NOTE:

This article was published in the Journal of Immersion and Content-Based Language Education (Doi: 10.1075/jicb.19022.agu). The printed version is accessible at:

https://www.jbe-platform.com/content/journals/10.1075/jicb.19022.agu

Are CLIL texts too complicated?

A computational analysis of their linguistic characteristics

Amaia Aguirregoitia Martinez, Kepa Bengoetxea Kortazar, & Itziar Gonzalez-Dios
University of the Basque Country (UPV/EHU)

This article presents a comparative study of the linguistic characteristics of some materials used to teach English as a foreign language, and Geography and History through English in a Content and Language Integrated Learning (CLIL) experience in the Basque Country with students aged 11–13. This paper analyzes and compares the contents of current textbooks using CohMetrix and AzterTest, which calculate stylistic and linguistic metrics regarding lexical and grammatical complexity, readability and coherence. Finally, the study suggests that there are significant differences mainly in vocabulary level, narrativity and cohesion, it identifies the potential difficulties of CLIL texts and offers advice on how to overcome them. Raising awareness of the complexity of some texts used in CLIL can provide a starting point for pedagogical adaptations and contribute to optimizing learning.

Keywords: CLIL, multilingual education, second language teaching, primary education, secondary education

1. Introduction

Practices of multilingual education have been implemented in the past (Franceschini, 2013), but only those belonging to certain high social status had access to it. Nowadays, societies are conscious of the importance of foreign language learning and, therefore, education systems all over the world are being adapted in an attempt to improve their students' foreign language proficiency. CLIL practices have emerged from the aspiration to enhance multilingualism. Based on socio-cultural and second language acquisition theories, it is considered that CLIL enables students to acquire an L2 through increased exposure to and

engagement with the L2, with content subjects providing the context and motivation for authentic, purposeful communication (Lyster & Ruiz de Zarobe, 2018).

Bilingual models were established in the education system of the Basque Autonomous Community (BAC, henceforth) more than 30 years ago and nowadays this system is committed to the development of a trilingual education system (L1 + L2) + L3 system where Basque and Spanish languages are used as official languages for teaching whereas English is introduced at an early age. Nowadays, 78.04% of the students of the compulsory education levels in the BAC study in what is called a D model (Eusko Jaurlaritza, 2018), an immersion program where Basque is used as a vehicular language for teaching (Cenoz, 2009). English has been introduced from an early age at schools and it is compulsory at primary education aligned with the guidelines of the Council of Europe (Beacco & Byram, 2007).

Even if it is assumed that every person has the capability of being multilingual at a different development level, this does not mean that similar competences will be achieved in all the languages. There is a general agreement that students using a language to learn content will develop different competences from those studying that language as a subject (Cummins, 2009). With the idea of "use language to learn and learn to use language", the Basque Government defined in 2010 a framework program (Eusko Jaurlaritza, 2010) to coordinate and evaluate the experimental introduction of CLIL at Primary and Secondary levels, which generated diverse practices at different schools. Both state and private schools have been introducing CLIL to teach mostly content in the subjects of Science, Geography and History or Art and Crafts in recent years. It involves a process that is generally curriculum-driven with the language curriculum arising from the content curriculum (Gierlinger, 2017).

The best known feature of the educational system in the BAC is multilingualism, and the Basque educational system pays particular attention to the development of an integrated syllabus of the three languages. As Cenoz (2009) explains, the results of multilingual programmes are associated with diverse educational variables and the sociolinguistic or socioeconomic context. Previous research has proved that well-implemented bilingual programs are highly effective in developing strong L2 skills at no cost either to students' abilities in the dominant language or their knowledge of curriculum content taught through the L2 (Cummins 2008; Huguet et al., 2008). Also, Lopez and Sichra (2008, p. 5) conclude that "findings from different countries provide empirical evidence related to indigenous bilingual children's overall academic achievement, active participation in learning and development of positive self-image, self-esteem and respect".

As we have mentioned, many schools in the BAC are introducing CLIL programs in an attempt to develop multilingualism, but introducing a foreign lan-

guage is not always necessarily positive. As Ruiz de Zarobe (2013) points out, the need to communicate in a foreign language may cause anxiety and it may have adverse effects on communication and self-esteem as well as detrimental consequences on content learning. Moreover, Heras and Lasagabaster (2015, p. 71) indicate that "we urgently need to find more efficient ways to teach and learn FLs and this is one of the major reasons why attention is being paid to content and language integrated learning (CLIL)".

There are many pedagogical aspects for a CLIL program to be successful (Meyer, 2010). In fact, CLIL has attracted attention of many fields such as Second Language Acquisition (SLA), Systemic Functional Linguistics, Discourse Analysis or Sociolinguistics, which represent different understandings of CLIL (Llinares & Whittaker, 2017). By integrating different approaches, Llinares et al. (2012) justify and describes the relevance of language in CLIL and the roles of language in representing the academic subject and organizing the social world of the classrooms.

Schleppegrell (2004) defines "the language of schooling", and she reflects on the many challenges that must be faced, particularly by non-native English speakers. Textbook selection is one of them, and consequently, the research presented in this paper is focused on analyzing the linguistic characteristics of a set of textbook excerpts. This empirical evidence will be used to clarify whether these materials used in the programs in the BAC are appropriate. In order to address this question, we analyze texts from the CLIL (Social Science, henceforth SS, Geography and History) and ESL (English as a Second Language) subjects based on the linguistic characteristics and readability, which have been a starting point to assess the complexity of texts of different genres, subjects and levels (Crossley et al. 2007; Crossley et al. 2008; Crossley et al. 2011; McNamara et al. 2014).

Our study aims at answering three main questions: (1) What are the main differences between the texts used in CLIL and the texts in Touchstone Applied Science Association (TASA) corpus (Zeno et al., 1995) of the Social Science subject for anglophones. (2) Are there linguistic differences between CLIL texts and ESL texts? (3) If so, can we use these characteristics to define practices to promote optimal comprehension in a real experience? To that end, we have used two tools to calculate a complete set of the most relevant linguistic metrics.

2. Method

2.1 Participants

For this research, we have analyzed materials used for instruction in a specific school of the BAC. This school has more than 2,000 students aged 2 to 18 years

and it adopts the immersion program in Basque. Students study English from age 4, but English as a vehicular language is introduced optionally from the 3rd grade of primary education. English is used to teach SS, and it continues as an optional vehicular language for that subject in the subsequent courses. However, many students go back to the instruction in Basque after their first year and attribute their difficulties to learning in a foreign language.

First of all, we analyzed some excerpts from the edited textbooks used to teach SS/History and ESL in 6th grade of primary education and 1st and 2nd grade of secondary education (11–13 years). These edited books are sold in Spain by a widely-known publishing house and have been selected by the school board. In Table 1 we present the number of excerpts, their word-length, subject and grade.

