Using global complexity measures to assess second language proficiency

Comparing CLIL and non-CLIL learners of English and Dutch in French-speaking Belgium

Amélie Bulon, Isa Hendrikx, Fanny Meunier & Kristel Van Goethem
Université catholique de Louvain (UCL)

Abstract. This study falls within the framework of an interdisciplinary project on Content and Language Integrated Learning (CLIL) in French-speaking Belgium. One of the project's aims is to compare the L2 language proficiency of CLIL and non-CLIL French-speaking learners of English and Dutch. In the present paper we focus on learners' global proficiency and use of different types of metrics to assess syntactic and lexical complexity in the learners' written productions. Using various computational tools, we extracted lexical and syntactic complexity scores for texts written by CLIL and non-CLIL pupils in their L2 (English or Dutch) and their L1 (French). These scores were then compared to investigate the potential influence of CLIL education on the pupils' language proficiency as CLIL programs provide more target language input than non-CLIL programs. We therefore hypothesized that CLIL pupils would display a more native-like competence in the target language, i.e. a more native-like level of syntactic and lexical complexity in their writing. As for the influence of CLIL programs on the L1, we did not expect any difference between the two groups. Our results show that our first hypothesis is only partly confirmed as the effect of CLIL on L2 complexity varies according to the language: while the Dutch texts written by CLIL pupils turn out to be more complex for nearly all measures, this was only the case for half of the measures in the English

texts. As initially expected for our second hypothesis, we found no influence of CLIL on the complexity of the pupils' L1.

1. Introduction

The present study has been carried out in the context of a concerted research action (ARC) Project entitled *Assessing Content and Language Integrated Learning (CLIL): Linguistic, Cognitive and Educational Perspectives*[i]. The project aims to investigate the influences of CLIL and other educational and motivational factors on the acquisition of a second language. Over 400 Belgian French-speaking fifth year primary and almost 500 fifth year secondary school pupils learning English or Dutch in immersion (CLIL) or traditional settings (non-CLIL) will be followed longitudinally for two consecutive school years (2015-2016 and 2016-2017).

The data types collected in the project will include various linguistic tasks, cognitive tests, socio-affective questionnaires and interviews in focus groups. The linguistic data from CLIL and non-CLIL learners will be analyzed and compared and the influence of cognitive, socio-emotional and pedagogical variables will be examined.

The aim of this paper is to examine the impact of CLIL education on the language proficiency of secondary school pupils through the use of a selection of complexity measures. The first section summarizes research on complexity measures, especially their relation to L2 proficiency and their validity as indicators of proficiency levels. The second section presents the research questions and hypotheses. The third section describes the data collection, the software and the selected complexity measures. The

fourth section is devoted to a pilot study related to a methodological issue regarding the influence of spelling/grammatical/punctuation mistakes on complexity scores. The final section presents the results and some concluding remarks.

2. Theoretical background

2.1. *Link between complexity and proficiency in L2*

SLA researchers have been investigating L2 proficiency for several decades, trying to identify what makes a L2 learner proficient and how L2 proficiency can best be measured (Housen, Kuiken & Vedder 2012).

L2 proficiency is defined by Thomas (1994, as cited in Bulté & Housen 2015:50) as "a person's overall competence and ability to perform in L2" which can be defined and analyzed according to three different components typically referred to as CAF, viz. Complexity, Accuracy and Fluency (Housen, Kuiken & Vedder 2012; Norris & Ortega 2009). In the CAF framework, L2 proficiency is mainly measured quantitatively, by means of a large number of frequencies, ratios and indices.

There is no overall agreement on a definition of complexity in L2 literature, hence the various meanings assigned to it across studies (Bulté & Housen 2015; Housen, Kuiken & Vedder 2012; Bulté & Housen 2012). Housen, Kuiken & Vedder (2012:2) report that complexity is "commonly characterized as the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2". As for Bulté & Housen, they further define complexity as

> *an absolute, objective and essentially quantitative property of language units, features*

> *and (sub) systems thereof in terms of (i) the number and the nature of discrete parts that the unit/feature/system consists of and (ii) the number and the nature of the interconnections between the parts (2015:50).*

In other words, a language feature or system is considered complex when it is made up of many components and when the relationships between these components are numerous and dense.

Bulté & Housen's taxonomy (2012) makes a distinction between relative and absolute complexity. The relative approach defines complexity by referring to "the mental ease or difficulty with which linguistic items are learned, processed or verbalized in the processes of language acquisition and use" (Hulstijn & De Graaff 1994, as cited in Bulté & Housen 2012:23). The absolute approach defines complexity in "objective, quantitative terms as the number of discrete components that a language system or feature consists of, and as the number of connections between the different components" (Bulté & Housen 2012:24). Absolute complexity can be further subdivided into discourse-interactional complexity, propositional complexity and linguistic complexity. In the present paper, the focus will be on absolute linguistic complexity in L2 writing.

