analysis may be induced by a widespread focus on instruction of the phoneme grapheme correspondences as a means for writing, the role played by morphological analysis must rather be attributed to the morphological awareness promoted by a salient morphology in Catalan.

It is worth noting that given the nature of the task children were not given a determined set of words to write but they were able to choose the words they wrote down. The fact that a fairly noticeable percent of misspelled words persists in 5th grade suggests that they either lack consciousness concerning their difficulties about the spelling of particular words, or if they are aware of having such difficulties, this acknowledgement does not refrain them from writing them down all the same. In this line, other research has pointed out that children showed sensitivity towards their spelling difficulties in a text writing task (Chenu & Jisa, 2009). Thus spelling is more than a merely school-learned ability. It requires the perception, integration and mapping of linguistic information onto orthographic segments and this process is key to the development of linguistic literacy (Ravid & Tolchinky, 2002)

The developmental pattern of spelling when number of errors of each type is computed in relation to the total number of errors made sheds light on the differences between each type of error developmental path. Our results concerning the distribution of errors of each type in relation to the total number of errors made in each grade support our initial predictions that the proportion of lexical errors would increase with schooling against a decrease in all other types of error. We had expected, however that the phonographic type of error would decrease sooner and more acutely than any other type of error since, on the one hand, the knowledge required to avoid these errors, the phoneme grapheme correspondences has been shown to be consolidated early for rather transparent orthographies. On the other hand, such correspondences are perceived as the basis for writing in Catalan and they are therefore explicitly taught and profusely practiced in school. However we did not find such a decrease in the proportion of phonographic errors over the total number of errors made. Our results suggest

that children's development in spelling does not follow a neat continuous line but rather takes places through steps forwards and backwards, through gaining efficiency in solving different types of spelling problems and then losing part it because they are incorporating new pieces of knowledge useful for other spelling problems. For instance, while the word *enfadat* /amfadat/ 'angry' is habitually misspelled *amfadat* in 1st grade following phonographic transcription, many 2nd graders show lexical knowledge and produce the normative *enfadat*. However, in 3rd grade, instances of the misspelled form *ambfadat* for the same *enfadat* occur, showing that children are overgeneralizing the conventional transcription of *amb* /am/ 'with' , that is, a high frequency preposition, to a segment morphologically different but phonologically equivalent segment. In fact, the proportion of morphological errors relative to the number of spelling errors is the only type of error to show a sustained decrease between 1st and 5th grade. Virtually no flexion or derivation affix in Catalan can be spelled correctly through straight PGCR application. Even the application of context dependence rules is insufficient or useless for rendering the correct spelling of such segments. Our results indicate, therefore, that a salient morphology in Catalan together with increased linguistic knowledge with schooling triggers morphological awareness, which in turn facilitates recognition of the morphological status of affixes. This recognition would lead to producing the correct spelling.

The possibility that children are solving inconsistent spellings on the basis of their lexical knowledge does not seem much plausible in the light of the results we obtained for children's spelling of word stems as compared to suffixes (we obtained a better spelling performance on affix than on stem spelling). Such recognition, however, clearly does not solve at once all the difficulties encountered by the writer and future research is needed in order to obtain a more fine-grained picture of the differences between spelling derivational and inflective suffixes, or between nominal and verbal inflection. In this line, we have initial evidence that children made fewer misspellings in nominal than in verbal flexion. The contrary was found for Spanish (very similar to Catalan in terms of morphological typology). However, in Catalan spelling an inflected noun (mostly plural nouns in our corpus) entails control of an inconsistent letter to sound correspondence (sabates /səbatəs/ 'shoes'), whereas spelling a verbal suffixation (mostly

infinitives in our corpus) requires solving the rarer circumstance (in Catalan) of writing a phonologically empty letter (*jugar* /ʒuγa/ 'to play'). Morpho-phonological recovery of this segment would afford an additional means to attain correct spelling in verbal flexions and also in some noun forms. Future research on the corpus will provide support for or against the claim that these types of cue does not necessarily result in faster spelling learning (Gillis & Ravid, 2006) perhaps because they entail a two step action, that is, recognition of the morphological status of the segment followed by the morphological manipulation of the word in order to recover the phonologically empty segment in another inflected form of the paradigm.

We could speculate on a possible effect of spelling instruction on spelling performance, at least in the lower grades when spelling is the major focus of writing practices and more attention is placed on the acquisition of phoneme grapheme correspondences. From 3rd grade on, the acquisition of the notational aspects of writing is considered to be completed and children are required/encouraged to produce complete pieces of texts and to gain increasing command of the discursive style of writing. This both entails and fosters writing new words recently incorporated in the child's lexicon. It may well be that the child turns to phonographic correspondences for new words, and this might be a consequence of an excessively narrow approach to spelling by spelling instruction practices. However, relying only on phonographic correspondences for writing words in Catalan is not sufficient to render accurate spelling and a more multifaceted analysis of the word is required when spelling is approached as an interface of the phonological, lexical and morphosintactic levels of language. However, in order to establish the exact extent of an effect of instructional practices, specific research needs to be planned that takes into account different instructional approaches (Rieben, Ntamakiliro, Gonthier, & Fayol, 2005)

Finally, we would like to make a few considerations on the corpus-based approach taken in this study. Although the compiled data has limitations in terms of between subjects comparisons since they neither produced the same number of words nor did the words produced by each child posit the same number and type of difficulties, we are convinced that an analysis approaching spelling research from looking at what happened when children were allowed to select and write

down as many words as they wanted has provided us provides us with real language data from which to obtain a general picture of what is and what is not a serious problem for both beginner and more skillful spellers. This in turn provides us with an extremely useful platform from which to discern where to carry out research in the future. The fact that the CesCa corpus contains written texts in addition to the vocabularies, both having been written by the same children, only enhances, we believe, its potential for future research.

CHAPTER 8

Conclusions and future directions

In this final chapter, I look back at what has been accomplished in this thesis. Overall, the main accomplishment has been to expand the understanding of the process of language development in Catalan as shown in written texts produced by school children and adolescence ranging from 5 to 16 years.

We designed a corpus, that provides us with some 240,000 vocabularies representing 5 different semantic fields (food, clothing, leisure activities, traits of personality and natural phenomena) and some 11,000 texts representing 4 different genres (narrative, argumentative, definitional and conversational) written in Catalan by some 2,300 schoolers native speakers of a variety of languages, attending 32 schools distributed cross-nationally (Catalonia).

By using a corpus-based approach, this thesis has advanced the understanding of noteworthy issues of corpus annotation of written texts produced by non-expert writers, a tasks that poses a series of particular problems such as lemmatization of non-normative words, unconventional segmentation of words, graphic variation, use of mixed languages, among others.

General trends on later language development in written Catalan

The lexicon stands as the domain that experiences the most marked growth throughout compulsory schooling, most particularly throughout gradeschool. In this period children produce increasingly longer texts. With age, children increase their world knowledge and this brings about an increase of their lexicon, they acquire new lexical items and constructions with which to name physical entities but also abstract concepts. They can elaborate on a diversity of topics, and can do so more profusely and more in detail. Also, they develop their own personal opinion and perspective and learn to introduce that personal stance in their texts. Finally, they become more fluent writers, more capable of planning, producing and reviewing their texts. All this is reflected in the length of the text.

In addition, children writers progressively adjust their lexical uses to the different genre-specific requirements. This affects the very text length: a definition is clearly shorter than an explanation, but also other lexical features such as word length and use of (complex) nominalizations which are much more likely to occur in more formal school based texts such as explanations and definitions than in less school like texts such as recommendations and jokes. However, writing a recommendation appears like an excellent opportunity for producing adjectives, even at young ages in elementary school and writing a jokes alleviates some of the burden of planning and allows young children to produce fairly long texts.

However, this path of growth suddenly recedes at secondary school. Unexpectedly, texts decrese in length, and this recession extends to other aspects of lexical development such as use of nominalizations and adjectives. Why this happens and what produces a second wind of growth at the end of this period, in 10th grade, remains an open question. It is worth noting that although planning and undertaking a grade by grade characterization of the lexical development is a considerable effort and takes a lot of time and resources, it has made it possible to obtain this detailed developmental pattern. Otherwise, the recession in both quantity and quality in the text-embedded lexical uses could be mask as simple stagnation or slowed down growth.

The domain of syntax use experiences less marked and slower development throughout compulsory schooling. There is a sustained but slight increase in number of words per clause as well as in complexity both at the noun phrase and clause levels. However, it is only by 10th grade were the differences become

significant. The differentiation of the syntax uses by type of texts develops at a similarly slow pace. Again, it is only by 10th grade when text-driven complex syntax appear, both in explanations and recommendations at the noun phrase level, but only in explanations at the clause level. Lexical and syntactic uses correlate, more so in 6th graders' texts than in their 2nd and 10th grade peers. However, the correlations found do not yield an overall neat pattern. Several studies have shown that syntactic development relies strongerly on size of the lexicon in early childhood in several languages (although typological properties of the language may condition the configuration of this relationship). Our results suggest that this relation continues but becomes more relaxed throughout middle childhood and adolescence. It would appear as if in spite of the fact that a solid lexical growth is in place that could be ready to sustain and clothe the complex syntactic uses, schoolers throughout compulsory schooling frame their texts in a syntactic architecture showing that has attained a rather stable state. In this state, instances of complex syntactic structures are found, but they are not a common feature. A different explanation for this scarcity of use of complex, low frequency syntactic structures could be that children are finding difficult to organize and cast the flow of their intended information into such complex architectures. That is, that they have a hard time attaining command over writing as a discourse style. Although, many widely used rubrics for teaching and assessing writing include some criteria for evaluating sentence construction as one component of a text's overall quality, and despite evidence that explicit instruction on aspects of sentence construction, especially sentence-combining, improved writing performance (Saddler & Graham, 2005), writing instruction has generally moved away from sentence-level exercises to a focus upon higher-level processes such as planning, organization, and the composition of authentic discourse (Connors, 2000). The relationship between syntax and writing quality is of an indirect nature and more complexity does not necessarily equal better writing: it is rather variety of sentence structure, not complexity of sentence structure that makes texts flow. However, it is important to gain better acquaintance about the amount and type of guidance that Catalan school children receive on use of syntax and on the characteristics of the written discourse

In contrast, the orthographic aspects of writing are more easily dealt with

as shown by the evident changes undergone by the spelling domain most particularly within the two first grades of elementary school. On the one hand, children have attained good command of phonographic correspondences by then and most worth noting they also show marked sensitiveness to the morphological composition of the word (rich in Catalan through both flexion and derivation) and use this knowledge in their morphology-based spellings that yield a dramatic decrease in the number of errors. All this notwithstanding, a percentage of 23% misspelled words remains by 5th grade. This is a considerable remain, given that the Catalan orthography is relatively transparent. Moreover, although with age an increasing percentage of these misspellings are caused by lack of orthographic and/or lexical knowledge, phonographic errors persist throughout. Children's spellings are sensitive to the effect of word frequency and lexical category as shown by the fact that they produce more errors when spelling adjectives a later developing grammatical category. In contrast, the rate of misspellings decreases when they write names of Natural phenomena. It is possible to think of a developing relationship (but this need to be further examined) between this lower rate of error and the fact than children learn many of these terms through text-book reading. Thus, at the initial steps of learning to write, children would focus almost exclusively on the phonographic relationships of the language as a means to render the written form of words. With increased experience with the written language, however, this would become itself a source of orthographic knowledge.