Table 1. Texts analyzed in the research by grade, number of texts and the number of words

Subject	Grade	Number of texts/units	Number of words
Social Science	6th Primary	3	4,672
ESL	6th Primary	18	3,823
Geography and History	1st Secondary	3	6,720
ESL	1st Secondary	10	2,517
History	2nd Secondary	3	6,473
ESL	2nd Secondary	9	3,019
TOTAL		**46**	**27,224**

2.2 Characteristics of genres in History

Measures should be interpreted on the basis of the genre characteristics. Particularly, genres in CLIL have previously been analyzed and described by Llinares et al. (2012) and Lorenzo (2013). Llinares and Whittaker (2010) analyzed the language used by CLIL secondary school students of History and that of students following the same syllabus in Spanish. History is a subject in which the role of language is especially present and students transit to cognitive academic language proficiency (CALP) with texts that tend to use descriptions of physical features, definitions of terms related to phenomena and typically they include sequential, factorial, causal and consequential explanations. Genres in History typically include a chronological structure of the narrations, descriptions of different aspects of society and explanations about the way things were at a particular time. These texts also use logically structured explanations of temporal and causal relations of historical events and cultures, they use specific terminology and they sometimes include historical arguments and discussions. More specific details can

be found in the list of genres for CLIL history courses defined by Lorenzo (2013), who has also conducted specific research on the evolution of the students' development of the academic language when writing about historical facts in English as L2 (Lorenzo et al., 2019). It might be expected that history texts, as a consequence of using academic language, will score high at nominalizations, cause-effect relations between clauses, lexical development, syntactic complexity or hypernymy and lower in familiarity or word concreteness (Lorenzo et al., 2019).

Leaving apart the specificities, when evaluating the results it should be considered that language in any CLIL text should promote the development of academic language skills, which means not being too poor to teach any content. It has also been noted that sometimes, unrealistic and unnecessarily simple language has resulted from simplification of L2 texts and language-adjustment techniques (Lorenzo, 2008). Convenient integration of language and content ultimately affects what students will learn, and subsequently, it is crucial that the appropriateness of the texts and their linguistic measures are evaluated without overlooking the specific characteristics and challenges of the genre.

2.3 Tools

Previous research used the computational tool Coh-Metrix (Graesser et al., 2014) to provide a better understanding of the linguistic characteristics of texts in the field of second language materials (Crossley et al., 2007). Coh-Metrix measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis. For our study we also considered the published table of index norms of school (K-12) and college grade level texts from Social Studies, Science and Language Arts genres (McNamara et al., 2014). The norms were calculated by comparing the TASA corpus of over 37,000 texts ranging in genre and school grade and college levels which facilitates the interpretation and comparison of collected data.

Coh-Metrix provides eight principal components (PCs) following the guidelines outlined by Der and Everitt (2008) in the form of Z-scores and percentile scores: narrativity (PC1), referential cohesion (PC2), syntactic simplicity (PC3), word concreteness (PC4), causal cohesion (PC5), verb cohesion (PC6), logical cohesion (PC7) and temporal cohesion (PC8). For each of these PCs, Z-scores and percentile scores are calculated. An analysis of PCs was conducted, yielding eight components that explained a striking 67.3% of the variability among the texts of the TASA corpus; the top five components explained over 50% of the variance. Most importantly, the components aligned with the language-discourse levels previously proposed in multilevel theoretical frameworks of cognition and comprehension (Kintsch, 1998; Snow, 2002; Graesser et al., 2011). The five main

linguistic dimensions are currently being used to analyze texts in K-12 for the Common Core literacy standards in the USA (CCSSONGA, 2010):

Narrativity (PC1).	Narrative text tells a story, with characters, events, places, and things familiar to the reader. This robust component is highly affiliated with word familiarity, world knowledge, and oral language. Informational expository texts on less familiar topics would lie at the opposite end of the continuum.
Referential cohesion (PC2).	This component includes Coh-Metrix indices that assess referential cohesion. High-cohesion texts contain words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader. Low-cohesion texts are typically more difficult to process because there are fewer threads that tie the ideas together for the reader. CLIL texts will only be understood with an appropriate level of coherence and cohesion (Whittaker et al., 2011).
Syntactic simplicity (PC3).	This component reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures, which are less challenging to process.
Word concreteness (PC4).	Texts that contain concrete and meaningful content words, which evoke mental images are easier to process and understand.
Causal cohesion (PC5).	This dimension reflects the degree to which the text contains causal, intentional, and temporal connectives and conceptual links. These connectives help the reader to form a more coherent and deeper understanding of the causal events, processes, and actions in the text.

Although this set of metrics offers relevant and contrasted information about the linguistic features of the texts, we also used AzterTest[1] (Bengoetxea et al., 2020), an open source linguistic and stylistic analysis tool. AzterTest calculates 164 measures[2] using the latest advances in Natural Language Processing (NLP) tools and linguistic resources, such as the parsers NLP-Cube (0.1.0.7) (Boros et al., 2018) and Stanford NLP neural pipeline (0.2.0) (Qi et al., 2018), wordfreq (Speer et al., 2018), FastText embeddings (Mikolov et al., 2018) and Oxford 5000 vocabulary list,[3] based on the Common European Framework of Reference for Languages

1. The application can be tested at http://161.35.202.53/ and the code is freely available at https://github.com/kepaxabier/AzterTest

2. Visit http://161.35.202.53/information.html for more complete information

3. https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000

(CEFR) standard. The validity of AzterTest has been described in Bengoetxea et al. (2020) where this tool was tested in a readability assessment scenario for English texts, outperforming the results obtained with Coh-Metrix's output.

We analyzed a corpus of 46 texts (see Table 1 above) using both tools to obtain all their linguistic metrics. First, we calculated all the metrics using Coh-Metrix. Second, we calculated the metrics with AzterTest. Third, we analyzed the results of both tools and we used the Wilcoxon signed rank test (also called the Mann-Whitney U test) to find significant differences between ESL and CLIL texts.

3. Results

In this section, we present the differences between CLIL texts and texts from the TASA corpus (RQ1); the differences between CLIL and ESL texts in different levels considering the main five Text Easability PCs obtained by Coh-Metrix (RQ2) and the linguistic features obtained by Coh-Metrix and AzterTest (RQ3) to get a more fine-grained analysis.

3.1 Analysis of principal components in CLIL texts and texts for native English speakers

Table 2 presents the main five Text Easability PCs with the percentile score and compares the texts of this study with TASA SS texts used by native English speaking students of the same ages. Percentile score (p-score) varies from 0 to 100, higher scores meaning the text is likely to be easier to read than other texts in the corpus. In Appendix Table A.1 we show Z-score values.

Some conclusions can be drawn from the measures obtained. In comparison with the samples from TASA the 6–8 grade-level band in SS, it is observed that:

a. CLIL texts score lower at narrativity, which points out that they might be even more difficult for the reader than texts aimed at native speakers of English.
b. Texts aimed at native speakers of English in the corpus show more referential cohesion than those of 6th grade of Primary and 2nd grade of Secondary. Furthermore, the difference is greater with the lowest level, 6th grade of Primary, where, according to O'Reilly and McNamara (2007), a greater cohesion is usually necessary.
c. Regarding syntactic simplicity, CLIL texts are simpler than the texts for native English speakers.
d. It is worth remarking that texts from 6th and 1st include a higher value of concrete words, which are easier to process and understand.