Despite the lack of consensus between L2 researchers on the definition of complexity, they all seem to acknowledge the link between complexity and proficiency. Likewise, Bulté & Housen (2012) postulate that complexity is often investigated in SLA research as one possible descriptor of L2 performance and L2 proficiency in order to assess the effect of some other variable (learner variables such as age, but also different types of instruction and learning contexts, etc.). In fact, complexity is rarely investigated for its own sake, and is mainly used to "(a) gauge

proficiency, (b) to describe performance, and (c) to benchmark development" (Ortega 2012:4).

Recent research provides supporting evidence that complexity, accuracy and fluency measures positively correlate with L2 proficiency. In an investigation of the relationship between L2 proficiency and the CAF of L2 production, Kim, Nam & Lee (2016) showed that the greater the scores in the CAF of learners' L2 writing, especially in complexity, the greater the proficiency level. Ebrahimi (2015) reported large correlations between CAF measures and oral proficiency scores. More specifically, positive correlations between lexical diversity measures and overall proficiency were found in Daller, Van Hout & Treffers-Daller (2003), Treffers-Daller (2013) and Crossley, Salsbury & McNamara (2013). Vyatkina (2012) confirmed that length-based complexity measures correlated well with proficiency levels. Seo's study (2009, as cited in Kim, Nam & Lee 2016: 156) revealed significant differences in number of words, clauses and morphemes per sentence depending on the learners' proficiency levels.

2.2. *Use of CAF measures in the CLIL context*

Several studies have been carried out to evaluate the language proficiency of learners in immersive settings (very often in comparison with learners in non-immersive settings) using global measures of complexity, accuracy and/or fluency. Jexenflicker & Dalton-Puffer (2010), for example, identified highly significant differences between the writing of CLIL and non-CLIL for all measures of grammar and syntactic complexity, except for the number of subordinate clauses. Gené-Gil et al. (2015) reported significant differences in the development of written complexity, accuracy and fluency of CLIL learners over a 3-year period (only in accuracy for non-CLIL learners). Martínez (2015) investigated the writing of learners following bilingual and non-bilingual programs

and noticed that the bilingual group surpassed the non-bilingual group in all the fluency, accuracy and lexical complexity measures. Pérez-Vidal & Roquet (2015) identified larger gains in accuracy, syntactic and lexical complexity in the writing of learners who received extra CLIL hours. While most of these studies worked with small samples and focused only on English as a target language, the present study involves more than 400 pupils learning English or Dutch.

2.3. *Lexical and syntactic complexity measures as valid indicators of L2 proficiency*

A number of measures abound for each of the three CAF components and numerous studies have been devoted to the assessment of their validity and reliability as indicators of L2 proficiency (e.g. Wolfe-Quintero et al 1998, Ellis & Barkhuizen 2005 and Malvern et al 2004 as cited in Bulté & Housen 2015; Ortega 2003). Housen & Kuiken (2009, as cited in Bulté & Housen 2015:44) write that both in L1 and L2 research, complexity proved to be a "valid and basic descriptor of L2 performance, as an indicator of proficiency and as an index of language development and progress". Hence in the present study our focus is on complexity rather than accuracy and fluency.

The most popular measures of complexity are lexical and syntactic. Lexical measures relate to lexical competence and can be categorized into breadth of knowledge measures (e.g. word frequency and lexical diversity/variation), depth of knowledge measures (e.g. hypernymy, polysemy and word associations) and accessibility of lexical items (e.g. word concreteness, word imageability and word familiarity) (Crossley & McNamara 2009).
As for syntactic measures, they mostly

> *seek to quantify one of the following in one way or another: the length of production units (i.e. clauses, sentences, and T-units […]), the amount of embedding or subordination, the amount of coordination, the range of surface syntactic structures, and the degree of sophistication or particular syntactic structures (Ortega 2003, as cited in Lu 2010:1-2).*

The present study is a first exploratory analysis of a part of the corpora we collected (cf. footnote 1) in order to have an idea of the global proficiency level of CLIL and non-CLIL pupils, hence the focus on linguistic complexity only. Nevertheless, following studies will combine quantitative and qualitative approaches to capture the influence of CLIL on L2 proficiency in a more comprehensive manner.

3. Research questions and hypotheses

In this study we intend to examine the impact of CLIL education on the language proficiency of secondary school pupils through the use of a selection of complexity measures. Our first research question is the following:

1. Do CLIL learners have a higher L2 proficiency level than non-CLIL learners? In other words, will we encounter higher scores for the selected complexity measures in the texts written by the CLIL pupils?