The results obtained in this study reencounter the well known phenomenon that overall language development in the school and adolescence years is the product of development in each domain of language. Although each domain follows a particular developmental pathway, there is reason to argue for an interaction between the development of each domain with the others. Thus, we have seen that young children are sensitive to the rich morphology of the Catalan language as shown in their spellings. Then, we have also seen that morphology plays a key role in later lexical development by allowing the acquisition (and interpretation) of morphologically complex words, usually sophisticated, abstract, literate words. We have found more moderate evidence of an interaction between the lexical and syntactic domain. Certainly we did find correlations between the two domains but the pattern of these correlations was not altogether clear. Further research is

needed expanding the number and the type of indicators of syntactic complexity, in order to obtain results more readily comparable to other studies in which a more neat pattern of correlations was found (Berman & Nir, 2009).

This work has several linguistic implications.

First, throughout compulsory schooling, language experience a remarkable development affecting each of its components. New lexical items and multiword constructions, many times difficult to reduce to a *single word*, are added to the child's lexicon. These lexical forms are characterized by growth in length (due in many cases to morphological complexity) and in sophistication and depth on meaning are added to the child's lexicon. Spelling also benefits from this increase in morphological awareness which leads to a dramatic decrease in misspellings in morphological segments. Another worth noting feature of the later developing lexicon is the notable growth of one particular lexical category, the adjective, which expands in size but also grows to include (morphological) complex exemplars. Use of adjectives contributes to the quality of the (written) expression and also has been related to increase the complexity of noun phrases, a preferred site for developing syntactic complexity according to our results. Further research is needed, however, in order to extend our undersatnding of the developmental path of acquisition of vocabulary depth. The use of synonims and antonims, and the path by which children add secondary or more abstract, figurative meanings to the terms they are already using in more primary ways will contribute relevant information. Also important, research must focus on the particular intricacies of the verbal paradigm, for instance, the uses of the subjunctive and conditional verbal modes. And there is a long way to go in the research on the development of morphology, a domain that has proven to underpin several aspects of language development in Catalan.

Second, a worth noting feature of the CesCa (written) corpus was the fact that we did not find in the written productions a significant rate of interaction between the many languages spoken by the participants. In addition, the multilingual condition of some of them had a favorable effect on some aspects of language development such as lexical growth. However, this work provides a

characterization of the features of the written language only. A different picture might be revealed in the spoken uses of the language, where code switing and mixing has been found to be more prevalent. It is of key importance, therefore, to extend the present work by adding information on the spoken uses of Catalan by schoolers as well as the relation between the two modalities of the language.

Third, this work adds support to the claim that a wide variety of different languages must be accounted for by research on language development (Hakuta & Bloom, 1986; Berman, 2004). Just as the typological characteristics of the Catalan morphology may underpin the relevance shown by morphological aspects on both lexical and spelling development in this language; this might be different in languages with a more sparse morphology (as in fact has shown to be the case with regard to spelling), the language typology may determine that some measures work better than others. Given the scarcity of research on later langage development in Catalan, this work served as an opportunity to test some wide-spread used lexical measures. Some, such as word length, use of nominalizations and use of adjectives proved as useful, adquate measures of text-embedded lexical developmenat in catalan.

However, other measures i.e., lexical density and Heylighen's F-measure (intended to evaluate the level of lexical formality of a text on the basis of the distribution of different lexical categories) did not work so well. Lexical density neither yielded a developmental pattern nor it discriminated by type of text. According to these results we have no concluding evidence regarding the suitability of this measure as an indicators of school based language development. However, as it has been discussed in the corresponding studies, that literate uses of Catalan require the use of a rich system of pronouns that is hard to get under command. These difficulties may underlie the unexpected finding that young children or participants with short acquaintance with the Catalan language produced denser texts than other participants, older and better acquainted with the language. As for Heylighen's F-measure of formality, we did not obtain evidence of developmental changes, although it discriminated finely by type of text

Nevertheless, further research is needed in order to confirm whether such lack of suitability of these two of measure for characterizing language development in Catalan is due to the typological characteristics of this language or, rather, to the

fact that the in general our participants yielded low rates of productivity, and produced rather short texts, a feature that discourages the use of other distributional measures such as lexical diversity.

Fourth, later language development has been characterized by the expansion of the linguistic repetoire just as much as by the imcreasing ability to deploy such repertoire in a wider range of communicative circumstances. Research must therefore focus on a wide range of genres, for different genres drive use (and growth subsequently) of different aspects of language. Our work adds clear evidence in the line that differences can be masked even between types of text subsumed under one same genre. For instance, although there is a considerable amount of research on the development of the definitional skills, very few works have included definition of words of different lexical category. However, our study proves that children use different strategies and different linguistic resources in each type of definition. In this work we have focus on the examination of local aspectes of linguistic expression. Future research will have to focus on the global aspects. It would be possible to undertake most of the purported studies on the CesCa corpus.

Fifth, corpus linguistics provides the linguistics community with a powerful tool for research. On the one hand, given the importance to test and confirm theoretical proposals and hypothesis with authentic uses of language in a diversity of circumnstances. This is particularly important, we think, with regard to the written productions of non-expert writers since no experimental design can capture the richness of the expressive resources deployed by speakers/writers. On the other hand, because it provides the researcher with the means for handling larger samples and range of variables. In the present studies we took a corpus-based approach, but we see it convenient, in future research to undertake corpus driven analyses, that is, to examine the range (and characterization) of lexical and syntactic patterns underlying the texts. Given the present results, this may be especially worthwhile with regard to obtaining a fine grained characterization of the syntactic structures used by compulsory school children.

This work has also educational implications

Disposing of a corpus of written texts by school children and adolescents currently attending compulsory schooling represents a significative contribution to the educationist community. It gives teachers the opportunity to explore a much wider sample of texts than they can dispose of in their habitual school setting, and it can become the referent against which, they can compare their student's written texts. The current analysis, and also all the future ones, provide the educationist community with a guideline of what aspects (local and global) are relevant for writing quality, and access to specific exemples of both good and bad text-embedded realization of each of the aspects under analysis. Most important is the fact that the available examples are not abstractions on language production but excerpts of texts actually produced by schoolers. We think that because all of this is much more tangible and specific than theoretical elaborations on the elusive notion of text quality, teachers can feel more safely lead to craking down the diversity of components contributing to a good text, the extent of that contribution, the types of difficuties the child writer can encounter,

All this research based on authentic texts should generate rich debate an interaction between the research and the educationist communities aiming at bridging the gap between the researchers' interests and the educationists' needs. Given the educational difficulties deriving from the multilingual condition of the school population, the results concerning the effcet of this multilingualism are of particular interest. The present results, as well as future results obtained from further analyses of the corpus should be taken into account by specialists in curricula design (programs, goals, teaching practices and methods of assessment should take into account the current reality as well as research-based validated indicators). Thus, the present results can inform educationist practicioners on the importance of 1) productivity, 2) text-embedded lexical uses, 3) text-embedded syntactic uses and 3) spelling on overall text quality.

1) We have obtained rather low rates of productivity (in number of tokens per text) throughout schooling (instances of just one-sentence long texts were not rare). The types of text representing typically school-based types of texts (explanations and definitions) showed stronger growth than less practiced genres

(recommendations and joke telling). However, overall growth from 1st graders' and 10th graders' written texts was lower than expected. In particular, we find stagnation in productivity and lexical growth throughout secondary education to be particularly upsetting. It is crucial that we find the cause of that stagnation so that researchers and educationist can tackle it together. It might be that highschoolers are failing to continue to develop content knowledge which would generate lack of provision of subject matter to write about. Or it might be that they lack strategies for text generation, that their planning do not take into account all the steps involved in skilful text writing. Learning to write skilfully takes a lot of practice with writing, but the particularities of the linguistic configuration of the Catalan schoolers population have caused that a strong focus is placed on oral communication as a means to counterbalance multilingual interaction. Our results point that writing can actually be a good way to leave switch and/or mix between codes aside.

2) The found stagnation somewhat affected measures of text-embedded lexical growth such as use of nominalizations and adjectives. A rich lexicon is key for precise expression and literate language uses are characterized for the ability to deploy in a genre-appropiate form an advanced lexicon containing low frequency morphologically complex words and constructionsa and specialized terms, and using abstract, figurative language realized through polysemous lexical items. An important aspect of developing a literate lexicon is the ability to use morphological knowledge as a source of autonomous word learning of potentially knowable words. The overall picture obtained in this study points to the importance of morphology and morphological awareness in the development of interrelated aspects of language. However, it is important that teaching practices ensure that learning of new complex words is realized in text-embedded contexts. To have some knowledge of a word may not sufice for producing good quality texts. Children must be given multiple opportunities to use their growing vocabularies in a variety of contexts, for a variety of purposes. The relative imperviousness of the written language to interaction between languages with regard to lexical development must be tested in other aspectes of language such as syntactic complexity.

3) The stagnation in productivity goes hand in hand with the moderate development of syntactic complexity. Although highschoolers produced overall longer texts than their younger peers, this inscrease was due to their writing more clauses more than to their producing longer clauses containing more complex (longer therefore) noun phrases or subordinated clauses. Our results point to an increase in syntactic complexity throughout compulsory schooling. In the period extending between 6th grade (end of elementary school) to 10th grade (end of secondary school), however, this increase is very moderate. Although mere use of complex syntax does not necessarily equals better quality, complex noun phrases indicate higher precision in denoting a particular entity or concept, just as use of (adverbial) subordinates serves the expression of additional information concerning, the time, the place and the circumstances of the exposed facts or events. Teachers should be aware of the relationship between certain grammatical resources and the effect they produce in the text in order to shift their instructional focus from the teaching of grammar to guiding their students on the actual use of these resources. Overall, the moderate increase in use of complex syntax may be due to persisting difficulties with the written language as a discourse style. It is important, we think, that children are given plenty of occasion to produce written texts for command of this modality of the language is key for adequate levels of literacy.

4) The teaching community has a lot to learn, we think, from the presently obtained results on development of spelling in school grade. Two facts may be of particular interest for them. First, the fact that phonographic errors persist through school grades even though a main focus if placed on the teaching of the phoneme grapheme correspondences as well as into practices that foster phonological analysis of the words. A phonographic approach to spelling words in Catalan, however, is largely insufficient for rendering conventional orthographic representations of words. Second, the fact that children show marked sensitivity to the morphological composition of the words they are spelling. It is therefore important that spelling practices be integrated in the overall reflection on the language instead of being taught as an additional or independent superficial feature of the language. It is important that spelling is considered one of the aspects, but not the only one or the main one, contributing to writing quality.

