Matching the data: developing a multilingual corpus of language development

Manon Buysse[1], Bastien De Clercq[2]

[1] Vrije Universiteit Brussel – manon.buysse@vub.ac.be
[2] Vrije Universiteit Brussel – bastien.de.clercq@vub.ac.be

Abstract

In recent years, there has been a growing awareness in the field of Second Language Acquisition (SLA) that research into some of the outstanding questions on language development is contingent upon the use of appropriate corpora (Housen & Kuiken, 2009 ; Larsen-Freeman, 2009 ; Verspoor et al., 2011 ; Verspoor et al., 2012 ; Vyatkina 2012). Previous studies in SLA have too often relied on cross-sectional corpora and are frequently biased towards English (Myles, 2005). Whereas cross-sectional corpora rely on group means to approximate developmental trajectories, longitudinal corpora allow for a more direct examination of linguistic development. However, the challenges posed by the constitution of longitudinal corpora from a practical perspective are numerous: data collection is often limited in terms of time span, group size is difficult to maintain because of participant dropout, and it may be difficult to obtain equivalent data without repetition effects. In multilingual corpora, these challenges are not only repeated for each additional language, but the corpora should also be as equivalent as possible. For corpora intended to reflect linguistic development, this equivalence should ideally be guaranteed for, amongst others, the participant background, the completed task and the proficiency levels or developmental stages. Additionally, research departments have often accumulated a considerable body of data which, though not initially conceived as multilingual corpora, may share enough characteristics to constitute a multilingual corpus. Reasons for integrating data into a larger corpus may arise from practical and ecological grounds, such as the amount of time and resources originally invested in the data gathering, but also from the untapped potential of these data. On the basis of a case study, we will discuss some of the challenges encountered when constituting a multilingual corpus for the purposes of SLA research, and possible solutions to these issues. More specifically, we will address the notion of corpus equivalence and to what extent it is both feasible and necessary. Additionally, we will consider a number of methods to triangulate linguistic proficiency. We will equally discuss the implications for research in terms of what can and what cannot be expected from such a multilingual corpus, in light of our own data: the need to assemble a multilingual corpus arose from two related PhD projects aiming to compare linguistic development in oral production for L2 English and French across a time frame spanning from early to fairly advanced proficiency.

Keywords: second language acquisition, second language proficiency, multilingual L2 corpora, comparability

1. Introduction

In the field of Second Language Acquisition (SLA), researchers typically study the development of (a part of) the linguistic system in a second language (L2) over a period of time. Over the past decades, one of the main empirical source for this type of research has been the use of learner corpora, i.e. either spoken or written data produced by learners of the particular second or foreign language under investigation. Whereas the main purpose in SLA research is to reveal "the mental representations and developmental processes which shape and constrain second language (L2) productions" (Myles, 2005:374), these processes are not easily observable, and the best way of attempting to uncover them is the study of their concrete realizations, i.e. what learners of a given language can do in their L2 productions.

To this end, researchers have often relied on cross-sectional corpora that are supposed to reflect different stages in the language learning process, even though these do not strictly speaking represent learner development. However, in recent years, there has been a call for longitudinal corpora, which more accurately represent the development of a L2 through dense data collection intervals. Recently, some researchers have equally prioritized the study of individual learner development, rather than the development of group tendencies (Housen & Kuiken, 2009 ; Larsen-Freeman, 2009 ; Verspoor et al., 2011 ; Verspoor et al., 2012 ; Vyatkina, 2012).

The challenges posed by the constitution of such a corpus are numerous. Firstly, longitudinal corpora are notoriously difficult to maintain, due to participant dropout. As a consequence, these corpora are typically limited over time. Moreover, longitudinal studies typically trace the linguistic development of only a limited number of participants. Additional problems pertain to task equivalence: in order to avoid a task-repetition effect, researchers need to find roughly equivalent tasks, which elicit the same type of language from the participant. These issues can be overcome in a cross-sectional corpus, since participants only complete the task once and do not have to be available for later data collection points.