Since CLIL programs provide greater exposure to the L2 and appear to be beneficial in terms of overall general competence (in particular in lexical and syntactic complexity), we expect higher

scores of complexity in the writing of CLIL learners, as has also been found in other studies[2] (see for instance the studies mentioned in section 2.3. or Navès & Victori 2010 and de Zarobe 2010).

Our second research question is the following:

2. Do CLIL programs also have an impact on the complexity of the learners' L1? If so, is it positive or negative?

Even though the influence of CLIL education on the L1 development appears to be a source of concern for parents (e.g. Pladevall-Ballester's 2015 study on CLIL in Catalonia), we expect CLIL and non-CLIL learners to perform equally well in French. Despite the fact that numerous parents believe that following between 50 up to 75% of the education in an L2 might be detrimental to the acquisition and mastery of their children's L1 (Van de Craen et al. 2013: 252), studies from the 1970's onwards have shown that CLIL programs do not seem to have a negative impact on the mother tongue (e.g. Braun & Vergallo 2010; Genesee 1989, as cited in Genesee 1991; Wesche 2002; Knell et al 2007; Marsh 2002). Some researchers even reported a positive influence of CLIL on the L1 (Vesterbacka 1991; Harley et al 1986, as cited in Wesche 2002).

4. Method

4.1. *Data collection*

The participants in the study are 438 French-speaking secondary school learners of Dutch or English from nine different schools in

Wallonia: Charleroi (n=48), Ciney (n=44), Tubize (n=27), Marche-en-Famenne (n=80), Wavre (n=54), Ottignies (n=87), La Louvière (n=34), Anvaing (n=20) and Tournai (n=44). These schools provide immersion programs in Dutch or/and English, along with traditional instruction: 229 pupils learn English (96) or Dutch (133) in immersion, while 209 learn English (97) or Dutch (112) in traditional settings. The participants' ages range from 15-18,9 and their mean age is 16,5. 207 (47%) of the learners are male and 231 (53%) female. All the pupils are in the fifth year of their secondary school education.

The learners performed two writing tasks between October 19th and November 9th 2015 in computer rooms in Louvain-la-Neuve. The tasks consisted in writing an e-mail of at least 15 lines on two possible topics (either their last holidays or a party they attended). They wrote the first e-mail in the foreign language (Dutch or English) in the morning and the second one in their mother tongue (French) in the afternoon. The task was timed (25 minutes to write one e-mail) and we made sure the pupils had no access to dictionaries or other reference tools. A few texts were lost due to technical problems, but as Table 1 shows, we were able to collect a total of 843 productions:

French (L1)	Dutch CLIL	Dutch non-CLIL	English CLIL	English non-CLIL
431	132	100	90	90

Table 1: Number of texts collected per condition (language – type of education)

As text type has been reported to have a great influence on lexical diversity (see for instance Yu 2009), we controlled for this by making all the pupils write on the two same topics (if a pupil had written about a party in the morning, s/he was given the other topic – holidays – in the afternoon, and vice versa).

4.2. *Computer software*

For the English texts, we used Coh-Metrix, a computational tool "which analyzes texts on over 200 measures of cohesion, language and readability" (Graesser, McNamara, Louwerse & Cai 2004:93). For the Dutch texts, we used a similar tool, T-Scan, which stands for *Software voor Complexiteits-Analyse van het Nederlands* (Pander Maat et al. 2014).

Both Coh-Metrix and T-Scan are based on various tools and resources such as part-of-speech taggers, syntactic parsers, frequency lists and other components developed in computational linguistics.

For the French texts, as software tools are scarce, we used Wordsmith Tools (Scott 2016) for basic word and sentence counts and then used more complex measures included in a software tool for French developed by François (2011).

As the computational tools offered a large number of complexity measures, some being similar for the three languages and some others specific to one or two language(s) only, a selection of shared indices had to be made (see following section).

4.3. *Complexity measures selected*

Among the measures offered by each tool, we selected both lexical and syntactic measures that were similar for English, Dutch and French. Because we had to rely on three different software programs to compute the complexity scores (one for each language under study), our choice was limited to the most relevant measures that could be computed by each program and compared for the three languages involved.

The total number of sentences and words per text are general measures of complexity. The more words and sentences in a text, the more complex and the higher the perceived quality of the text (Reid 1990 as cited in Crossley, Kyle, Allen, Guo & McNamara 2014; Ferris 1994; Frase et al. 1999).

Sentence length (in number of words) was chosen as a syntactic complexity measure. It is often used in L2 and L1 research to measure linguistic proficiency in general (Iwashita, Brown, McNamara & O'Hagan 2008 and Tavakoli & Foster 2008 for L2 research; Brown 1973 as cited in Bulté & Housen 2012 and Hunt 1965 for L1 research) and has been reported to be a good indicator of L2 text quality by various researchers (Reid 1990 as cited in Crossley, Kyle, Allen, Guo & McNamara 2014; Frase et al., 1999; Grant & Ginther, 2000).