Additionally, we have initial findings of a relation between source of input (text-book) and spelling rate. Although further research is needed in order to confirm this trend, it brings up a debate with interesting potential for the teaching community: the relationship between reading and writing, the role of reading as a source for learning to write and for general language development.

Overall, our results point to the need of providing school children the opportunity to reflect about and practice with all the domains of the language: the lexical pieces whether monoword or multiword, the morphological devices, the syntactic devices, the constraints of the spelling system and the orthography ; but also they must think and discover the relation of one domain with the others. No single domain contributes uniquely to writing quality. Rather, the child writer needs to learn to use them all in a skilful way in order to shape the content of his texts in a meaningful expressive form. Abundant and purposeful practice with the written language looks as the only possible way to foster this kind of competence in both the written and the spoken modalities of his language.

Future research would benefit from including educational intervention that can lead to research-grounded evidence on the extent of the impact of instructional practices on text writing performance and on general aspects of school-based language development.

References

Albert, M. & Tolchinsky, L. (2010). Un robón no es lo mismo que un senyor que roba: el desarrollo de la estructura formal y semántica en la definición de diferentes categories morfológicas de palabras. Communication presented at the 6th International Conference on Language Acquisition, Barcelona

Alegría, J. & Mousty, P. (1996). The development of spelling procedures in French-speaking, normal and reading-disabled children: Effects of frequency and lexicality. *Journal of Experimental Child Psychology*, *63*(2), 312-338.

Anderson, R. & Freebody, P. (1981**)**. Vocabulary knowledge. In J. Guthrie (Ed.) *Comprehension and teaching research reviews* (pp. 711-71). Newark: International Reading Association.

Anglin, Jeremy (1970). *The Growth of Word meaning* . MIT Cambridge, Mass.

Anglin, J. (1993). Vocabulary development: A morphological analysis. *Monographs of the society for research in child development, 58*(10, Serial 238), 1-165.

Aronoff, M. (1994). Spelling and culture. In W.C. Watt (ed.) *Writing systems and cognition*. Dordrecht: Kluwer.

Aronson, K. & Thorell, M. (2002). Adult-chold talk and reaccentuation in Children's play. In S. Blum-Kulka and C. Snow (Eds.) *Talking with adults the contribution of multi-party talk to language development*. Mahwah NJ: Lawrence Earlbaum.

Auer, P. (1984). *Bilingual conversation*. Amsterdam, Netherlands: John Benjamins.

Auer, P. (1998). *Code-switching in conversation: language, interaction, and identity*. London, England: Routledge.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda*. Washington, DC: National Academy Press.

Baayen, R.H., McQueen, J., Dijkstra, T. & Schreuder, R. (2003). Frequency effects in regular inectional morphology: Revisiting Dutch plurals. In R.H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 355-390). Berlin: Mouton de Gruyter.

Baayen, H. R. & Renouf, A. (1996). Chronicling The Times: Productive Lexical Innovations in an English Newspaper. *Language*, *72*, 69-96.

Banks, J. (1993). Approaches to multicultural curriculum reform. In J. Banks and C. Banks (Eds.), *Multicultural education: Issues and perspectives*. Boston, MA: Allyn & Bacon.

Bar-Ilan, L., & Berman, R. (2007). Developing register differentiation: The Latinate-Germanic divide in English. *Linguistic*, *45*(1),1-35.

Barata, A. (2010). *Visual writing*. Cambridge Scholars Publishers.

Barry, C. (1994). Spelling routes (or roots or rutes). In G. D. A. Brown & N. C. Ellis (Eds.) *Handbook of spelling: Theory, process and intervention* (pp. 27-49). New York: John Wiley & Sons.

Bates, E,. & Goodman, J. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real time processing. In G. Altmann (Ed.), Special issue on the lexicon, *Language and Cognitive Processes, 12*(5/6), 507-586.

Beers, S. F., & Nagy, W. (2009). Syntactic Complexity as a Predictor of Adolescent Writing Quality: Which Measures? Which Genre? *Reading and Writing: An Interdisciplinary Journal, 22*, 185-200.

Beers, S. F., & Nagy, W. (2010). Writing development in four genres from grades three to seven: Syntactic complexity and genre differentiation. *Reading and Writing: An Interdisciplinary Journal*, 24, 183-202.

Bender, L. (1968). Analysis of a Barya Word list. *Anthropological Linguistics*, 10(9), 1-24.

Berko, J. (1958). The Child's Learning of English Morphology. *Word*, 14, 150-177.

Berman, R. A. (2004). *Language development across childhood and adolescence: Psycholinguistic and crosslinguistic perspectives.* Amsterdam: John Benjamins.

Berman, R. A. (2004). Between emergence and mastery: The long developmental route of language acquisition. In R. A. Berman (Ed.) *Language development across childhood and adolescence. Trends in language acquisition research*, (Vol 3, pp. 9-34). Amsterdam: John Benjamins.

Berman, R. A. (2005). Introduction: Developing discourse stance in different text types and languages. *Journal of Pragmatics, 37*(2), 105-124.

Berman, R. A. (2007). Developing language knowledge and language use across adolescence. In E. Hoff & M. Shatz, (Eds.) *Handbook of language development* (pp. 346-367). London: Blackwell Publishing.

Berman, R. A. (2008). The psycholinguistics of developing text construction. *Journal of Child Language*, *35*, 735–771.

Berman, R, & Katzenberger, I. (2004). Form and function in introducing narrative and expository texts: A developmental perspective. In R. A. Berman & B. Nir-Sagiv (Eds.) *Discourse Processes*, *38*(1), 57-94.

Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes, 43*(2), 79-120.

Berman, R. A. & Nir, B. (2009). Cognitive and linguistic factors in evaluating text quality. Global versus local. In V. Evans & S. Pourcel (Eds.) *New Directions in Cognitive Linguistics.* Amsterdam: John Benjamins.

Berman, R. A., & Ravid, D. (2008). Becoming a literate language user: Oral and written text construction across adolescence. In D. R. Olson & N. Torrance (Eds.), *Cambridge handbook of literacy* (pp. 92–111). Cambridge: Cambridge University Press.

Berman, R. A. & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study.* Hillsdale, NJ: Lawrence Erlbaum.

Berman R. A. & Verhoeven L. (2002). Cross-linguistic perspectives on the development of text-production abilities: speech and writing. *Written Language and Literacy, 5*(1), 1-43.

Berninger, V, Yates, C., Cartwright, A., Rutberg, J., Remy, E. & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal*, *4*, 257–280

Berninger, V., & Winn, W. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.) *Handbook of Writing Research* (pp 96- 114). New York: The Guilford Press.

Bialystock, E. (2001). *Bilingualism in Development: Language, Literacy, and Cognition*. Cambridge: CUP

Biber, D. (1988). *Variation across spoken and written English.* Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation. A cross-linguistic comparison*. Cambridge, MA. Cambridge University Press.

Biber, D (2007). On the complexity of discourse complexity: a multi-dimensional analysis. In T.A. van Dijk (Ed.) *Discourse studies* [Vol. 1] (pp. 127-157). London: Sage.

Blodgett, E. G. & Cooper, E. B. (1987). *Analysis of language of learning* Linguisystems.

Blum-Kulka, S. (2010). Explanations in naturally occurring peer talk: Conversational emergence and function, thematic scope, and contribution to the development of discursive skills. *First Language, 30*(3/4), 440-460.

Bokamba, E.G. (1989). Are there syntactic constraints on code-mixing? *World Englishes, 8*, 277-292.

Boleda, G. (2006). *Automatic acquisition of semantic classes for adjectives*. PhD dissertation. Universitat Pompeu Fabra.

Borer, H. (2005). *In name Only.* New York: Oxford University Press.

Borromba: Diccionari bàsic il·lustrat (1976) [Illustrated Dictionary] La Galera, Barcelona.

Bialystok, E., & Feng, X. (2010). Language proficiency and its implications for monolingual and bilingual children. In A.Y. Durgunoglu & C. Goldenberg (Eds.), *Dual language learners: The development and assessment of oral and written language.* (pp. 121-138). New York: Guilford Press.

Brown, R. (1958). *Words and things*. Glencoe. Ill: Free Press.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Burke, D., MacKay, D., Worthley, J. S. & Wade, E. (1991). On the tip of the tongue: What causes word finding failure in young and older adults? *Journal of Memory and Language,* 30, 237-246.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford. Oxford University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, J. & Noonan, M. (2002). *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson*. Amsterdam: John Benjamins.

Byrne, A., MacDonald, J. & Buckley, S. (2002). Reading, language and memory skills: A comparative longitudinal study of children with Down syndrome and their mainstream peers. *British Journal of Educational Psychology*, *72*, 513-529.

Cain, K., Towse, A. & Knight, R. (2009). The Development of Idiom Comprehension: An Investigation of Semantic and Contextual Processing Skills. *Journal of Experimental Child Psychology*, *102(3)*, 280-298.

Cameron, C. A., Hunt, A. K., & Linton, M. J. (1988). Medium effects on children's story rewriting and story retelling. *First Language*, *8*, 3–18.

Caravolas, M., & Bruck, M. (1993). The effect of oral and written language input on children's phonological awareness: A cross-linguistic study. *Journal of Experimental Child Psychology*, *55*, 1-30.

Carrillo, M.S. Alegria, J. & Marin, J. (in press). On the acquisition of some basic spelling mechanisms in a deep (French) and a shallow (Spanish) system. *Reading and Writing*.

Carlisle J. & Nomanbhoy, D. (1993). Phonological and morphological development. *Applied psycholinguistics*, *14*, 177 -195

Carrillo, M.S. Alegria, J. & Marin, J. (in press). On the acquisition of some basic spelling mechanisms in a deep 8French) and a shallow (Spanish) system.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanded, L. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10,* 159-199.

Caselli, M. C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language*, *26*, 69-111.

Cenoz, J. (1997) *The influence of bilingualism on multilingual acquisition: some data from the Basque country*. Actas do I simposio internacional sobre o bilinguismo.

Civit, M. (2003). *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*. Procesamiento del Lenguaje Natural, Colección Monografías, 3. ISBN: 84-600-9944-X

Chafe, W. (1994). *Discourse, consciousness, and time: The flow of language in speech and writing.* Chicago, IL: Chicago University Press.

Chafe, W., & Danielewicz, I. (1987). Properties of spoken and written language. In R. Horowitz & S. I. Samuels (Eds.), *Comprehending oral and written language* (pp. 83-113). San Diego: Academic Press.

Chenu, F. & Jisa, H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development", *AILE (Acquisition et Interaction en Langue Etrangère)*, 9(1), 17-38

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, E., Hetch, B. & Mulford, R. C. (1986). Coining complex compounds in English. *Linguistics*, *24*, 7-29.