Table 2. Text easability PCs in SS genre (Percentile scores)

Text easability PCs	TASA (native) social science 6th–8th	CLIL texts (Social science)		
		6th grade (Prim.)	1st grade (Sec.)	2nd grade (Sec.)
	M	M	M	M
Narrativity	33.42	11.26	13.36	7.15
Referential cohesion	41.25	18.12	53.42	24.95
Syntactic simplicity	63.18	70.36	75.60	71.63
Word concreteness	65.56	65.67	72.74	65.28
Causal cohesion	55.28	38.05	48.29	51.79

Note: PC=Principal Component; M=Mean

e. Regarding causal cohesion, texts for native English speaking children have a higher causal cohesion than CLIL texts, particularly in the 6th grade.

It might be concluded that CLIL texts are comparatively simpler syntactically but narrativity and referential and causal cohesion might be improved.

3.2 Analysis of principal components in ESL and CLIL texts

This section presents the comparison of the English and CLIL texts regarding the five Text Easability PCs (Table 3) in percentile score (The Z-scores are given in Table A.2.).

CLIL texts are significantly less narrative in all the degrees, meaning that the vocabulary is less related to world knowledge and oral language. Moreover, they use less concrete words and have more causal cohesion in the three grades than ESL texts. The referential cohesion and syntactic simplicity of CLIL texts are lower in 6th and 2nd grades. This indicates that CLIL texts (a) include language which is less affiliated to everyday conversations and more difficult to the reader, which might be associated to the genre and to the world and domain knowledge; (b) use less referential cohesive devices in two of the grades; (c) are syntactically more challenging, which will be analyzed in the syntactic complexity subsection; (d) contain more abstract words that can be related to the domain knowledge; and (e) in line with expectations, CLIL texts require the reader to form a more

Table 3. Coh-metrix easability PCs in ESL and CLIL texts (Percentile scores)

	6th grade of primary			1st grade of secondary			2nd grade of secondary		
Text easability PCs	**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**	
	M	**M**	***p***	**M**	**M**	***p***	**M**	**M**	***p***
Narrativity	**43.23**	**11.26**	**0.015**	**59.89**	**13.36**	**0.006**	**54.34**	**7.15**	**0.016**
Referential cohesion	**58.17**	**18.12**	**0.001**	46.99	53.42	0.811	30.02	24.95	0.711
Syntactic simplicity	73.99	70.36	0.450	63.87	75.60	0.370	74.65	71.63	0.853
Word concreteness	73.81	65.67	0.339	73.95	72.74	1	83.60	65.28	0.194
Causal cohesion	25.05	38.05	0.184	41.63	48.29	1	52.16	51.79	0.863

Note: PC=Principal Component; M=Mean; $p = p$ value

coherent and deeper understanding of the causal events, processes, and actions in the text in a higher degree than English texts.

This PC analysis offers an overview of the main characteristics of the texts, but does not get into a fine-grained level. In the following section, we study in depth the linguistic and stylistic features of the texts.

3.3 Differences between ESL and CLIL texts

To obtain a broader view of the linguistic characteristics, we selected a set of metrics from Coh-Metrix and AzterTest including descriptive metrics, lexical diversity, word information, polysemy and hypernymy, morphological information, syntactic complexity, referential cohesion, connectives, and causal/intentional cohesion. We analyzed them for significant differences between CLIL and ESL at $p < .05$ level, calculated the correlation (Pearson) among measures and grouped those very highly correlated (0.9 or −0.9). Next subsections describe these features.

3.3.1 *Descriptive metrics*

Descriptive metrics are used for initial assessment of text complexity, although not without controversy (Crossley et al., 2007). Concerning word length (Table 4), words are significantly longer in SS texts in all the grades. In general, shorter words are easier to read and word length serves as a common proxy for

word frequency. In addition, we can see in Table 4 the average number of words in each sentence within the text without stopwords, which is significantly higher in CLIL texts than in the English texts in all the grades.

Table 4. Descriptive metrics

	6th grade of primary			1st grade of secondary			2nd grade of secondary		
	ESL texts	CLIL texts		ESL texts	CLIL texts		ESL texts	CLIL texts	
Descriptive metrics	**M**	**M**	***p***	**M**	**M**	***p***	**M**	**M**	***p***
Letters in words without stopwords	**5.57**	**6.22**	**0.030**	**5.68**	**6.23**	**0.034**	**5.60**	**6.70**	**0.009**
Words per sentence without stopwords	**3.93**	**5.36**	**0.010**	**4.56**	**6.18**	**0.033**	**4.27**	**6.99**	**0.009**

Note: PC=Principal Component; M=Mean; *p*=*p* value

3.3.2 *Lexical diversity*

Regarding lexical diversity metrics (Table 5), we focus the analysis on Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005) and *vocd* (McCarthy & Jarvis, 2007) metrics, which are analogous to the well-known type-token ratio but overcome the potential confound of text length. These metrics indicate that the vocabulary in ESL texts is significantly less diverse than CLIL texts in 6th grade. However, in the higher levels, the mean values for lexical diversity are greater in ESL texts, but there is no significant difference.

Table 5. Lexical diversity

	6th grade of primary			1st grade of secondary			2nd grade of secondary		
	ESL texts	CLIL texts		ESL texts	CLIL texts		ESL texts	CLIL texts	
Lexical diversity	**M**	**M**	***p***	**M**	**M**	***p***	**M**	**M**	***p***
vocd*	**61.29**	**79.02**	**0.046**	78.00	74.02	0.575	89.08	78.24	0.209
MTLD	47.33	57.71	0.306	66.81	53.50	0.286	76.07	65.06	0.145

Note:
* Measures obtained by Coh-Metrix; M=Mean; *p*=*p* value

3.3.3 *Word information*

Regarding word information, AzterTest calculates several metrics related to vocabulary knowledge. In this paper, we have selected the incidence of A1, A2, B1, B2 and C1 level words and the number of rare content words with a word frequency below 4.