Regarding the lexical complexity measures, we chose word length (in letters and syllables) since it can give information about learners' proficiency level (Grant & Ginther 2000). We also used the lexical diversity measures TTR (Type-Token Ratio) and MTLD (Measure of Textual Lexical Diversity), which also appear to be good predictors of overall L2 proficiency (Daller, Van Hout & Treffers-Daller 2003 as cited in Treffers-Daller, Parslow & Williams 2016; Treffers-Daller 2013; Crossley, Salsbury & McNamara 2013; Crossley, Salsbury & McNamara 2014; Crossley, Salsbury & McNamara 2011; Yu 2009). For English and Dutch, TTRs were computed for content words and for all words. For French, TTRs are computed for inflected forms and lemmas.

We are aware of the fact that Type-Token Ratio is sensitive to text length and tends to decrease as the length of the text increases. More recent indices have been developed and Koizumi (2012) found that the *Measure of Textual Lexical Diversity*[3] **(**MTLD) was

least affected by text length compared to TTR and other recent indices (Guiraud index and D), when used with texts of at least 100 tokens. That is why we chose to include it along with TTR for English and Dutch (we were unable to compute this score for the French texts). In addition, as our texts are similar in terms of text lengths, the length effect – if any – could only be minimal.

5. Pilot study: do spelling/grammar/punctuation mistakes in texts affect complexity measures?

It has been shown that spelling/grammatical/punctuation mistakes in the writing of the learners might modify some measures significantly (e.g. increase in TTR as misspelt words are considered as 'new' words) (Treffers-Daller, Parslow & Williams 2016; Yu 2009). In order to examine the effect of mistakes on the computation of the complexity scores, we randomly selected 20 English texts, 20 Dutch texts and 20 French texts and corrected them. Any misspelt word, typing error or grammatical mistake such as basic subject-verb agreement were corrected. Punctuation errors (e.g. missing full stops, no space following commas, etc.) were also corrected. We then compared the scores calculated for the corrected texts with those for the original texts using Spearman correlation coefficient.

		DUTCH		**ENGLISH**		**FRENCH**	
		Orig.	**Cor.**	**Orig.**	**Cor.**	**Orig.**	**Cor.**
Sentences per text	**Mean**	24,5	24,3	20,0	20,7	23,4	22,3
	Median	25	25	20,5	21,0	23,0	20,0
	Variance	60,579	55,145	32,05	33,08	87,50	80,43
		r_s =,99, p <.001		r_s = 0,96, p <.001		r_s = 0,92 p <.001	
Words per text	**Mean**	230,6	225,8	276,4	278,2	317,5	317,2
	Median	258,5	247	267,0	273,5	279,0	281,5
	Variance	4139,0	3892,0	5019,50	4945,01	7432,05	7460,87

		r_s = 0,91, p <.001		r_s = 1,00, p <.001		r_s = 1,00, p <.001	
Words per sentence	**Mean**	10,2	10,0	14,8	14,3	25,6	19,4
	Median	8,9	9,2	14,1	13,6	19,1	18,0
	Variance	17,9	12,2	27,95	24,67	713,6	28,2
		r_s = 0,99, p <.001		r_s = 0,99, p <.001		r_s = 0,80, p <.001	
Syllables/morphs per word	**Mean**	1,2	1,2	1,30	1,29	-	-
	Median	1,2	1,3	1,29	1,29	-	-
	Variance	0,004	0,004	0,002	0,002	-	-
		r_s = 0,87, p <.001		r_s = 0,99,p <.001		-	
Letters per word	**Mean**	4,3	4,3	3,87	3,86	4,1	4,1
	Median	4,3	4,3	3,9	3,89	4,1	4,1
	Variance	0,037	0,032	0,026	0,022	0,04	0,03
		r_s = 0,97, p <.001		r_s = 0,95, p <.00		r_s = 0,95, p <.001	
TTR content words	**Mean**	0,75	0,75	0,73	0,73	0,76	0,76
	Median	0,73	0,73	0,73	0,73	0,76	0,77
	Variance	0,006	0,008	0,003	0,005	0,001	0,001
		r_s = 0,95, p <.001		r_s = 0,96,p <.001		r_s = 0,93, p <.001	
TTR all words	**Mean**	0,52	0,52	0,51	0,51	0,67	0,66
	Median	0,51	0,50	0,51	0,50	0,67	0,67
	Variance	0,005	0,005	0,002	0,002	0,001	0,001
		r_s = 0,84, p <.001		r_s = 0,99, p <.001		r_s = 0,9, p <.001	
MTLD all words	**Mean**	61,0	59,7	68,1	68,3	-	-
	Median	62,3	60,6	64,4	63,3	-	-
	Variance	196,1	174,4	317,04	372,65	-	-
		r_s = 0,91, p <.001		r_s = 0,98,p <.001			

Table 2: Comparison between the original and corrected English, Dutch and French texts (green = significant correlation)

As can be seen from Table 2, all measures computed on the original and the corrected measures correlate significantly. Correcting the texts did not significantly alter the complexity measures in any of the languages.