Coirier, P., Gaonac'h, D. & Passerault, J.M. (1996). *Psycholinguistique textuelle : une approche cognitive de la compréhension et de la production des textes*. Paris : Colin.

Cordero Alonso, J. A., (2002). *L'aprenentatge del català com a segona llengua. Evolució dels aspectes lèxics al llarg de l'escolaritat obligatòria*. Unpublished manuscript Report to the Department of Culture of the Catalan Government. Barcelona.

Corsaro, W. A. (1985). *Friendship and peer culture in the early years*. Noorwood, NJ: Abblex.

Cosme C. (2008). A corpus-based perspective on clause-linking patterns in English, French and Dutch. In C. Fabricius-Hansen & W. Ramm (Eds.), *Subordination vs. coordination in sentence and text from a cross-linguistic perspective*. Amsterdam, Benjamins.

Costa, A., Hernandez, M. & Sebastian-Galles, N. (2008). Bilingualism aids conflict

resolution: Evidence from the ANT task. *Cognition* , *106*, 59-86.

Crystal, D. (1996). *Rediscover grammar*. Essex, England: Longman.

Cristofaro, S. (2003). *Subordination*. Oxford: Oxford University Press.

Cunningham, A. E., Perry, K. E., Stanovich, K. E., & Share, D. L. (2002). Orthographic learning during reading: examining the role of self-teaching. *Journal of Experimental Child Psychology*, *82*, 185–199.

Cunningham, A. E., & Stanovich, K. E. (1997). Early acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33(6)*, 934–945.

de Jong, P. F. & Share, D. L. (2007). Orthographic learning during oral and silent reading. *SSSR Journal, 11,* 55-71.

Defior, S., & Alegria, J. (2005). Conexion entre morfosintaxis y escritura: Cuando la fonologıa es (casi) suficiente para escribir. [Morphosyntax and spelling connection: when phonology is (almost) enough for spelling]. *Revista de Logopedia, Foniatrı´a y Audiologı´a*, *25*(2), 51- 61.

Defior, S. & Serrano, F., (2005). The initial development of spelling in Spanish: From global to analytical. *Reading and Writing*, 18, 81–98.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge: Cambridge University Press.

Dockrell, J., Lindsay, G. & Connelly, V. (2009). The impact of specific language impairment on adolescents' written text. *Exceptional Children*, *75*(4), 427–446.

Dockrell, J. & Messer, D. A. (2004). Lexical acquisition in the early years. In R. A. Berman (Ed.), *Language Development across Childhood and Adolescence.* Trends in Language Acquisition Research Series (3) (pp. 35-52). Amsterdam, Netherlands: John Benjamins Publishing Company.

Dressler, W. (2004). Degrees of grammatical productivity in inflectional morphology. *Italian Journal of Linguistics*, 15, 31-62.

Du Bois, J. (2003). *Preferred Argument Structure: Grammar as Architecture for Function*. Amsterdam: Benjamins.

Durkin, K., Crowther, R.D. & Shire, B. (1986). Children's processing of polysemous vocabulary in school. In K. Durkin (Ed.), *Language Development in the School Years*. Croom Helm, London.

Ehri, L. (1997). Learning to read and learning to spell are one and the same, almost. In C. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to Spell: Research, Theory and Practice Across Languages*(pp. 237-269). Mahwah, NJ: Erlbaum.

Ellis, R. (2002). A metaphorical analysis of learner beliefs. In P. Burmeister, T. Piske & A. Rohde (Eds.), *An integrated view of language development: Papers in honor of Henning Wode*. Trier, Germany: Wissenschaftlicher Verlag.

Ellis, W. S. & Evans, J.L. (2002). The role of processing limitations in early identification of specific language impairment. *Topics in Language Disorders*. *22*(3), 15-29.

Fayol, M. (1991). From sentence production to text production. *European Journal of Psychology of Education*, (special issue on Writing), 101-119.

Feldman, L.B., Frost, R., & Pnini, T. (1995). Decomposing words into their constituent morphemes: Evidence from English and Hebrew. *Journal of Experimental Psychology: Learning. Memory and Cognition*, 21, 947–960.

Fenson, L., Dale, Ph.S., Reznick, J.S.; Thal, D., Bates, E.; Hartung, J.P., Pethick, S. & Reilly, J.S. (1993). *Mac Arthur Communicative Development Inventories: User's guide and technical manual***.** San Diego, Singular Publishing Group (2nd edition).

Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research*, *47*, 1301–1318.

Friedman, N., & Novogrotsky, R. (2004). The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child Language*, *31*, 661-681.

Frost, R. (1992). Orthography and phonology: The psychological reality of orthographic depth. In P. Downing, S. Lima & M. Noonan (Eds.), *The linguistics of literacy* (pp. 225-274. Ansterdam: Benjamins.

Frost, R., Katz, L. & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance* , *13*, 104-115.

Garcia, N. (2007). Phonological, orthographic, and morphological contributions to the spelling development of good, average, and poor spellers. Unpublished doctoral dissertation. University of Washington.

Garcia, N. P., Abbott, R. D., & Berninger, V. W. (2010). Predicting poor, average, and superior spellers in Grades 1 to 6 from phonological, orthographic, and morphological, spelling, or reading composites. *Written Language & Literacy*, *13*, 61–98.

Garrett, P. B. (1999). *Language socialization, convergence, and shift in St. Lucia, West Indies*. Unpublished doctoral dissertation, New York University, New York.

Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English-language learners. *Exceptional Children*, *66*(4), 454–470.

Gillis, S., & Ravid, D. (2006). Typological effects on spelling development: a crosslinguistic study of Hebrew and Dutch. *Journal of Child Language*, *33,* 621-659.

Ginsburg, H. & Opper, S. (1988). *Piagett's theory of intellectual develop*ment, New York: Prentice Hall.

Gleitman, L. (1990). Structural sources of verb learning. *Language Acquisition*, *1*, 1-63.

Goetz, P. J. & Shatz, M. (2000). When and how peers give reasons: justifications in the talk of middle school children. *Journal of Child Language*, *26*(03), 721-748.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. (2005). *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Gollan, T. H. & Ferreira, V. S. (2009). Should I stay or should I switch? A cost-benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *35*, 640–665.

Goodwin, M. A. (1990). *He-said-she-said: Talk as social organization among black children*. Bloomington, IN: Indiana University Press.

Graham, S., Berninger, V., Abbott, R., Abbott, S., & Whitaker, D. (1997). The role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, *89*{1), 170-182.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*, Longman.

Halliday, M.A.K. & Hassan, R. (1985), *Language: Context and Text*. Burwood, Vic;

Deaken University.

Harley, T. A. (1995). *The psychology of language.* Hove: Psychology Press.

Hayes, J.R., & Chenoweth, N. (2006). Is Working Memory Involved in the Transcribing and Editing of. Texts*? Written Communication*, *23*(2), 135.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinbert (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hepburn, S., Fidler, D.J., & Rogers, S. (2008). Autism symptoms in toddlers with Down syndrome. *Journal of Applied Research in Intellectual Disabilities*, *21*, 48-57.

Heylighen, F. & Dewaele, J.M. (1999). *Formality of language: Definition measurement and behavioral determinants*. Internal report, Center Leo Apostel, Free University of Brussels.

Hickmann, M. (2003). Coherence, cohesion, and context in narrative development: some comparative perspectives. In S. Strömqvist & L. Verhoeven (Eds.), *Relating events in narrative: Typological and contextual perspectives* (pp. 281-306). Hillsdale, NJ: Lawrence Erlbaum.

Hillis, A. & Caramazza, A. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain*, *114*, 2081-2094.

Holmes, V., M., Stowe L. A., & Cupples, L. (1989). Lexical expectations in parsing complement verb sentences. *Journal of Memory and Language*, *28*, 668 689.

Hoyle, S. & Adger C. (1998) Introduction. En S. Hoyle y C. Adger (Eds.) *Kid Talk: Strategic language use in later childhood* (pp. 3-22). Nueva York: Oxford University Press.

Hudson, J. & Shapiro, L. (1991). From knowing to telling: The development of children´s scripts, stories, and personal narratives. In A. McCabe & C. Peterson (Eds.), *Developing narrative structure* (pp. 89-135). New Jersey: Lawrence Erlbaum Ass.

Hunt. K. W. (1970). Syntactic maturity in school children and adults. In *Monographs of the Society for Re search in Child Development*. University of Chicago. Press, Chicago, IL.

Hyltenstam, K. (1988) Lexical characteristics of near-native second language learners of Swedish. *Journal of Multilingual and Multicultural Development*, *9*. 67-84

Idescat: http://www.idescat.cat/cat/poblacio/poblcensling.html

Jackendoff, R. (2002) *Foundations of language*. Oxford University Press, Oxford.

Jescheniak, J. D. & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 824-843.

Jisa, H. (2004). Growing into academic French. In R. A. Berman (Ed.), *Language development across childhood and adolescence* (pp. 135-162). Amsterdam: John Benjamins.

Jisa H. & Mazur A. (2006). L'expression de la causalité : une étude développementale, Les savoirs savants aux savoirs enseignés. (Vol. 8, pp. 33-60) Namur : Presse Universitaire de Namur.

Jisa, H., Reilly, J.S., Verhoeven, L.,Baruch, E., & Rosado, E. (2002). Passive Voice construction in written texts: A cross-linguistic study. *Written Language and Literacy*, *2*(5),163-182.

Jisa, H. & Viguie, A. (2005). A developmental perspective of the role on in written and spoken expository tests. *Journal of Pragmatics* [Special issue on discourse stance].

Johansson, V. (1999). Word frequencies in speech and writing: a study of expository discourse. In A. Ravid (Ed.), *Working papers in developing literacy across genres, modalities, and languages*, (vol. I, pp. 182- 98). Tel Aviv: Tel Aviv University Press.

Johanson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers, 53*, 61-79. Lund University, Departament of Linguistics and Phonetics.

Joshi, R. (2005). Vocabulary: A critical component of comprehension. *Reading and Writing Quarterly*, *21*, 209-219.

Joshi, R. M. (2010). Role of orthography in literacy acquisition and literacy problems among monolinguals and bilinguals. In D. Aram and O. Korat (Eeds.) *Literacy Development and Enhancement Across Orthographies and Cultures*, (pp.167-176). New York, NY: Springer.

Con formato: Fuente: +Cuerpo (Calibri), 12 pto

Joshi, R. M., Treiman, R., Carreker, S., & Moats, L. C. (2009). How words cast their spell. *American Educator*, 6-43.

Kamhi, A. G. & Catts, H. W. (1989). Language and reading: Convergences, divergences, and development" En A. G. Kamhi & H. W. Catts (Eds.), *Reading disabilities: A developmental language perspective* (pp. 1-34). Boston, MA: College Hill.

Karmiloff-Smith, A. (1986). Some fundamental aspects of language development after age five. In P. Fletcher and M. Garman (Eds.), *Language Acquisition: Studies in first language development*. (pp. 455-474). Cambridge: Cambridge University Press.