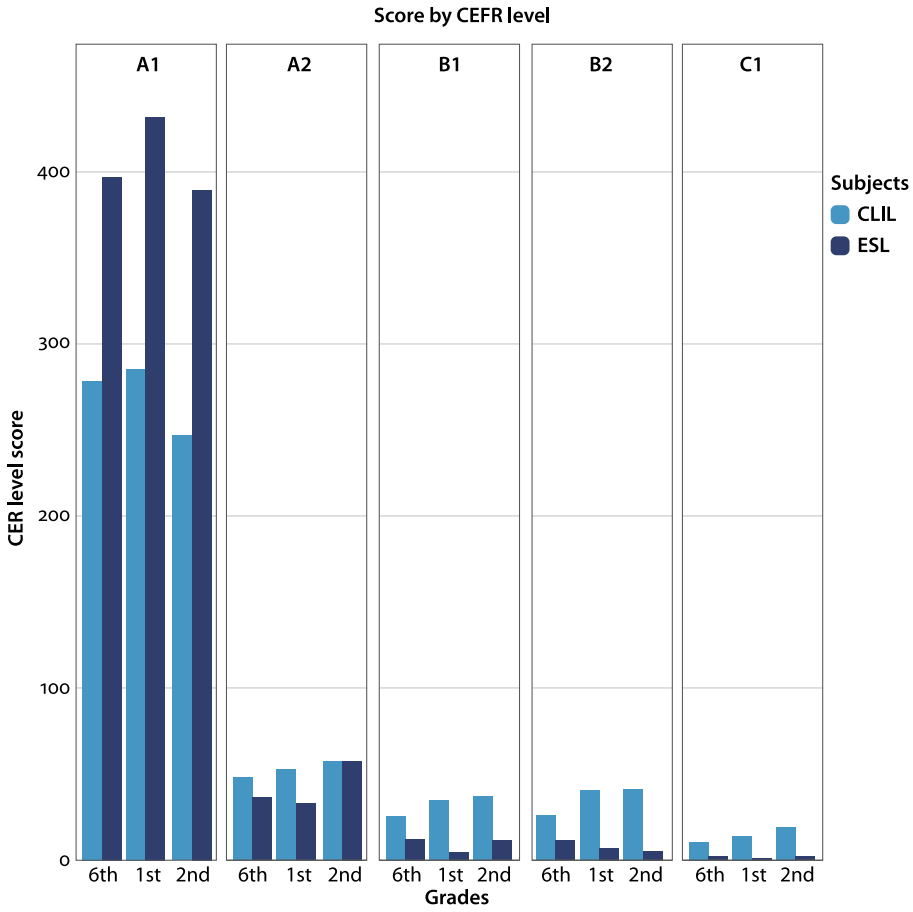

Figure 1. Distribution of words by CEFR level

Regarding the distribution of words according to CEFR levels (Figure 1), the incidence of A1 words is higher in ESL texts. The incidence of B1 and higher-level words, which can be considered intermediate and advanced difficulty, is superior in CLIL texts. All the differences between ESL and CLIL texts differ significantly in all the degrees except for the A2 level, where the difference is significant only in 2nd degree.

Regarding rare words (Table 6), rare adjective incidence results demonstrate that History texts of every grade include a higher number of adjectives that are less frequent in everyday speech, and therefore, pedagogical issues should be considered to enhance understanding the additional information provided by these adjectives. We can find adjectives as *"fallow, feudal, uncultivated"* in CLIL texts, which are less common than *"conscious, clean, polluted"*, in the ESL texts. Similarly, rare adverb incidence is significantly higher in the first two grades and the mean value is higher in all of them. We find examples as *"respectively"* or *"particularly"* in CLIL texts, in contrast to English text examples such as *"really"* or *"usually"*. Indeed, the incidence of rare content words supports the idea that SS texts use more difficult words which might require an additional comprehension effort.

Table 6. Word information

	6th grade of primary			**1st grade of secondary**			**2nd grade of secondary**		
Word information	**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**	
	M	**M**	***p***	**M**	**M**	***p***	**M**	**M**	***p***
Number of rare nouns (i)	57.87	68.85	0.533	43.67	71.06	0.111	41.60	67.49	0.100
Number of rare adjectives (i)	11.51	18.22	0.244	**5.46**	**25.97**	**0.033**	10.00	18.09	0.209
Number of rare verbs (i)	**5.35**	**13.31**	**0.019**	**6.35**	**18.15**	**0.021**	**7.05**	**21.73**	**0.063**
Number of rare adverbs (i)	**0.23**	**0.37**	**0.013**	**0.28**	**0.90**	**0.017**	0.69	0.98	0.148
Number of rare content words (i)	74.96	100.76	0.221	**55.78**	**116.10**	**0.006**	**59.35**	**108.31**	**0.036**

Note: M=Mean; *p*=*p* value; i=incidence per 1000 words

These results indicate that vocabulary is exceptionally important for understanding CLIL texts, because they include less common words than ESL texts and those words will appear repeatedly in the texts. Evidently, less familiar words in SS texts require additional effort and time to process.

3.3.4 *Polysemy and hypernymy*

For semantic information, AzterTest calculates the polysemy and hypernymy measures. In Table 7 we focus on the polysemy of nouns and verbs and hypernymy of nouns.

Table 7. Polysemy and hypernymy

	6th grade of primary			1st grade of secondary			2nd grade of secondary		
Polysemy and hypernymy	**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**	
	M	**M**	***p***	**M**	**M**	***p***	**M**	**M**	***p***
Polysemy of nouns and verbs (m)	**7.75**	**5.77**	**0.030**	7.33	6.66	0.286	**7.00**	**5.94**	**0.009**
Hypernymy of nouns (m)	**7.80**	**7.06**	**0.010**	7.65	6.85	0.160	7.36	7.22	1

Note: M=Mean; $p = p$ value; m=mean value

Regarding polysemy, we found significant differences between CLIL and ESL texts of 6th and 2nd grade and concerning hypernymy for nouns, significant differences were observed in 6th grade. A lower value for polysemy is an indicator of word concreteness. This measure scores less in SS and History texts, where we find words like "bourgeoisie", which are typically more specific and conceptually more difficult words. Hypernym level values are also lower for CLIL texts, and therefore students are required to understand superordinate words, which have a broader meaning than the ones used in the ESL texts.

3.3.5 *Morphological information*

AzterTest calculates several metrics related to Morphological Information. In Table 8 we focus on the results for verbs.

The analysis of the morphosyntactic metrics shows that in English texts, the incidence of verbs is always significantly higher and that basic forms (present, infinitive) are predominant. As it might be expected, English texts tend to show a higher variability in the use of verb tense depending on the grade and the contents worked on. In the 1st grade, SS contents describe the natural and human-made environment and use present tense more frequently. Conversely, irregular past tense verbs are less frequent because they are introduced in the ESL curriculum. On the other hand, in the 2nd grade SS texts, which narrate historical events, we find out that the use of present tense and infinitive is reduced, past tense is significatively more frequent and that the mean values for irregular verbs in past

Table 8. Morphological information

Morphological information	6th grade of primary			1st grade of secondary			2nd grade of secondary		
	ESL texts	CLIL texts		ESL texts	CLIL texts		ESL texts	CLIL texts	
	M	M	*p*	M	M	*p*	M	M	*p*
Verb (i)	**147.35**	**107.02**	**0.024**	**151.88**	**117.69**	**0.048**	**154.56**	**115.41**	**0.009**
Verbs in present (i)	**85.44**	**44.87**	**0.030**	54.59	68.70	0.468	64.29	17.80	0.063
Verbs in future (i)	10.65	7.03	0.644	19.95	7.32	0.160	16.30	9.25	0.481
Verbs in past (i)	14.83	39.40	0.096	47.61	30.07	0.932	**30.30**	**75.64**	**0.036**
Irregular verbs in past (i)	8.01	11.63	0.878	31.89	6.74	0.349	17.90	23.20	0.481
Verbs in infinitive (i)	38.10	13.12	0.061	37.37	10.36	0.076	**45.83**	**12.08**	**0.009**
Verbs in gerund (i)	**1.05**	**10.33**	**0.008**	9.64	7.94	1	9.94	8.93	1

Note: M=Mean; *p*=*p* value; i=incidence per 1000 words

tense is higher, which might compromise understanding of these verbs that have previously been identified as less commonly used.