6. Results

Since the pilot study showed that spelling, grammatical and punctuation mistakes did not result in many significant differences and as the correction of more than 800 texts would have been

highly time-consuming, we went ahead with the analysis of the original texts. This said, all 843 texts were nonetheless corrected for punctuation mistakes (missing spaces after full stops and commas were added) as this could be done semi-automatically. In this manner, we expect to increase accuracy as the pilot study showed significant differences in the number of words and the number of sentences in the English texts after correction. For the Dutch and the French texts, we expect the accuracy to improve slightly yet not significantly, as can be assumed from the pilot study.

6.1. *Comparison CLIL/non-CLIL*

Table 3 presents the median scores per complexity measure for the texts written by the CLIL and the non-CLIL pupils (Table 5 containing the statistics of comparison between the CLIL and the non-CLIL scores is included in the appendices).

		DUTCH			ENGLISH			FRENCH		
		non-CLIL		**CLIL**	**non-CLIL**		**CLIL**	**non-CLIL**		**CLIL**
Sentences per text	**Median**	21,0	<	25,5	22,5	=	22,0	18,0	=	19,0
	variance	50,72		41,86	64,23		54,32	59,03		53,65
Words per text	**Median**	194,5	<	274,0	267,5	<	328,0	288,0	=	296
	variance	3571,47		3339,65	6930,62		4071,49	6836,66		5248,65
Words per sentence	**Median**	9,0	<	10,7	11,0	<	14,5	14,9	=	14,8
	variance	6,14		9,77	29,71		35,20	51,66		20,64
Syllables/morphs per word	**Median**	1,23	<	1,27	1,30	>	1,28	-	-	-
	variance	0,004		0,002	0,002		0,002	-		-

Letters per word	**Median**	4,27	<	4,35	3,86	=	3,84	4,03	=	4,06
	variance	0,04		0,03	0,29		0,02	0,04		0,04
TTR content words	**Median**	0,77	=	0,77	0,70	=	0,71	0,68	=	0,68
	variance	0,006		0,004	0,12		0,004	0,001		0,002
TTR all words	**Median**	0,54	>	0,51	0,49	=	0,49	0,76	=	0,77
	variance	0,005		0,003	0,01		0,003	0,001		0,001
MTLD	**Median**	58,5	<	73,0	60,2	<	70,4	-	-	-
	variance	323,82		258,74	205,67		269,42	-		-

Table 3: Median scores per measure for each language group (< CLIL significantly higher score, > CLIL significantly lower score)

A closer look at Table 3 shows us that most significant differences can be found in the texts written in Dutch; only half of the scores diverge significantly in English and no significant differences are observed in the pupils' native language, French.

Using the Mann Whitney test[4] to calculate the differences between the scores, we found that the Dutch texts written by the CLIL pupils were more complex than those produced by the non-CLIL pupils for all measures but TTR content words – which remained insignificant – and TTR all words, which was significantly lower in the texts written by CLIL pupils.

Examining the English texts, those written by the CLIL pupils were significantly more complex in terms of words per text, sentences per texts and MTLD. Regarding the other four measures, no significant differences were encountered. In contrast, the non-CLIL learners used significantly longer words (with regard to the number of syllables per word) compared to the CLIL learners.

Investigating the French texts, we found no significant differences between the pupils in CLIL and in non-CLIL. This finding seems to indicate that CLIL education does not have any (positive or negative) influence on the complexity of the pupils' mother tongue.

7. Conclusions

The present study aimed at assessing the L2 proficiency level of secondary school pupils learning Dutch or English in CLIL and non-CLIL settings using a set of complexity measures. We hypothesized a higher level of L2 complexity in the texts written by the CLIL pupils. Regarding the L1 texts, we did not expect to find any difference between the pupils in CLIL and in non-CLIL.

On the basis of our pilot study, we can argue that, methodologically speaking, spelling/grammatical/punctuation mistakes do not significantly alter the selected complexity scores computed by T-scan, Coh-Metrix and François' (2011) software.