Karmiloff-Smith, A. (1992). *Beyond Modularity*. MIT Press.

Katz, L. & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), Orthography, Phonology, Morphology, and Meaning. (pp. 67-84). Amsterdam: Elsevier Science Publishers B.V.

Katz, L. & Frost, S. J. (2001). Phonology constrains the mental orthographic representation. *Reading & Writing*, *14*, 297-332.

Kendeou, P., & van den Broek, P. (2007). Interactions between prior knowledge and text structure during comprehension of scientific texts. *Memory and Cognition*, *35*, 1567–1577.

Krashen, S. (1996). *Every person a reader*. Culver City, CA: Language Education Associates.

Laaha, S., & Gillis, S. (2007). *Typological perspectives on the acquisition of noun and verb morphology* [Antwerp Papers in Linguistics 112] Antwerp: University of Antwerp.

Larson, V. L. & McKinley, N. L. (2003). *Communication solutions for older students: Assessment and intervention strategies*. Wisconsin, Thinking Publications.

Laufer, B. & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing, 16*(1), 33-51.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16/3.

Leech, G. (1992). Corpora and Theories of Linguistic Performance. Directions in Corpus

Linguistics. Proceedings of Nobel Symposium 82, 105-122.

Leong, C.K., & Ho, M.K. (2008). The role of lexical knowledge and related linguistic components in typical and poor language comprehenders of Chinese. *Reading and Writing: An Interdisciplinary Journal, 21*, 559-586.

Leté, B., Peeremean, R., & Fayol, M. (2008). Consistency and word-frequency effects on spelling among first- to fifth-grade French children: A regression-based study *Journal of Memory and Language*, 58, 952–977.

Lieven E. V., & Tomasello M. (2008). Children's first language acquisition from a usage-based perspective: In P. Robinson and N. Ellis (Eds). *Handbook of cognitive linguistics and second language acquisition* (pp. 168-196). New York: Routledge.

Loban, W. (1963). *The language of elementary school children*. (NCTE Research report n. 1) Urbana, IL. National Council of Teachers of English.

Loban, W. (1976). *Language development: Kindergarten through grade twelve.* (Research Rep. No. 18). Urbana, IL: National Council of Teachers of English.

Lust, B., Foley, C., & Dye, C. (2008). The first language acquisition of complex sentences. In E. Bavin (Ed.), *Handbook of child language* (pp. 463–505). Cambridge: Cambridge University Press.

Lyons, J. (1968). *Introduction to theoretical linguistcs*. Cambridge University Press.

Lyons, J. (1977). *Semantics* (Vol. II). Cambridge: Cambridge University Press.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Science*. *14*, 348-356

Malvern, D., Richards, B., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basing-stoke: Palgrave Macmillan.

Marchman, V., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, *21*, 339–366.

Marchman, V., & Thal, D. (2005). Words and grammar. In M. Tomasello & D. I. Slobin (Eds.), *Beyond nature–nurture: Essays in honor of Elizabeth Bates* (pp. 139–164). Mahwah, NJ: Lawrence Erlbaum.

Marin, J., Carrillo, M., & Alegria, J. (1999). El proceso de aprendizaje de la escritura

en español y frances: Un estudio comparativo (Learning to spell in Spanish and French: A comparative study). IV Simposio de Psicolingıstica (IV Symposium in Psycholinguistic), Miraflores de la Sierra (Madrid).

Marjanovic-Umek, L., Feknoja-Peklaj, U. & Podlesek, A. (2012). Characteristics of early vocabulary and grammar development in Slovenian-speaking infants and toddlers: a CDI-adaptation study. *Journal of child language*, *20*(3), 233-246.

McKeown, M.G., Beck,I. L., Omanson, R. C. & Perfetti, C. A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior* , *15*(1), 3–18.

Miller, G., Beckwith, R., Fellbaum, C, Gross, D. & Miller, K. (1990). Introduction to WordNet: An Online Lexical Database. *International Journal of Lexicography*, *3*(4), 235-244.

Moffett, J. (1968). *Teaching the universe of discourse*. Boynton, Cook.

Mousty, P., & Alegria, J. (1999). L'acquisition de l'orthographe: donn´ees comparatives entre enfants normo-lecteurs et dyslexiques. *Revue Française de Pedagogie*, *126*, 7–22.

Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge, MA: Cambridge University Press.

Myers-Scotton, C. (1995). A lexically based model of code-switching. In L. Milroy & P. Muysken (Eds), *One speaker, two languages*, (pp. 233-256). Cambridge, MA: Cambridge University Press.

Myers-Scotton, C. (2001). Explaining aspects of codeswitching and their implications, in J. Nicol (Ed.), *One mind, two languages: Bilingual language procession* (pp. 84-116). Oxford: Blackwell.

Myers, L. & Botting, N. (2008). Literacy in the mainstream inner-city school: Its relationship to spoken language. *Child Language Teaching and Therapy*, *24*(1), 95-114.

Nagy, W., & Anderson, R. (1984). How many words are there in printed school English? *Reading Research Quarterly*, *19*, 304-330.

Nagy,W. E. & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for instruction. In M. McKeown & M. Curtis (Eds.), *The Nature of Vocabulary Acquisition*, (pp. 19–35). Hillsdale, NJ: Erlbaum.

Nation, K., Cocksey, J., Taylor, J. S. H. & Bishop, D. V. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry*, *51*(9), 1031-1039

Nation, K., Snowling, M. J. & Clarke, P. J. (2007). Dissecting the relationship between language skills and learning to read: semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension. *Advances in Speech-Language Pathology*, *9*, 131-139.

Nelson N.W. , Roth, F.P., Van Meter, A.M. (2009.)Written composition instruction and Intervention for students with language impairment. In G. A. Troia (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices*. (pp. 187–212.) New York: Guilford.

Nippold, M. A. (1998). *Later language development*. Austin, TX: PRO-ED

Nippold, M. A. (2002). Lexical learning in school-age children, adolescents and adults: A process where language and literacy converge. *Journal of Child Language*, *29*, 474-478.

Nippold, M. A. (2004). Research on later language development: International perspectives. In R. A. Berman (Ed.), *Language Development across Childhood and Adolescence*, Amsterdam-Philadelphia: John Benjamins.

Nippold, M. A. (2007). *Later language development: School-age children, adolescents, and young adults*. Austin, TX: Pro-Ed.

Nippold, M. A., Hesketh, L. J., Duthie, J.K. &Mansfield, T.C. (2005). Conversational versus expository discourse: A study of syntactic development in children, adolescents, and adults. *Journal of Speech, Language, and Hearing Research*, *48*, 1048–1064.

Nippold, M., Mansfield, T., & Billow, L. (2007). Peer conflict explanations in children, adolescents, and adults: Examining the development of complex syntax. *American Journal of Speech-Language Pathology*, *16*,1-10.

Nippold, M., & Sun, L. (2008). Knowledge of Morphologically Complex Words: A Developmental Study of Older Children and Young Adolescents. *Language, Speech, & Hearing Services in Schools*, *39*(3), 365-373.

Nippold, M. A., & Taylor, C. L. (2002). Judgments of idiom familiarity and transparency: A comparison of children and adolescents. *Journal of Speech, Language, and Hearing Research, 45*, 384-391.

Nir-Sagiv, B. (2005). *Word length as a criterion of text complexity: A cross linguistic developmental study*. Paper presented at triennial conference of the International Association for the Study of Child Language [IASCL], Berlin (July).

Nir-Sagiv, B., Bar-Ilan, L., & Berman, R. A. (2008). Vocabulary development across adolescence: Text-based analyses. In I. Kupferberg & A. Stavans (Eds.), *Studies in language and language education: Essays in honor of Elite Olshstain* (pp. 47-74). Jerusalem: Magnes Press.

Nunes, T. & Bryant, P. (2006). *Improving literacy through teaching morphemes* London: Routledge.

Nunes, T., Bryant, P., & Bindman, M. (1997). Morphological spelling strategies: Developmental stages and processes. *Developmental Psychology*, *33*(4), 637-649,

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English*, 40(4), 392-412.

Olson, D. (1994). *The world on paper*. Cambridge: Cambridge Univversity Press.

Olson, D. R, & Astington, J. W. (1990). Talking about text: How literacy contributes to thought. *Journal of Pragmatics*, *14*, 557-573.

O'Toole, C & Fletcher, P. (2012). Profiling vocabulary acquisition in Irish. *Journal of Child Language*, *39*(1), 205-220.

Owens R. E. (2008). *Language development: An introduction*. New York: Pearson Education.

Pacton, S. & Fayol, M. (2004). Learning to spell in a deep orthography: the case of French. In R. Berman (Ed.), *Language development across childhood and adolescence*. Amsterdam: Benjamins.

Payrató, L. (1996). *Català col·loquial. Aspectes de l'ús corrent de la llengua catalana*. València, Espanya: Universitat de València.

Paugh, A. L. (2001). *Creole day is every day: Language socialization, shift, and ideologies in Dominica, West Indies*. Unpublished doctoral dissertation, New York University, New York.

Paulesu, E., Démonet, J. F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N. Cappa, S. F,. Cossu, G., Habib, M. C. D., Frith, C. D., & Frith, U. (2001). Dyslexia: Cultural diversity and biological unity. *Science*, *291*, 2165-2167.

Perera, J., Aparici, M., Fité, M. & Busquets, J. (2010), L'empremta del castellà en les produccions lingüístiques dels catalanoparlants: una anàlisi evolutiva. En K. Faluba & I. Szijj (Eds.), *Actes del XIV Col·loqui Internacional de Llengua i Literatura Catalanes*, Vol. III (pp. 257-272). Barcelona: Publicacions de l'Abadia de Montserrat.

Peskin, J. & Olson, D.R. (2004). On reading poetry: Implications for later language development. In R. A. Berman (Ed.), *Language development across childhood and adolescence* (pp. 53-81). Amsterdam: John Benjamins.

Peters, A. (1997). *Language typology, prosody, and the acquisition of grammatical morphemes*. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition* (Vol. 5, pp. 135-197). Mahwah: Erlbaum.

Pinker, S. (1999). *Words and Rules*. London: Phoenix

Pollard, C. & Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press: Chicago.

Pontecorvo, C. & Zucchermaglio, C. (1988): Modes of Differentiation in Children's Writing Construction. *European Journal of Psychology of Education,* 3(4), 371–384.

Poplack, S. (1987). Contrasting patterns of code switching in two communities. In E. Wande, J. Andwars, B. Nordberg, L. Steensland, & M. Thelander (Eds.), *Aspects of bilingualism: Proceedings from the fourth Nordic symposium on bilingualism* (pp. 51-76). Uppsala: Borgström, Motala.

Purcell-Gates, V. (1988). Lexical and syntactic knowledge of written narrative held by well-read-to kindergartners and second graders. *Research in the teaching of English*, *22*, 128-160.