Tense and aspect repetition, which are quite prevalent in History books, are sometimes considered as an indicator of temporal cohesion. However, we can find examples with tense repetition that hide causal and temporal relationships in this example from the texts: *"Political decisions, such as wars and border changes force populations to move and create demographic voids and concentrations"*. In this case, they are all present tense but the student should not only understand the meaning but also use the semantic content of *"force"* to deduce the causal and temporal relationships, so tense repetition does not necessarily mean that the text is more understandable. Similarly, if the text does not include temporal expressions such as a period of time, adverbial phrases or adverbs of time, it might be confusing.

The following lines are an excerpt from the 2nd grade textbook: *"They emerged in the 9th century, following the division of the Carolingian Empire, and were located in the centre of the continent, where they were divided into feudal territories."*. In this example, the gerund is used to indicate that it happened after the division of the empire, but as we know, gerund might evoke events which are not only simultaneous, but also prior or subsequent with respect to a main verb. Consequently, it is essential to shed light in unclear temporal relationships between evoked events. This ambiguity cannot be found in the English texts, which express

much clearer temporal relationships as in *"Previously, each social class would present its petitions and requests to the king"*.

Another important aspect related to the tense and aspect repetition is that events in some SS and History passages are sometimes bound to a generic time. This time is descriptive or reportative, where the order of descriptions, as well as the internal event representations, is static or timeless and the most frequent tense is present tense with no salient aspectual cues.

3.3.6 *Syntactic complexity*

AzterTest calculates several metrics related to syntactic complexity. In Table 9 we present the results for verb phrase incidence, prepositional phrase incidence, mean of left embeddedness, incidence of relative subordinate clauses and incidence of negative forms.

Table 9. Syntactic complexity

	6th grade of primary			1st grade of secondary			2nd grade of secondary		
	ESL texts	CLIL texts		ESL texts	CLIL texts		ESL texts	CLIL texts	
Syntactic complexity	M	M	*p*	M	M	*p*	M	M	*p*
Verb phrase (i)	**147.59**	**107.02**	**0.024**	**151.88**	**117.69**	**0.048**	**154.56**	**115.41**	**0.009**
Prepositional phrase (i)*	**101.35**	**129.33**	**0.010**	**98.61**	**127.77**	**0.014**	**95.81**	**127.47**	**0.036**
Number of modifiers per NP (m)	**0.92**	**1.29**	**0.026**	**0.90**	**1.32**	**0.014**	**0.91**	**1.27**	**0.016**
Agentless passive voice verbs (i)	2.56	5.55	0.163	**1.05**	**12.34**	**0.005**	**0.31**	**16.25**	**0.004**
Left embeddedness (m)	2.18	2.72	0.078	2.53	3.22	0.160	2.71	3.30	0.209
Relative subordinate clauses (i)	7.38	9.49	0.338	**0.97**	**13.98**	**0.007**	7.37	13.92	0.194
Negative form (i)	6.26	2.28	0.387	**13.11**	**1.69**	**0.048**	**11.40**	**0.59**	**0.016**

Note:
* Measures obtained by Coh-Metrix; M=Mean; *p*=*p* value; i=incidence per 1000 words; m=mean value

Verb phrase incidence values demonstrate that History samples have less grammatical complexity in that aspect, since verb phrases such as *"should have come"* for example are less likely to be used in this type of text.

Another characteristic is that prepositional phrase incidence is higher in History books and they score higher in the number of modifiers per noun phrase, which is considered a measure of syntactic complexity (Biber et al, 2011). Generally, academic texts use, among others, modified noun phrases and they are much more common in informational written registers than in other registers. We can find the example *"Meanwhile, in mountains located in the temperate zone, there is deciduous woodland"* whereas in the English texts it is frequent to find adjectives as premodifiers, as in *"He had a nice home and a very good life"*. Scores also support that History texts samples share some characteristics of academic texts, such as a high frequency of nouns and more complex noun phrases (Biber et al., 1999).

There is a higher ratio of agentless passive forms in the CLIL texts, like *"In this time period a new cultural and scientific movement developed, it was called Humanism."*. Students will probably find them more difficult to interpret, because passive voice is more difficult to process than active voice (Just & Carpenter, 1987), with an extra load in the case of agentless sentences because they depend on readers' assumptions.

In general, the analysis of syntactic information suggests that the CLIL texts might be more complex than ESL texts in some concrete aspects. Even if the statistical analysis does not show significant differences, there are many examples in the CLIL texts of sentences with high left embeddedness. In grade 6, sentences in the CLIL texts have a greater mean value for left embeddedness ratio (number of words before the main verb) than the ESL texts, which is considered a measure of syntactic complexity because it implies a higher working memory load. We can find *"Lisa's favourite subject is ICT."* in the English texts and sentences with higher left embeddedness such as *"Some cities on the river valleys of Mesopotamia, Egypt, China and India grew thanks to trade in the Metal Age. In these cities, systems of social organization and communication emerged"* in the CLIL texts. Moreover, the use of relative subordinate clauses is also higher for this grade in SS. The incidence of the negative forms is, however, higher in the ESL texts.

3.3.7 *Referential cohesion*

The current version of AzterTest measures 8 forms of referential cohesion between sentences. In Table 10 we report on the argument overlap and content word overlap in both adjacent (local) and all sentences (global). These types of cohesive links have been shown to aid in text comprehension and reading speed (Kintsch & van Dijk, 1978).

Table 10. Referential cohesion

Referential cohesion	6th grade of primary			1st grade of secondary			2nd grade of secondary		
	ESL texts	CLIL texts		ESL texts	CLIL texts		ESL texts	CLIL texts	
	M	M	*p*	M	M	*p*	M	M	*p*
Local argument overlap (m)	**0.36**	**0.20**	**0.030**	0.37	0.39	0.932	0.29	0.31	1
Global argument overlap (m)	**0.19**	**0.06**	**0.015**	**0.25**	**0.13**	**0.022**	0.20	0.09	0.164
Local content word overlap (m)	**0.05**	**0.02**	**0.027**	0.04	0.04	0.929	0.02	0.03	0.698
Global content overlap (m)	**0.02**	**0.00**	**0.047**	0.02	0.01	0.076	0.02	0.01	0.178
Local argument overlap (m)	**0.36**	**0.20**	**0.030**	0.37	0.39	0.932	0.29	0.31	1

Note: M=Mean; *p*=*p* value; m=mean value

Regarding the results of referential cohesion (see Table 10), we find that values for global and local argument overlap and content word overlap are significantly higher in English texts of 6th grade, and not significantly higher in the rest of the grades. In the example *"Some languages don't conjugate verbs. Other languages have many different conjugations"*, from the English text, the word "language" is repeated, whereas in the CLIL example *"The countries involved in World War I considered the Germans responsible for starting it, so they forced Germany to pay for the disasters caused by the war. These costs were too great for (…)"* the word "costs" refers to "the disasters of the war".