As for the comparison between CLIL and non-CLIL pupils, the results vary according to the language. For Dutch, we found that the texts written by CLIL pupils were significantly more complex than those written by non-CLIL pupils, except for TTR content words and TTR all words. For English, we observed that the texts written by CLIL pupils were more complex regarding three of the complexity measures under study and that the non-CLIL pupils used significantly more syllables per word. We found no influence of CLIL on the complexity of French, which is in line with our hypothesis that CLIL does not have a negative impact on the L1.

Our results show that CLIL appears to have more impact in fostering L2 complexity in Dutch than in English. One possible

explanation for this may be that non-CLIL learners of English are more proficient in the L2 than non-CLIL learners of Dutch are. The greater availability of English outside school and/or its greater attractiveness as a L2 may also play a role (De Le Vingne 2014; Berns et al. 2007), so that the impact of CLIL in English is less evident compared to CLIL in Dutch. Another potential explanation is that non-CLIL learners of Dutch would have more difficulty acquiring Dutch than non-CLIL learners of English acquiring English, one potential reason being that French shares more (syntactic, morphological and lexical) properties with English than with Dutch (Pierce 2012; Fox 2002; Pasquarella et al. 2011; Laufer & Paribakht 1998; Forlot & Beaucamp 2008).

Further analyses will soon be carried out to investigate the potential influence of socio-linguistic variables such as gender, age and number of years in CLIL education, and cognitive variables (e.g. IQ). In addition, we will carry out a thorough analysis of specific linguistic constructions (phraseological language and intensifying constructions) in comparative corpora of L1 and L2 speakers of English, Dutch and French (of which the learner sub-corpora used in the present paper, form a part). A following step is to integrate these different quantitative and qualitative measures of L2 proficiency to gain a more comprehensive understanding of the characteristics of the different interlanguages. Since our project is longitudinal, we will replicate this study in spring 2017 in order to track any possible evolution.

REFERENCES

Berns, M., Claes, M. T., de Bot, K., Evers, R., Hasebrink, U., Huibregtse, I., Truchot, C. & van der Wijst, P. (2007) English in Europe. In M. Berns, K. de Bot & U. Hasebrink, éds, *In the presence of English: Media and European Youth*, Springer, New-York, 15-42.

Braun, A. & Vergallo, E. (2010) Influences de l'immersion linguistique sur la maîtrise du français. *Education et Formation* 292, 155-165.

Bulté, B. & Housen, A. (2012) Defining and Operationalizing L2 Complexity. In A. Housen, F. Kuiken & I. Vedder, éds, *Dimensions of L2 Performance and Proficiency – Investigating Complexity, Accuracy and Fluency in SLA*. John Benjamins, Amsterdam, 21-46.

Bulté, B. & Housen, A. (2015) Evaluating short-term changes in L2 complexity development. *Círculo de lingüística aplicada a la comunicación* 63, 42-76.

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L. & McNamara D. S. (2014) Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation, *The Journal of Writing Assessment* 7 (1).

Crossley, S. A. & McNamara, D. S. (2009) Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing* 18 (2), 119-135.

Crossley, S. A., Salsbury, T. & McNamara, D. S. (2011) Predicting the proficiency level of language learners using lexical indices. *Language Testing* 29 (2), 243-263.

Crossley, S. A., Salsbury, T. & McNamara, D. S. (2013) Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M.Daller, éds, *Human ratings and automated measures*. John Benjamins, Amsterdam, 47-105.

Daller, H., Van Hout, R. & Treffers-Daller, J. (2003) Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24 (2), 197-222.

De le Vingne, C. (2014) Immersie-Onderwijs met het Nederlands als Doeltaal. Onderzoek naar de attitudes en motivatie van de ouders van immersie- en niet-immersie-leerlingen uit Luikse secundaire scholen. Université de Liège.

de Zarobe, Y. R. (2010) Written production and CLIL: An empirical study. In C. Dalton-Puffer, T. Nikula & U. Smit, éds, *Language use and language learning in CLIL classrooms*, John Benjamins, Amsterdam, 191-212.

Ebrahimi, E. (2015) The effect of dynamic assessment on complexity, accuracy and fluency in EFL learners' oral production. *International Journal of Research Studies in Language Learning*, 4 (3), 107-123.

Ferris, D. R. (1994) Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly* 28 (2), 414-420.

Forlot, G. & Beaucamp, J. (2008) Heurs et malheurs de la proximité linguistique dans l'enseignement de l'anglais au primaire. *Ela, Etudes de Linguistique Appliquée* 1, 77-92.

Fox, H. J. (2002) Phrasal cohesion and statistical machine translation. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 304-311.

François, T. (2011) Les apports du traitement automatique du langage à la lisibilité du français langue étrangère, PhD thesis, Université catholique de Louvain.

Frase, L. T., Faletti, J., Ginther, A. & Grant, L. (1998) Computer analysis of the TOEFL Test of Written English. *ETS Research Report Series* 1998 (2), 1-24.