Pustejovsky, J. & Boguraev, B. (1993). Lexical knowledge representation and natural language processing. *Artificial Intelligence,* 63, 193-223.

Pustejovsky, J. (1995). *The Generative Lexicon.* MIT Press: Cambridge Mass

Rampton, B. (1995). Language crossing and the problematization of ethnicity and socialization. *Pragmatics*, *5*(4), 483-513.

Ragnarsdóttir, H., & Strömqvist, S. (2005). The development of generic mathur/man for the construction of discourse stance in Icelandic and Swedish. *Journal of Pragmatics, 37*, 143-155.

Rampton, B. (1998). Speech community. In J. Verscheren, J. Östman, J. Blommaert & C. Bulcaen (Eds.), *Handbook of pragmatics* (pp. 1-34). Amsterdam/Philadelphia, Netherlands: John Benjamins.

Ravid, D. (2001). Learning to spell in Hebrew: phonological and morphological factors. *Reading and Writing*, *14*, 459–85.

Ravid, D. (2004). Derivational morphology revisited: Later lexical development in Hebrew. In R. A. Berman (Ed.), *Language development across childhood and adolescence* (pp. 53-81). Amsterdam: John Benjamins.

Ravid, D. (2005). Hebrew orthography and literacy. In R. M. Joshi & P. G. Aaron (Eds.), *Handbook of orthography and literacy* (pp. 339–363). Mahwah, NJ: Lawrence Erlbaum.

Ravid, D. (2012). *Spelling morphology: The psycholinguistics of Hebrew spelling*. Springer.

Ravid, D. & Avidor A. (1998). Acquisition of derived nominals in Hebrew: developmental and linguistic principles. *Journal of Child Language, 25*, 229-266.

Ravid, D., & Berman, R. (2006). Information density in the development of spoken and written narratives in English and Hebrew. *Discourse Processes*, *41*(2), 117-149.

Ravid, D. & Saban, R. (2008). Syntactic and meta-syntactic skills in the school years: A developmental study in Hebrew. In I. Kupferberg & A. Stavans (Eds.), *Language education in Israel: Papers in Honor of Elite Olshtain.* (pp. 75-110). Jerusalem: Magnes Press.

Ravid, D. & Tolchinsky, L. (2002). Developing linguistic literacy: A comprehensive model. *Journal of Child Language, 29*, 417.447.

Ravid, D., Levie, R. & Avivi-Ben Zvi, G. (2003). Morphological disorders. In L. Verhoeven & H. van Balkom (Eds.), *Classification of developmental language disorders: Theoretical issues and clinical implications* (pp. 235-260). Mawhah, NJ: Erlbaum.

Ravid, D. & Saban, R. (2008). Syntactic and meta-syntactic skills in the school years: A developmental study in Hebrew. In I. Kupferberg & A. Stavans (Eds.), *Studies in*

language and language education: Essays in honor of Elite Olshstain (pp. 47-74). Jerusalem: Magnes Press.

Ravid, D. & Berman, R.A. (2009). Developing linguistic register across text types: The case of Modern Hebrew. *Pragmatics and Cognition, 17*, 108-145.

Ravid, D. & Berman, R. A. (2010). Developing noun phrase complexity at school age: A cross-linguistic text-embedded analysis. *First Language, 30*(1), 1-29.

Ravid, D. & Cahana-Amitay, D. (2004). Verbal and nominal expressions in narrating conflict situations in Hebrew. *Journal of Pragmatics* [Special issue on discourse stance], *37*(*2*), 157-184.

Ravid, D. & Levie, R. (2010). Hebrew adjectives in later language text production. *First Language*, *30*(1), 27-55.

Ravid, D. & Nir, B. (2000). On the development of the category of Adjective in Hebrew. In M. Beers, B. van den Bogaerde, G. Bol, Jan de Jong & C. Rooijmans (Eds.), *From sound to sentence: studies on first language acquisition* (pp. 113-124) Groningen: Center for Language and Cognition.

Ravid, D., & Schiff, R. (2006). Morphological abilities in Hebrew-speaking gradeschoolers from two socio-economic backgrounds: An analogy task. *First Language*, *26*, 381–402.

Ravid, D. & Sclesinger, Y. (1995). Factors in the selection of compound-type in spoken and written Hebrew. *Language Sciences*, *17*, 147-179.

Ravid, D. & Tolchinsky, L. (2002). Linguistic Literacy. *Journal of Child Language , 29*, 217-247.

Ravid, D., & Zilberbuch, S. (2003). Morpho-syntactic constructs in the development of spoken and written Hebrew text production. *Journal of Child Language*, *30*, 1–24.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, NJ: Erlbaum.

Reilly, J & Anderson, D. (2002). The acquisition of non-manual morphology in ASL. In G. Morgan & B. Woll (Eds.), *Current developments in the study of signed language acquisition* (pp. 159-181). John Benjamins.

Read, C., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic spelling. Cognition, *24*, 31-44

Reilly, J.S., Zamora, A., & McGivern, R.F. (2005). Developing Perspective in English: The Acquisition of Stance. *Journal of Pragmatics* ,*37* (2), 185-208.

Reitsma, P. (1983a). Printed word learning in beginning readers. *Journal of Experimental Child Psychology*, *36*, 321–339.

Reitsma, P. (1983b). Word-speci**W**c knowledge in beginning reading. *Journal of Research in Reading*, *6*, 41–56.

Reyes, A., Prosso, P., Martí, M.A, Taulé, M. (2009). Características y rasgos afectivos del humor: Un estudio de reconocimiento automático del humor en textos escolares en catalán. *Procesamiento del Lenguaje Natural*, *43*, 235-243.

Richards, T., Berninger, V. & Fayol, M. (2009). FMRI activation differences between 11-year-old good and poor spellers' access in working memory to temporary and long-term orthographic representations. *Journal of Neurolinguistics*, 22, 327-353.

Rieben, L., Ntamakiliro, L., Gonthier, B. & Fayol, M. (2005). Effects of various early writing practices on reading and spelling. *Scientific studies of reading*, *9*(2), 145–166.

Riedemann, H. (1996). Word length distribution in English press texts. *Journal of Quantitative Linguistics, 3*, 265-271.

Rumelhart, D. E.,& McClelland, J. L. (1986).On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Sampson, G. (2001). *Empirical Linguistics*. London: Continuum

Sampson, G. (2005). *The "language instinct" debate*. London: Continuum

Schecter, B., & Broughton, J. (1991). Developmental relationships between psychological metaphors and concepts of life and consciousness. *Metaphor and Symbolic Activity*, *6*, 119-143.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Schiffrin, D. (1994). *Approaches to Discourse*. Malden, Mass.: Blackwell.

Schlyter, S. 1993: The weaker language in bilingual Swedish–French children. In K. Hyltenstam & A. Viberg (Eds), *Progression and regression in language*. Cambridge: C.U.P.

Con formato: Fuente: +Cuerpo (Calibri), 12 pto

Scinto, L. M. (1986). *Written language and psychological development*. Orlando, FL: Academic Press.

Scott, C. M. (1988). Spoken and written syntax. In M. A. Nippold (Ed.). *Later Language Development: Ages nine through nineteen*. (pp. 49-95). Boston, MA: Little, Brown and Company.

Scott, M. C. (2004). Syntactic ability in children and adolescents with language and learning disabilities. In R. A. Berman (Ed.), *Language development across childhood and adolescence* (pp. 135-162). Amsterdam: John Benjamins.

Scott, C. M., & Stokes, S. L. (1995). Measures of syntax in school-age children and adolescents. *Language, Speech, and Hearing Services in Schools, 26*, 309-319.

Scott, C., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse in school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research*, *43*, 324-339.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*, 143–174.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*, 151–218.

Share, D. L. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology*, *72*, 95–129.

Share, D. L. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology*, *87*, 267–298.

Share, D.L. & Shalev, C. (2004). Self-teaching in normal and disabled readers. *Reading and Writing*, *17*, 1-31.

Slobin, D. (1985). Crosslinguistic evidence for the language making capacity. In D-I Slobin (Ed.), *The crosslinguistic study of language acquisition*. (Volume 2: Theoretical issues, pp. 1157-1256). Hillsdale, NJ: Lawrence Erlbaum.

Snow, C. E., (1990). The development of definitional skill. *Journal of Child Language*, *17,* 697–710.

Snow, C., & Uccelli, P. (2009). The challenge of academic language. In D. Olson & N.

Torrance (Eds.), *Cambridge handbook of literacy* (pp. 112-135) Cambridge, MA : Cambridge University Press.

Stahl, S. A. & Fairbanks, M. M. (1986). The Effects of Vocabulary Instruction: A Model-based Metaanalysis. *Review of Educational Research*, *56*, 72-110.

Stolt, S., Haataja, L., Lapinleimu, H. & Lehtonen, L. (2009). Associations between lexicon and grammar at the end of the second year in Finnish children. *Journal of Child Language*, *36*, 779–806.

Stromqvist, S. (1999). Production rate profiles. In S. Stronqvist & Ahlsen (Eds.), *The process of writing -- a progress report*. Gothenburg papers in theoretical linguistics 83 (pp. 53-70). Sweden, Gothenburg: Gotëborg University. Department of Linguistics.

Strömqvist, S., Nordqvist, A., & Wengelin, A. (2004). Writing the frog story: Developmental and cross-modal perspectives. In S. Strömqvist & L. Verhoeven (Eds.), *Relating events in narrative: Typological and contextual perspectives* (pp. 359-394). Mahwah, NJ: Lawrence Erlbaum.

Szmrecsányi, B. M. (2004). *On operationalizing syntactic complexity*. JADT 2004: 7eme journees internationalles d'analyse statistique des donnees textuelles.

Teberosky, A. (1992). *Aprendiendo a escribir*. Barcelona, ICE-HORSORI

Teddiman, L. (2009). Contextuality and beyond: Investigating an online diary corpus. Proceedings of the Third International ICWSM Conference.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, *10* (*1*), 1–13.

Thordardottir,E., Ellis, S. & Evans, J. (2002). Continuity in lexical and morphological development in Icelandic and English-speaking 2-year-olds. *First Language*, *22*, 3-28.

Tincoff, R. & Hauser, M. (2006). Cognitive basis for language evolution in nonhuman primates. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics*, (pp. 553-538). Amsterdam: Elsevier 2nd edition.

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: Benjamins

Tolchinsky, L. (1992). *Aprendizaje del lenguaje escrito*. Procesos evolutivos e implicaciones didácticas. Barcelona, Anthropos.

Tolchinsky, L. (2004). The nature and scope of later language development. In R. A. Berman (Ed.), *Language development across childhood and adolescence* (pp. 135-162). Amsterdam: John Benjamins.

Tolchinsky, L., & Rosado, E. (2005). The effect of literacy, text type, and modality on the use of grammatical means for agency alternation in Spanish. *Journal of Pragmatics*, *37*, 209–238.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Totereau, C., Thevenin, M. G., & Fayol, M. (1997). Acquisition de la morphologie du nombre à l'écrit en français. In L. Rieben, M. Fayol & C. A. Perfetti (Eds.), Des orthographes et leur acquisition (pp. 147-165). Lausanne: Delachaux & Niestlé.