3.3.8 *Connectives (logical cohesion)*

The current version of AzterTest measures the incidence of all connectives and logical, temporal, causal, conditional and adversative/contrastive connectives. We present these results in Table 11.

A previous study by Anderson et al. (1983) demonstrated that an event is processed faster when it is introduced in the text by an adverbial that indicates a time frame within the typical duration of that event.

Temporal connectives incidence is higher in SS texts than in the English texts of the first two grades, which means that ideas require cohesive links and that clauses provide clues about text organization. Therefore, CLIL texts of 6th and 1st grade require from the students an additional effort to establish the right temporal

Table 11. Connectives

	6th grade of primary			1st grade of secondary			2nd grade of secondary		
	ESL texts	CLIL texts		ESL texts	CLIL texts		ESL texts	CLIL texts	
Connectives	**M**	**M**	***p***	**M**	**M**	***p***	**M**	**M**	***p***
Logical connectives (i)	42.78	54.15	0.096	57.87	55.84	0.811	53.92	65.99	0.1
Temporal connectives(i)	6.54	13.17	0.093	14.15	17.38	0.672	16.43	16.26	0.727
Causal connectives (i)	3.04	10.19	0.073	**6.12**	**11.48**	**0.048**	4.99	14.07	0.093
Conditional connectives (i)	1.62	0.74	0.549	1.16	0.33	0.501	1.25	1.28	0.148
Adversative / contrastive connectives (i)	7.46	5.83	0.724	**9.34**	**3.45**	**0.027**	11.29	4.38	0.063
All connectives (i)	**61.47**	**84.11**	**0.023**	88.66	88.49	0.937	87.90	102.00	0.281

Note: M=Mean; *p*=*p* value; i=incidence per 1000 words

relationships between the ideas and understand all the connections in the texts. In second grade, temporal connectives are present at the same level for both types, so it can be said that the ESL syllabus seems to adapt progressively to the CLIL level and achieves an acceptable balance at this point.

Concerning causal connectives, they are more frequent in CLIL texts and their use is significantly higher in 1st grade, which is an evidence of the cause and effect relationships present in the History books that students need to be able to interpret.

Finally, the use of logical connectives is higher for CLIL texts in 6th and 2nd and lower for conditional, adversative and contrastive connectives (except for the conditional connectives in 2nd). If we consider all connectives, their use is significantly higher in CLIL texts of 6th grade and the measure has a higher mean value also in 2nd grade.

3.3.9 *Causal/Intentional cohesion*

Causal/Intentional cohesion is measured in Coh-Metrix by calculating two ratio indices that reflect the necessity of connectives: the ratio of causal particles to

causal verbs (SMCAUSr) and the ratio of intentional particles to intentional verbs (SMINTEr). Table 12 shows the values of these measures.

Table 12. Causal/Intentional cohesion

	6th grade of primary			1st grade of secondary			2nd grade of secondary		
Causal / Intentional cohesion	**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**		**ESL texts**	**CLIL texts**	
	M	**M**	***p***	**M**	**M**	***p***	**M**	**M**	***p***
Causal particles to causal verbs (r)*	0.19	0.39	0.103	0.18	0.35	0.075	0.19	0.42	0.194
Intentional particles to intentional verbs (r)*	**0.42**	**1.22**	**0.038**	0.50	0.97	0.160	0.64	1.08	0.138

Note:
* Measures obtained by Coh-Metrix; M=Mean; *p* = *p* value; r=ratio

SS texts, and significantly those of 6th grade, show a tendency to have a higher ratio of intentional particles (in order to, so that, by means of) to intentional verbs, or verbs that signal state changes, events, actions and processes as opposed to states. In general, these findings demonstrate that since intentional particles provide cohesion, the reader does not have to infer the relationships.

We also found at this level, even when there is no significant difference, that History texts show a tendency to a higher ratio of causal particles to causal verbs. This is one of the highest indicators of causal coherence, meaning that when intentional verbs are used (kill, impact) the texts include causal particles (because, consequently) to establish connections.

In sum, the main difficulty with CLIL texts is the vocabulary, since words are less common, they have a broader meaning and they are more sophisticated. This assertion has additionally been corroborated by the CEFR level score. Moreover, CLIL texts are less narrative and cohesive than the texts in TASA corpus and the ESL texts as seen in the PCs, and they have longer sentences and longer words than ESL texts. It has also been observed that tense and aspect repetition in CLIL texts may hide temporal and causal relationships. Finally, CLIL texts include more causal and temporal connectives and a higher ratio of intentional particles to intentional verbs and of causal particles to causal verbs.

4. Discussion

The results of this analysis, although limited by the size and diversity of samples, suggest that there are significant differences in the linguistic characteristics of the textbooks of SS/History and English also compared to texts in TASA corpus. Students must overcome language difficulties and unfamiliar discourse structures and grammatical forms, which will definitely impair the intended gains of CLIL. The complexity of textbooks can compromise the attempted increase in motivation and improvement in overall language competence because students will feel discouraged and uncertain of their abilities of understanding (Otwinowska & Foryś, 2017).

Even when compared to texts addressed at English native speakers we have found that narrativity and cohesion were lower in CLIL texts. The study leads us to advise against the use of these edited textbooks without adjustments, adaptations or taking appropriate measures to create effective CLIL lessons. However, the teachers should not assume the burden of making these texts accessible to students, even when they play an irreplaceable role by facilitating and stimulating meaning identification, checking relevant words, initiating a discussion to clarify and validate understanding and explaining and providing feedback on problematic language forms.