Gené-Gil, M., Juan Garau, M. & Salazar-Noguera, J. (2015) Writing development under CLIL provision. In M. Juan-Garau & M. Salazar-Noguera, éds, *Content-based Language Learning in Multilingual Educational Environments*. Springer International Publishing, Cham, Switzerland, 139-161.

Genesee, F. (1991) Second language learning in school settings: Lessons from immersion. In A. G. Reynolds, éd, *Bilingualism, multiculturalism, and second language learning: The McGill conference in honour of Wallace E.* Lambert. Lawrence Erlbaum Associates, New-York, 183-201.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments & computers* 36 (2), 193-202.

Grant, L. & Ginther, A. (2000) Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing* 9 (2), 123-145.

Housen, A., Kuiken, F. & Vedder, I. (2012) Complexity, Accuracy and Fluency: Definitions, Measurement and Research. In A. Housen, F. Kuiken & I. Vedder, éds, *Dimensions of L2 Performance and Proficiency –*

Investigating Complexity, Accuracy and Fluency in SLA, John Benjamins, Amsterdam, 1-20.

Hunt, K. W. (1965) Grammatical Structures Written at Three Grade Levels. NCTE Research Report No.3

Iwashita, N., Brown, A., McNamara, T., O'Hagan, S. (2008) Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29, 24-49.

Jexenflicker, S., & Dalton-Puffer, C. (2010). The CLIL differential: Comparing the writing of CLIL and non-CLIL students in higher colleges of technology. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), Language use and language learning in CLIL classrooms. John Benjamins, Amsterdam, 169-190.

Kim, Y., Nam, J. & Lee, S-Y (2016) Correlation of proficiency with complexity, accuracy and fluency in spoken and written production: evidence from L2 Korean. *Journal of the National Council of Less Commonly Taught Languages*, 19, 147-181.

Knell, E., Haiyan, Q., Miao, P., Yanping, C., Siegel, L.S., Lin, Z. & Wei, Z. (2007) Early English Immersion and Literacy in X'ian, China. *The Modern Language Journal* 91 (3), 395-417.

Koizumi, R. (2012) Relationships between text length and lexical diversity measures: can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction* 1 (1), 60-69.

Laufer, B. & Paribakht, T. S. (1998) The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning* 48 (3), 365-391.

Lu, X. (2010) Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15 (4), 474-469.

Marsh, D. (2002) CLIL/EMILE – The European Dimension: Actions, Trends and Foresight Potential. University of Jyväskylä, Finland. https://jyx.jyu.fi/dspace/handle/123456789/47616, accessed July 21, 2016.

Martínez, A. C. L. (2015). Analysis of the Written Competence of Secondary Education Students in Bilingual and Non-Bilingual Programmes. In Conference proceedings. ICT for language learning. libreriauniversitaria.it edizioni, 499-503.

McCarthy, P. M. & Jarvis, S. (2010) MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42 (2), 381-392.

Navés, T. & Victori, M. (2010) CLIL in Catalonia: An Overview of Research Studies. In D. Lasagabaster & Y. R. de Zarobe, éds, *CLIL in Spain: Implementation, Results and Teacher Training*, Cambridge Scholars, Newcastle, 30-54.

Norris, J. M. & Ortega, L. (2009) Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics* 30 (4), 555-578.

Ortega, L. (2003) Syntactic Complexity Measures and their Relationship to L2 proficiency: A Research Synthesis of College-level L2 writing. *Applied Linguistics* 24 (4), 492-518.

Ortega, L. (2012) Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi, éds, *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Walter de Gruyter, Berlin, 127-155.

Pander Maat, H. Kraf, R., Dekker, N., Sloot, K. van der, Bosch, A. van den, Gompel, M. van & Klein, S. (2014) *Handleiding T-Scan*. Available at http://webservices-lst.science.ru.nl/

Pasquarella, A., Chen, X., Lam, K., Luo, Y. C. & Ramirez, G. (2011) Cross-language transfer of morphological awareness in Chinese-English bilinguals. *Journal of Research in Reading* 34 (1), 23-42.

Pérez-Vidal, C. & Roquet, H. (2015) CLIL in Context: Profiling Language Abilities. In M. Juan-Garau & M. Salazar-Noguera, éds, *Content-based Language Learning in Multilingual Educational Environments*. Springer International Publishing, Cham, Switzerland, 237-255.

Pierce, A. E. (2012) *Language acquisition and syntactic theory: A comparative analysis of French and English child grammars*. Springer Science & Business Media.

Pladevall-Ballester, E. (2015) Exploring Primary school CLIL perceptions in catalonia: students', teachers' and parents' opinions and expectations. *International journal of bilingual education and bilinguisim* 18 (1), 45-59.