Treiman, R., & Kessler, B. (2005). Writing systems and spelling development. In M. J. Snowing & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 120–134). Oxford: Blackwell.

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language, 28,* 649-667.

Van Valin, R. D. (2006). On the relationship between syntactic theory and models of language processing. In A. Bornkessel & M. Schlesewsky (Eds.), *Semantic Role Universals and Argument Linking: Theoretical, Typological and Psycho-/neurolinguistic Perspectives* (pp. 263–302). Mouton de Gruyter, Berlin.

Verhoeven, L, Aparici, M., Cahana-Amitai, D., van Hell, J., Kriz, S, Viguie´-Simon, A. (2002). Clause packaging in writing and speech: a cross-linguistic developmental analysis. *Written Language and Literacy*, *5*, 135–162.

Weissenborn, J. & Höhle, B. (2000). *Approaches to bootstrapping: phonological, lexical, syntactic and neurophysiological aspects of early language acquisition*. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Weizman, Z. O., & Snow, C. E. (2001). Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, *37*, 265–279.

Wimmer, G., & Altmann, G. (1996). The theory of word length distribution: Some results and generalizations. In P. Schmidt (Ed.), *Glottometrika, 15* (pp. 112-133). Trier, Germany:WVT.

Zentella, A. M. (1997). *Growing up bilingual: Puerto Rican children in New York*. Oxford: Blackwell Publishers.

Ziegler, J., Bertrand, D., To´th, D., Cse´pe, V., Reis, A., Faı´sca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*, 551–559.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled Redding across languages: a psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3-29.

APPENDICES

APPENDIX 1

Criteria for lemmatization

---

1 General criteria

a) Lemmas were assigned following the normative uses of language as sanctioned by the Dictionary of the Catalan Studies Institute (DIEC), the Dictionary of the Enciclopèdia Catalana (DEC), the Diccionari Català-Valencia-Balear by Alcover-Moll that gathers dialectal variants of Catalan as well as the TERMCAT terminological database that includes recently-created terms .

b) Expressions written in a language other than Catalan were lemmatized in the equivalent Catalan form.
labavo Spanish for *bany* 'bathroom'→ *bany* 'bathroom', 'petting' → *magrejar-se*.

c) Expressions in another language with a frequency of use beyond 200 in the corpus and produced within the Catalan or Catalan/Spanish groups were lemmatized as such
tonto → *tonto* 'silly' (for Catalan *brètol* 'silly` with very low ratio of use)

d) Expressions showing code mixing were lemmatized as such
puenting → *puenting* 'badging jumping'(from Spanish *puente* 'bridge' and English suffix 'ing'), whose Catalan equivalent should be *ponting* 'badging jumping' (from Catalan *pont* 'bridge'

e) Common use trade marks were lemmatized as such

play station → *play station*)

f) Uncompleted complex forms were lemmatized by its lexical base
Pa de 'bread of' → *pa* 'bread'

g) Indecipherable expressions were lemmatized "?". When the Indecipherable expression follows a light verb in a complex verbal construction it was lemmatized by the light verb + X.
*anar a **** 'to go to ***' → *anar a X* 'to go to X'

2. Morphological characteristics of lemmas

h) In most cases, the semantic field determines the lemma's morphological category. Thus, when possible, lemmas are nouns for Clothing and Food items, verbs for Leisure activities and adjectives for Personality traits. Only for Natural phenomena can lemmas be either nouns or verbs indistinctly.

i) Cases were found when it was not possible to assign the primed category or when similar syntactic constructions were lemmatized in slightly different ways. We refer to those cases in 4.

j) Adjectives were lemmatized in their masculine, singular form

divertides [feminine plural] 'fun'→ *divertit* [masculine singular], 'fun'

k) Double genre nouns were lemmatized in their masculine form
mico [masculine singular] 'monkey' and *mica* [feminine singular] 'monkey' were all lemmatized as *mico* [masculine singular] 'monkey'

l) Nouns that can be used in singular and plural form indistinctly were lemmatized in their singular form
cigrons [masculine plural] 'chickpeas' → *cigró* [masculine singular] 'chickpea'

m) Verbs that can be used pronominally were lemmatized in their pronominal form only when there is difference between the meanings of the pronominal form and the unpronominal form
trobar 'to find (something)' → lemma *trobar* 'to find (something)' ; *trobar-se* 'to meet (with each other)' → lemma *trobar-se* 'to meet (with each other)'

3. Semantic attribution

n) If an expression keeps no semantic relation with any of the proposed semantic fields the expression was used as such for the lemma
Gato copión → *gato copión* 'copycat'

4. Verbal multiword constructions: [V + N/ infinitive]

Determining whether a verb is full or light in a verb+noun construction can be a difficult task. The grammatical category of the semantic field can be helpful in determining when a verb is used as a support verb or as a full lexical item. Thus within Leisure activities the verb *anar* 'to go'='to go to' in *anar a Barcelona* can work as lemma whereas within the field of Clothing the same verb *anar* 'to go'='to wear' in *anar amb xancletes* 'to go in thongs'='to wear thongs' has no semantic weight and the expression was assigned *xancletes* 'thongs' as lemma.

p) For verbal multiword construction (verb+noun) simple infinitive verbal lemmas were priorized when possible. Light verbs such as *fer* 'to do/to make' / *anar* 'to go'='to go to' / *donar* 'to give' were substituted by the verb representing the conflated meaning of the light verb and the following noun:

[Light verb + N → Full verb]

fer/donar abraçades 'to make/to give hugs' = 'to hug' → *abraçar* 'to hug', *donar un vol* 'to give a walk'='to take a walk'→ *voltar* 'to take a walk'

This criterion did not apply for cases where the light verb and the noun shared the same stem without, however, keeping a strict semantic relation.

donar un tomb 'to give a walk'→ **tombar* 'to turn upside down'

q) Light verbs followed by a noun or infinitive expressing an activity were lemmatized by the activity lemma according to the following schema:

[light verb + V infinitive → V infinitive / light verb + N (activity) → N activity]

anar a nedar 'to go to swim'='to go swimming' → *nedar* 'to swim', *anar a futbol* 'to go to play football' = 'to go playing football'→ *futbol* 'football'

r) Verbs expressing a general idea of activity (such as *jugar* 'play') followed by the noun of the specific activity, were avoided in favor of the noun.

jugar a futbol 'to play football' → *futbol* 'football'

s) Expressions containing a movement verb followed by nouns referring to places were lemmatized by multiword lemmas containing the movement verb plus a variable LLOC 'PLACE':

[movement verb + N (place) → V (prep) + LLOC 'PLACE']

anar a Barcelona 'to go to Barcelona' → *anar a LLOC* 'to go to PLACE'

t) In conjunction with activity nouns, some grammatical verbs like *fer* 'to do/to make) add specific signification that must be preserved.

[light verb + N (activity) → N (activity)] but [*Fer* 'to do/to make'+ N (activity) → *Fer + N* 'to do/ to make'+ N (activity)]

fer amics 'to make friends' → *fer amics* 'to make friends'

5 Nominal multiword constructions: [(Adj+) N + (Adj/SP)]

These are nominal phrases modified either by a preceding or postponed adjective or by a prepositional phrase or by both at the same time.

u) In noun+adjective constructions, if the adjective modifies the noun without creating a new entity, the noun functions as a lemma. If however, a new entity is created, both terms are used as lemma.

molta calor 'very hot'→ *calor* 'hot', *onada forta* 'high wave'→ *onada* 'wave'

but:

fil 'thread'→*fil* 'thread' , *fil dental* 'dental thread' = 'dental floss'→ *fil dental* 'dental thread'='dental floss'

Decisions were particularly problematic due to vagueness of the referent and to the difficulty of establishing the classificatory or qualifying function of the adjective. For instance, *ulleres de sol* 'sunglasses', *ulleres* 'glasses' and *ulleres de submarinisme* 'goggles' show little ambiguity in their referents and can be considered as three different lemmas. Cases like *faldilla* 'skirt' and *faldilla curta* 'short skirt'='miniskirt' show more blurred referential limits that can result in an inconsistent lemmatization.

Table 6. Stepwise regression analysis for 2nd grade mean clause length

2nd grade

joke telling			B	SE	beta	R2	ΔR2
	step 1					.061	.004
		text length	-.007	.015	-.061		
	step 2					.274	0.08*
		text length	.038	.026	.334		
		vocabulary size	6.414	3.058	.477		
	step 3					.573	0.33**
		text length	.046	.023	.402		
		vocabulary size	7.352	2.790	.547		
		word length	-1.078	.453	-.306		
		nominalizations	-.042	.067	-.074		
		adjectives	.312	.424	.084		
		formality	.099	.023	.534		
recommendation							
	step 1					.322	0.104*
		text length	.063	.024	.322		
	step 2					.364	.133
		text length	.028	.035	.141		
		vocabulary size	-3.378	2.450	-.248		
	step 3					.627	0.394**
		text length	.021	.033	.107		
		vocabulary size	-2.972	2.190	-.218		
		word length	-.388	.142	-.317		
		nominalizations	.063	.044	.157		
		adjectives	.028	.255	.013		
		formality	.045	.011	.468		
explanation							
	step 1					.230	.053

	text length	.041	.023	.230		
step 2					.251	.063
	text length	.060	.033	.338		
	vocabulary size	2.556	3.268	.148		
step 3					.277	.077
	text length	.061	.035	.345		
	vocabulary size	2.428	3.598	.140		
	word length	.083	.425	.028		
	nominalizations	-.032	.049	-.088		
	adjectives	.236	.408	.077		
	formality	.006	.022	.038		

Table 7. Stepwise regression analysis for 2nd grade noun phrase

2nd grade

joke telling		B	SE	beta	R2	AR2
step 1					.115	.013
	text length	.001	.001	.115		
step 2		-.100	.241		.145	.021
	text length	.002	.002	.246		
	vocabulary size	.163	.241	.158		
step 3					.294	.086
	text length	.003	.002	.322		
	vocabulary size	.280	.249	.272		
	word length	-.046	.040	-.170		
	nominalizations	-.007	.006	-.172		
	adjectives	-.036	.038	-.127		
	formality	.001	.002	.043		
recommendation						
step 1					.180	.032
	text length	.006	.005	.180		
step 2					.180	.033
	text length	.007	.007	.188		

		vocabulary size	.027	.464	.011		
	step 3					.456	0.208*
		text length	.009	.007	.248		
		vocabulary size	.095	.449	.039		
		word length	.007	.029	.034		
		nominalizations	.003	.009	.037		
		adjectives	-.013	.052	-.033		
		formality	.007	.002	.398		
explanation							
	step 1					.051	.003
		text length	.002	.004	.051		
	step 2					.101	.010
		text length	-.001	.006	-.043		
		vocabulary size	-.391	.589	-.129		
	step 3					.271	.074
		text length	.000	.006	.015		
		vocabulary size	-.202	.632	-.067		
		word length	-.068	.075	-.131		
		nominalizations	.005	.009	.072		
		adjectives	.105	.072	.196		
		formality	.002	.004	.063		