In order to enhance provision of textbooks that are more suitable for L2 learners, there needs to be improved communication between experienced CLIL teachers and publishers. In the interim, before such improved CLIL texts can be made available, we suggest that teachers be supported towards adopting an active approach to making the current texts more accessible. For example, through using simplification techniques such as replacing difficult words with simpler ones that reflect the content, remove unnecessary information, divide the text into sentences that include one idea and use simple structure sentences (subject + verb + high-frequency words) change passive tenses into active ones, change phrasal verbs to simpler ones, add logical connectors or replace idiomatic language. Lorenzo (2008) analyzes language adjustment and it has been taxonomized defining three different stances on adapting subject area texts for the bilingual classroom: simplification, elaboration and rediscursification. The conclusions of Lorenzo's research show that different adjustment techniques result in different sentence length and different cognitive demands and that lessons on input modification in foreign language teacher courses may also be a useful idea. In this regard, automatic text simplification is a research line in natural language processing (Saggion, 2017; Alva-Manchego et al., 2020) and its outcomes can help teachers in the task of adapting texts. Moreover, other text modifications

and adaptations can be automatized by means of the tools presented as LagunTest (Gonzalez-Dios et al, 2020) or tools like Texamen.[4]

CLIL texts also show a significantly lower lexical diversity than English texts but the words used in these texts are less frequent, thereby teachers should focus on explicit vocabulary instruction and use language carefully, taking time to introduce this specialized vocabulary. We suggest that the L1 should be appreciated as a potentially valuable tool in bilingual learning situations (Nikula & Moore, 2019) and the use of translanguaging techniques, as current trends on multilingual education advocate for a holistic approach integrating languages to reinforce each other (Cenoz, 2019).

Providing key vocabulary, code-switching and offering alternative ways of expressing ideas at key points of the lesson can also benefit students. Prepositional phrases, agentless passive voice, temporal and causal connectives are additional difficulties. Visual instructional materials and methods to identify causal and temporal relationships are also recommended based on the results. Students can achieve a better understanding using visuals, graphs, mindmaps for key concepts, timelines, model making, diagrams, sequences or using gestures or drama to learn relationships. In a similar way, using similes, analyzing similarities and differences, retelling the story or recognition and description tasks can promote comprehension and connections with new knowledge.

Regarding the development of the different skills, Lorenzo (2013) has identified that the CLIL teacher is also responsible for developing multilingual academic literacy and for keeping a balance between language reception and production and for promoting productive skills. Students' productions allow teachers to detect weaknesses, identify conceptual misunderstandings and they stimulate their cognitive processes. Oral work, which is often disregarded in CLIL lessons, should encourage expressive language use and diverse techniques such as pair-working, verbal interactions in groups, note-taking and oral sharing or guided cooperative questioning can be applied. The role of classroom interaction, feedback, questions, interactional scaffolding and a dialogic teaching approach are of great importance to create a learner-centred scenario (Llinares et al., 2012; Mercer et al., 2019). The long list of demands that the CLIL teachers must satisfy require specific formative programs to develop all the mentioned pedagogical and linguistic skills. As Coyle (2010, p.viii) states: "without appropriate teacher education programs the full potential of CLIL is unlikely to be realized and the approach unsustainable".

4. See http://www.texamen.com

5. Conclusion

The results of the study suggest that in this CLIL experience case, substantial linguistic differences exist between texts used for SS and History aimed at L2 learners and those of English as a subject. It is evident that the former requires a particular academic register as previous studies have described, but this research has allowed us to identify and quantify these statistically significant differences, which not only explains the difficulties perceived by the students, but it also provides support for the detection of potential learning obstacles and for the design of most favorable teaching strategies. Furthermore, the analysis points out that in certain aspects, CLIL texts might be even more difficult than those texts in a corpus of texts for native speakers, which clearly indicates that it is crucial to adapt them to the students' level. It may be stated that, in the case under study, CLIL texts are more linguistically complicated and that teaching them requires conscious effort and particular actions which also require specific teacher education programs.

Although additional research is needed with a greater sample, the present study demonstrates how computational tools can be of assistance to teachers and material developers to identify potential difficulties in the specific case of CLIL experiences. Further studies will allow us to investigate the results from a wider range of CLIL programs in different genres, subjects and languages.

References

Anderson, A., Garrod, S. C., & Sanford, A. J. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *Quarterly Journal of Experimental Psychology*, *35A*, 427–440. https://doi.org/10.1080/14640748308402480

Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 1–87. https://doi.org/10.1162/coli_a_00370

Beacco, J. C. & Byram, M. (2007). *From linguistic diversity to plurilingual education: Guide for the development of language education policies in Europe*. Council of Europe.

Bengoetxea, K., Gonzalez-Dios, I., & Aguirregoitia, A. (2020). AzterTest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural*, *64*, 1–8. https://doi.org/10.26342/2020-64-7

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar spoken and written English*. Longman.

Boroş, T., Dumitrescu, S.D., & Burtica, R. (2018). NLP-Cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 171–179). Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-2017

Cenoz, J. (2009). *Towards multilingual education: Basque educational research from an international perspective.* Multilingual Matters. https://doi.org/10.21832/9781847691941

Cenoz, J. (2019). Translanguaging pedagogies and English as a lingua franca. *Language Teaching*, *52*(1), 71–85. https://doi.org/10.1017/S0261444817000246

Coyle, D. (2010). Foreword. In D. Lasagabaster & Y. Ruiz de Zarobe (Eds.) *CLIL in Spain: Implementation, results and teacher training*, pp. vii–viii. Cambridge Scholars Publishing. https://doi.org/10.1093/elt/ccr056

Crossley, S.A., Allen, D.B., & McNamara, D.S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, *23*(1), 84–101. https://doi.org/10125/66657

Crossley, S.A., Greenfield, J., & McNamara, D.S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, *42*(3), 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, S.A., Louwerse, M.M., McCarthy, P.M., & McNamara, D.S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15–30. https://doi.org/10.1111/j.1540-4781.2007.00507.x

Cummins, J. (2008). Teaching for transfer: Challenging the two solitudes assumption in bilingual education. *Encyclopedia of language and education*, *5*, 65–75. https://doi.org/10.1111/j.1540-4781.2007.00507.x

Cummins, J. (2009). Multilingualism in the English-language classroom: Pedagogical considerations. *TESOL quarterly*, *43*(2), 317–321. https://doi.org/10.1002/j.1545-7249.2009.tb00171.x

Der, G., & Everitt, B.S. (2008). *A handbook of statistical analyses using SAS*. Chapman and Hall CRC.. https://doi.org/10.1201/9781584887850

Eusko Jaurlaritza (2010). Proceso de experimentación del marco de educación trilingüe. *Documento marco, 2010–2011*. Departamento de educación. http://www.hezkuntza.ejgv.euskadi.eus/r43-2459/es/contenidos/informacion/dig_publicaciones_innovacion/es_dig_publ/adjuntos/19_hizkuntzak_500/500013c_Pub_EJ_experimentacion_MET_c.pdf

Eusko Jaurlaritza (2018). Matrikula: Bilakaera irudiak. http://www.euskadi.eus/contenidos/informacion/graficosmatr1314/eu_def/adjuntos/Grafikoak17_18/EAE_eredua_OBLI.pdf

Franceschini, R. (2013). "History of multilingualism." In *the Encyclopedia of Applied Linguistics*, edited by C.A. Chapelle. 1–9. Blackwell Publishing.

Gierlinger, E.M. (2017). Teaching CLIL? *Journal of Immersion and Content-Based Language Education*, *5*(2), 187–213. https://doi.org/10.1075/jicb.5.2.02gie

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Graesser, A.C., McNamara, D.S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, *115*(2), 210–229. https://doi.org/10.1086/678293

Gonzalez-Dios, I., Bengoetxea, K., & Aguirregoitia, A. (2020). LagunTest: A NLP based application to enhance reading comprehension. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding Difficulties (READI)*, pp. 63–69. European Language Resources Association.