Scott, M. (2016) WordSmith Tools version 6, Stroud: Lexical Analysis Software.

Tavakoli, P. & Foster, P. (2008) Task design and second language performance: The effect of narrative type on learner output. *Language Learning* 58, 439-473.

Treffers-Daller, J. (2013) Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller, éds, *Vocabulary Knowledge: Human Ratings and Automated Measures*. John Benjamins, 79-105.

Treffers-Daller, J., Parslow, P. & Williams, S. (2016) Back to basics: how measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics* amw009.

Van de Craen, P., Surmont, J., Ceuleers, E. and Allain, L. (2013) How policies influence multilingual education and the impact of multilingual education on practices. In Berthoud, A.C., Grin, F. and Lüdi, G. (eds). *Exploring the Dynamics of Multilingualism. The DYLAN project.* Amsterdam: John Benjamins, 343 – 372.

Vesterbacka, S. (1991) Ritualised routines and L2 acquisition: Acquisition strategies in an immersion program. *Journal of Multilingual and Multicultural Development* 12 (1-2), 35-43.

Vyatkina, N. (2012) The development of second language writing complexity in groups and individuals: a longitudinal learner corpus study. *The Modern Language Journal*, 96 (4), 576-598.

Wesche M. B. (2002) Early French immersion: How has the original Canadian model stood the test of time? In P. Thorsten, A. Rohde, H. Wode, & P. Burmeister, éds, *An Integrated View of Language Development: Papers in Honor of Henning Wode.* Wissenschaftlicher Verlag Trier, Trier, 357-378.

Yu, G. (2009) Lexical diversity in writing and speaking task performances. *Applied Linguistics* 31 (2), 236-259.

APPENDICES

	DUTCH CLIL / non-CLIL	**ENGLISH CLIL / non-CLIL**	**FRENCH CLIL / non-CLIL**
Sentences per text	*U* = 4234,0 *z* = -4,68, *p* <.05* *r* = -0,307	*U* = 3937,0 *z* = -0,32, *p* >.05 *r* = -0,024	*U* = 21647,5 *z* = -1,10 *p* >.05 *r* = -0,053
Words per text	*U* =1909,0 *z* = -9,27, *p* <.05* *r* = -0,601	*U* = 2254,5 *z* = -5,14, *p* <.05* *r* = -0,383	*U* = 20838,5 *z* = -1,73, *p* >.05 *r* = -0,083
Words per sentence	*U* = 4257,5 *z* = -4,63, *p* <.05* *r* = -0,301	*U* = 2539,5 *z* = -4,32, *p* <.05* *r* = -0,322	*U* = 22350,5 *z* = -0,06, *p* >.05 *r* = -0,003
Syllables/morphs per word	*U* = 3889,0 *z* = -5,36, *p* <.05* *r* = -0,352	*U* = 3250,5 *z* = -2,28, *p* <.05** *r* = -0,170	
Letters per word	*U* = 5034,0 *z* = -3,09, *p* <.05* *r* = -0,203	*U* = 3896,0 *z* = -0,44, *p* >.05 *r* = -0,033	*U* = 20641,0 *z* =-1,88, *p* *p* >.05 *r* = -0,091
TTR content words	*U* = 6124,5 *z* = -0,94, *p* >.05 *r* = -0,062	*U* = 3655,5 *z* = -1,129, *p* >.05 *r* = -0,084	*U* = 23031,5 *z*=-0,03, *p* >.05 *r* = -0,001
TTR all words	*U* = 5188,5 *z* = -2,79, *p* <.05** *r* = -0,183	*U* = 3762,0 *z* = -0,82, *p* >.05 *r* = -0,061	*U* = 22869,5 *z* = -0,15, *p* >.05 *r* = -0,007
MTLD all words	*U* = 3587,0 *z* = -5,95, *p* <.05* *r* = -0,391	*U* = 2380,0 *z* = -4,78, *p* <.05* *r* = -0,356	

Table 4: Comparison between the complexity scores computed for the texts written by the CLIL and the non-CLIL pupils, Mann-Whitney (*CLIL significantly higher score, ** non-CLIL significantly higher score)

1 For more information about the project : https://uclouvain.be/fr/instituts-recherche/ilc/assessing-content-and-language-integrated-learning-clil.html

[2] It is important to note that these studies do not control for IQ or any other similar variable – we intend to include a cognitive variable in further analyses

3 MTLD is calculated as the "mean length of sequential word strings in a text that maintain a given TTR value" (McCarthy & Jarvis 2010:385)

[4] Most of the scores for the English texts, half of the scores for the French texts and some of the scores for the Dutch texts are not normally distributed (Kolmogorov-Smirnov and Shapiro-Wilk tests of normality are significant), hence the choice for the non-parametric Mann-Whitney test.