Table 8. Stepwise regression analysis for 2nd grade clause complexity

2nd grade

joke telling		B	SE	beta	R2	AR2
	step 1				0.01	0
		text length	.000	.001	-.010		
	step 2				0.06	0.004
		text length	.000	.001	-.097		
		vocabulary size	-.058	.130	-.105		
	step 3				0.389	0.152
		text length	-.001	.001	-.130		

		vocabulary size	-.050	.128	-.092		
		word length	.009	.021	.063		
		nominalizations	-.001	.003	-.032		
		adjectives	.008	.019	.051		
		formality	-.003	.001	-.404		
recommendation							
	step 1					.172	.029
		text length	.004	.003	.172		
	step 2					.173	.030
		text length	.005	.004	.191		
		vocabulary size	.043	.312	.027		
	step 3					.287	.082
		text length	.003	.005	.135		
		vocabulary size	-.039	.324	-.024		
		word length	.020	.021	.136		
		nominalizations	.001	.007	.017		
		adjectives	.029	.038	.112		
		formality	-.002	.002	-.168		
explanation							
	step 1					.110	.012
		text length	.001	.001	.110		
	step 2					.137	.019
		text length	.002	.002	.199		
		vocabulary size	.124	.198	.121		
	step 3					.486	0.237*
		text length	.001	.002	.062		
		vocabulary size	.011	.193	.011		
		word length	.055	.023	.314		
		nominalizations	.004	.003	.200		
		adjectives	.006	.022	.031		
		formality	-.004	.001	-.415		

Table 9. Stepwise regression analysis for 6th grade mean clause length

6th grade

joke telling			B	SE	Beta	R2	ΔR2
	step 1					,003	,000
		text length	,000	,012	,003		
	step 2					,031	,001
		text length	,004	,019	,042		
		vocabulary size	,687	2,957	,050		
	step 3					,481	0,232**
		text length	,021	,019	,229		
		vocabulary size	3,861	2,840	,281		
		word length	1,551	,626	,357		
		nominalizations	-,117	,098	-,163		
		adjectives	-,051	,407	-,017		
		formality	,062	,026	,321		
recommendation							
	step 1					,253	,064
		text length	,045	,023	,253		
	step 2					,280	,078
		text length	,029	,028	,162		
		vocabulary size	-2,635	2,777	-,151		
	step 3					,705	0,496**
		text length	,016	,026	,088		
		vocabulary size	-,348	2,170	-,020		
		word length	-,367	,288	-,133		
		nominalizations	,111	,079	,153		
		adjectives	-,281	,230	-,154		
		formality	,086	,014	,644		
explanation							
	step 1					,014	,000

	text length	-,002	,017	-,014		
step 2					,215	,046
	text length	-,038	,028	-,293		
	vocabulary size	-8,738	5,262	-,352		
step 3					,594	0,353**
	text length	-,024	,024	-,185		
	vocabulary size	-6,200	4,549	-,250		
	word length	-,172	,893	-,023		
	nominalizations	,014	,102	,015		
	adjectives	-,287	,379	-,088		
	formality	,169	,035	,558		

Table 10. Stepwise regression analysis for 6th grade noun phrase

6th grade

joke telling		B	SE	beta	R2	ΔR2
	step 1				,105	,011
	text length	-,001	,002	-,105		
	step 2				,105	,011
	text length	-,001	,002	-,098		
	vocabulary size	,017	,375	,010		
	step 3				,433	0,188*
	text length	-,001	,002	-,110		
	vocabulary size	-,236	,372	-,135		
	word length	-,041	,082	-,074		
	nominalizations	,040	,013	,434		
	adjectives	-,065	,053	-,173		
	formality	-,002	,003	-,097		
recommendation						
	step 1				,224	,050

		text length	,006	,003	,224	
	step 2					,227 ,051
		text length	,005	,004	,198	
		vocabulary size	-,113	,414	-,044	
	step 3					,453 0,205*
		text length	,008	,005	,293	
		vocabulary size	,018	,401	,007	
		word length	-,021	,053	-,052	
		nominalizations	,029	,015	,277	
		adjectives	-,078	,042	-,292	
		formality	,001	,003	,048	
explanation						
	step 1					,142 ,020
		text length	,002	,002	,142	
	step 2					,169 ,028
		text length	,000	,003	,025	
		vocabulary size	-,346	,500	-,148	
	step 3					,427 ,182
		text length	,000	,003	-,021	
		vocabulary size	-,347	,481	-,149	
		word length	-,156	,094	-,219	
		nominalizations	,018	,011	,211	
		adjectives	,092	,040	,300	
		formality	,003	,004	,093	

Table 11. Stepwise regression analysis for 6th grade clause complexity

6th grade

joke telling			B	SE	beta	R2	ΔR2
	step 1					0,267	0,071*
		text length	,001	,000	,267		
	step 2					0,279	0,078
		text length	,001	,001	,164		
		vocabulary size	-,070	,111	-,131		
	step 3					0,329	0,108
		text length	,001	,001	,256		
		vocabulary size	-,034	,119	-,064		
		word length	,014	,026	,084		
		nominalizations	-,003	,004	-,096		
		adjectives	-,010	,017	-,085		
		formality	,001	,001	,150		
recommendation							
	step 1					,061	,004
		text length	,001	,002	,061		
	step 2					,163	,027
		text length	,003	,003	,174		
		vocabulary size	,290	,251	,189		
	step 3					,367	,135
		text length	,007	,003	,437		
		vocabulary size	,220	,250	,143		
		word length	,047	,033	,195		
		nominalizations	-,012	,009	-,188		
		adjectives	-,053	,026	-,332		
		formality	-,001	,002	-,120		
explanation							
	step 1					,100	,010
		text length	,001	,001	,100		
	step 2					,144	,021
		text length	,000	,001	-,035		
		vocabulary size	-,204	,258	-,170		

	step 3				,335	,112
		text length	,000	,001	-,064	
		vocabulary size	-,275	,258	-,229	
		word length	,036	,051	,098	
		nominalizations	,005	,006	,112	
		adjectives	-,003	,021	-,020	
		formality	-,004	,002	-,251	

Table 12. Stepwise regression analysis for 10th grade mean clause length

10th grade

joke telling			B	SE	beta	R2	ΔR2
	step 1					,114	,013
		text length	,013	,015	,114		
	step 2					,191	,036
		text length	,044	,030	,379		
		vocabulary size	4,485	3,825	,305		
	step 3					,357	,128
		text length	,039	,032	,334		
		vocabulary size	3,813	3,861	,259		
		word length	,056	,467	,018		
		nominalizations	-,098	,090	-,154		
		adjectives	,220	,394	,082		
		formality	,051	,026	,279		
recommendation							
	step 1					,216	,047
		text length	,018	,011	,216		
	step 2					,216	,047
		text length	,018	,017	,207		
		vocabulary size	-,146	2,475	-,012		
	step 3					,542	0,293*
		text length	-,011	,018	-,135		

		vocabulary size	-1,717	2,290	-,137		
		word length	,307	,374	,109		
		nominalizations	,119	,058	,285		
		adjectives	,073	,147	,077		
		formality	,031	,018	,264		
explanation							
	step 1					,111	,012
		text length	,007	,008	,111		
	step 2					,250	,063
		text length	-,015	,014	-,243		
		vocabulary size	-5,288	3,027	-,419		
	step 3					,373	,139
		text length	-,016	,015	-,275		
		vocabulary size	-6,652	3,269	-,527		
		word length	,397	,539	,110		
		nominalizations	-,027	,062	-,066		
		adjectives	-,164	,133	-,164		
		formality	,032	,018	,229		

Table 13. Stepwise regression analysis for 10th grade noun phrase

10 grade

joke telling			B	SE	beta	R2	ΔR2
	step 1					,047	,002
		text length	-,001	,002	-,047		
	step 2					,064	,004
		text length	,000	,004	,027		
		vocabulary size	,155	,479	,085		
	step 3					,249	,062
		text length	,001	,004	,077		
		vocabulary size	,182	,493	,101		
		word length	,040	,060	,102		

		nominalizations	,002	,012	,026		
		adjectives	-,037	,050	-,112		
		formality	,004	,003	,174		
recommendation							
	step 1					,073	,005
		text length	-,002	,003	-,073		
	step 2					,084	,007
		text length	-,001	,004	-,023		
		vocabulary size	,210	,650	,065		
	step 3					,284	,081
		text length	-,003	,005	-,136		
		vocabulary size	,045	,672	,014		
		word length	,061	,110	,084		
		nominalizations	-,005	,017	-,049		
		adjectives	-,016	,043	-,064		
		formality	,009	,005	,300		
explanation							
	step 1					,014	,000
		text length	,000	,001	,014		
	step 2					,322	0,104*2
		text length	,006	,003	,522		
		vocabulary size	1,382	,539	,601		
	step 3					,345	,119
		text length	,006	,003	,582		
		vocabulary size	1,558	,602	,678		
		word length	-,077	,099	-,118		
		nominalizations	,001	,011	,007		
		adjectives	,005	,025	,027		
		formality	-,001	,003	-,056		

Table 13. Stepwise regression analysis for 10th grade clause complexity

10th grade

joke telling			B	SE	beta	R2	ΔR2
	step 1	(Constante)				0,006	0
		text length	,000	,001	,006		
	step 2	(Constante)				0,199	0,04
		text length	,003	,002	,351		
		vocabulary size	,444	,290	,398		
	step 3	(Constante)				0,559	0,312**
		text length	,000	,002	,019		
		vocabulary size	,254	,260	,228		
		word length	,115	,032	,479		
		nominalizations	-,001	,006	-,011		
		adjectives	,059	,027	,287		
		formality	-,003	,002	-,187		
recommendation							
	step 1	(Constante)				,070	,005
		text length	-,001	,001	-,070		
	step 2	(Constante)				,138	,019
		text length	,001	,002	,068		
		vocabulary size	,271	,299	,182		
	step 3	(Constante)				,240	,058
		text length	,000	,002	-,009		
		vocabulary size	,304	,314	,205		
		word length	-,041	,051	-,122		
		nominalizations	-,003	,008	-,064		
		adjectives	,009	,020	,081		
		formality	,003	,002	,187		
explanation							
	step 1					,057	,003
		text length	,000	,001	-,057		
	step 2					,096	,009
		text length	-,001	,001	-,178		

	vocabulary size	-,175	,299	-,144		
step 3					,259	,067
	text length	-,001	,002	-,209		
	vocabulary size	-,159	,326	-,132		
	word length	,020	,054	,059		
	nominalizations	,002	,006	,057		
	adjectives	,004	,013	,037		
	formality	-,003	,002	-,238		