Heras, A., & Lasagabaster, D. (2015). The impact of CLIL on affective factors and vocabulary learning. *Language Teaching Research*, *19*(1), 70–88. https://doi.org/10.1177/1362168814541736

Huguet, À., Lasagabaster, D., & Vila, I. (2008). Bilingual education in Spain: Present realities and future challenges. *Encyclopedia of language and education*, 1672–1682. https://doi.org/10.1007/978-0-387-30424-3_127

Just, M.A., & Carpenter, P.A. (1987). *The psychology of reading and language comprehension*. Allyn & Bacon, Inc.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, W., & Van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Llinares, A., & Morton, T. (Eds.). (2017). *Applied linguistics perspectives on CLIL*. John Benjamins Publishing Company. https://doi.org/10.1075/lllt.47

Llinares, A., Morton, T., & Whittaker, R. (2012). *The roles of language in CLIL*. Cambridge University Press.

Llinares, A. & Whittaker, R.. (2010). Writing and speaking in the history class. In C. Dalton-Puffer & T. Nikula (Eds.) *Language use and language learning in CLIL classrooms*, (pp.125–124). John Benjamin Publishing. https://doi.org/10.1075/aals.7.14dal

Lorenzo, F. (2008). Instructional discourse in bilingual settings. An empirical study of linguistic adjustments in content and language integrated learning. *Language Learning Journal*, *36*(1), 21–33. https://doi.org/10.1080/09571730801988470

Lorenzo, F. (2013). Genre-based curricula: multilingual academic literacy in content and language integrated learning. *International Journal of Bilingual Education and Bilingualism*, *16*(3), 375–388. https://doi.org/10.1080/13670050.2013.777391

Lorenzo, F., Granados, A. & Ávila, I. (2019). The development of cognitive academic language proficiency in multilingual education: Evidence of a longitudinal study on the language of history. *Journal of English for Academic Purposes*, *41*, 100767. https://doi.org/10.1016/j.jeap.2019.06.010

López, L.E., Sichra, I. (2008). *Intercultural Bilingual Education Among Indigenous Peoples in Latin America*. In Hornberger, N.H. (eds). Encyclopedia of Language and Education. Springer. https://doi.org/10.1007/978-0-387-30424-3_132

Lyster, R., Ruiz de Zarobe, Y. (2018). Introduction: instructional practices and teacher development in CLIL and immersion school settings. *International Journal of Bilingual Education and Bilingualism*, *21*(3), 273–274. https://doi.org/10.1080/13670050.2017.1383353

McCarthy, P.M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P.M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD) [Doctoral dissertation]. The University of Memphis.

McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Mercer, N., Wegerif, R., & Major, L. (2019). *The Routledge international handbook of research on dialogic education*. Routledge. https://doi.org/10.4324/9780429441677

Meyer, O. (2010). Towards quality CLIL: Successful planning and teaching strategies. *PULSO. Revista de Educación*, *33*, 11–29.

Mikolov, T., Grave, É., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp 52–55. European Language Resources Association.

Nikula, T., & Moore, P. (2019). Exploring translanguaging in CLIL. *International Journal of Bilingual Education and Bilingualism*, *22*(2), 237–249. https://doi.org/10.1080/13670050.2016.1254151

O'Reilly, T., & McNamara, D.S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal*, *44*(1), 161–196. https://doi.org/10.3102/0002831206298171

Otwinowska, A., & Foryś, M. (2017). They learn the CLIL way, but do they like it? Affectivity and cognition in upper-primary CLIL classes. *International Journal of Bilingual Education and Bilingualism*, *20*(5), 457–480. https://doi.org/10.1080/13670050.2015.1051944

Qi, P., Dozat, T., Zhang, Y., & Manning, C. 2018. Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160–170. https://doi.org/10.18653/v1%2FK18-2016

Ruiz de Zarobe, Y. (2013). CLIL implementation: From policy-makers to individual initiatives. *International Journal of Bilingual Education and Bilingualism*, *16*, 231–243. https://doi.org/10.1080/13670050.2013.777383

Saggion, H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–137. https://doi.org/10.2200/S00700ED1V01Y201602HLT032

Schleppegrell, M.J. (2004). *The language of schooling: A functional linguistics perspective.* Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410610317

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Rand.

Speer, R., J. Chin, A. Lin, S. Jewett, & Nathan, L. (2018). *Luminosoinsight/wordfreq:v2.2.* https://doi.org/10.5281/zenodo.1443582

Whittaker, R., Llinares, A., & McCabe, A. (2011). Written discourse development in CLIL at secondary school. *Language Teaching Research*, *15*(3), 343–362. https://doi.org/10.1177/1362168811401154

Zeno, S., Ivens, S.H., Millard, R.T., & Duvvuri, R. (1995). *The educator's word frequency guide.* Touchstone Applied Science Associates.

Appendix A. Analysis of principal components

Table A1. Z cores of text easability PCs in SS genre

Text Easability PCs	TASA (native) Social Science 6th–8th	CLIL texts (Social Science)		
		6th grade (Prim.)	1st grade (Sec.)	2nd grade (Sec.)
	M	M	M	M
Narrativity	−0.50	−1.25	−1.15	−1.48
Referential cohesion	−0.26	−0.91	0.09	−0.77
Syntactic simplicity	0.40	0.55	0.71	0.59
Word concreteness	0.53	0.42	0.72	0.43
Causal cohesion	0.23	−0.31	−0.05	0.04

Note: PC = Principal Component; M = Mean; *p* = *p* value

Table A2. Z scores of the Coh-Metrix easability PCs in ESL and CLIL texts

Text Easability PCs	6th grade of primary			1st grade of secondary			2nd grade of secondary		
	ESL texts	*CLIL texts*		*ESL* texts	*CLIL* texts		*ESL texts*	*CLIL texts*	
	M	*M*	*p*	*M*	*M*	*p*	*M*	*M*	*p*
Narrativity	***−0.18***	***−1.25***	***0.0181***	***0.27***	***−1.15***	***0.0069***	***0.11***	***−1.48***	***0.0160***
Referential cohesion	***0.32***	***−0.91***	***0.0077***	*−0.06*	*0.09*	*0.8111*	*−0.59*	*−0.77*	*0.7110*
Syntactic simplicity	*0.74*	*0.55*	*0.4508*	*0.38*	*0.71*	*0.3706*	*0.69*	*0.59*	*0.9262*
Word concreteness	*0.82*	*0.42*	*0.3654*	*0.76*	*0.72*	*1*	*1.23*	*0.43*	*0.1947*
Causal cohesion	*−0.87*	*−0.31*	*0.1849*	*−0.31*	*−0.05*	*1*	*0.10*	*0.04*	*0.8636*

Note: PC = Principal Component; M = Mean; *p* = *p* value