The explicit cross-linguistic comparison of learner development in multilingual corpora is, however, still a relatively new area in SLA. In addition to the representative portrayal of learner development, multilingual corpora should also offer a degree of cross-linguistic comparability. The notion of comparability has more frequently been discussed in contrastive and translational corpus linguistics (Teubert, 1996 ; Laviosa, 1997 ; McEnery & Xiao, 2007). While some researchers conclude that full comparability is only possible for translation corpora (Aarts, 1998), monolingual corpora can still be compared if they consist of texts "sampled from different languages which are comparable in sampling criteria" (Xiao, 2007: 5). However, contrastive cross-linguistic research into learner language has typically only involved comparisons of either different stages of L2 development, L2 productions by speakers with different L1s, or L2 productions with native speaker productions (Tono, 2003 ; Granger, 2003).

Multilingual corpora of L2 development thus incorporate features of contrastive, comparable corpora, in that they need to be matched in terms of task-related, language-related and learner-related variables (Tono, 2003), and of developmental corpora, in that they represent various stages of language development. In addition to the requirements of comparability posed by contrastive research, the notion of comparability is extended to the data collection points, the time span and language development in a multilingual L2 corpus. Furthermore, if studies of a limited number of participants are partly limited to a description of general linguistic mechanisms, but cannot be generalized to give an impression of developmental tendencies (Granger, 2002), this limitation is even more amplified in the comparison of development in multilingual corpora.

Throughout the years, research departments have often gathered data and created corpora for a variety of research purposes. The gathering of these data is often the result of significant investments of time and resources and their potential for future research may remain untapped if they are discarded in favour of newly created corpora. In view of the difficulties of creating a corpus satisfying developmental and contrastive criteria, it may thus be more advisable to rely on previously gathered data instead.

In this paper, we will elaborate how we have, in the face of practical limitations, aligned two corpora to create a multilingual corpus representing four general proficiency levels in L2

French and English. We will discuss some of the encountered difficulties, potential solutions and the consequences for future analyses carried out on these data.

In the following sections, we will first describe the data that have been used as well as the research projects for which they will be employed. Next, we will present our considerations in creating the corpus, and finally discuss the implications for our research, as well as recommendations for future research.

2. Background: research projects and multilingual data

The data that will be presented were gathered over a number of years in light of several research projects on second language acquisition carried out at the Vrije Universiteit Brussel. In total, data from three research projects evaluating linguistic proficiency in mainstream Foreign Language Teaching (FLT) were used. In Flanders (Belgium), pupils are officially required to learn French and English at school. Foreign language education in French typically starts in the fifth year of primary school (ages 9-10), while English is introduced in the first or second year of secondary school (ages 11-13). At the end of secondary school, pupils will have received roughly 930 hours of French classes and 540 hours of English classes. In spite of this discrepancy in terms of exposure, the official final attainment levels set by the Flemish Ministry of Education are identical for both languages. Moreover, research has found that the head start for French may not result in a higher proficiency level for L2 French than for English (Spoelders, 1997).

The data cover a relatively wide period of L2 learning (6 years) by learners of the same linguistic and educational background. The data are oral narrative retellings of the "Frog Story" (Mayer 1969) in French and English by native speakers of Dutch, transcribed in CHAT format (MacWhinney, 2001). The participants relied on illustrations of the picture story and were asked to tell the story in a 10 minute interview with a researcher. Table 1 summarizes the general composition of the corpora in relation to their original research project.

Code	Language	Type	Age	Number	Original study
E1	English	L2	11-12	145	L2 French and English at the onset of secondary education in Flanders (unpublished)
F1	French	L2	11-12	171	
F3	French	L2	14-15	28	Dutch-language education in Brussels (Welcomme 2012)
E6	English	L2	17-18	87	L2 French and English at the end of secondary education in Flanders (De Clercq & Simoens, forthcoming)
F6	French	L2	17-18	131	

Table 1. Composition of the corpora

Two of these projects aimed at analyzing linguistic proficiency in L2 French and English, one at the onset of secondary school (E1, F1), the second at the end of secondary school (E6, F6), yielding four datasets. The fifth dataset was gathered in light of a study that aimed at evaluating the linguistic proficiency in Dutch and French of students in Brussels, the officially bilingual region in Belgium. This study also gathered data from a control group of 3rd year students in Flanders (F3). These were added in order to fill the gap between the F1 and F6 groups; due to the homogeneity of these two groups and the distance between them both in terms of cognitive development and years of L2 schooling of the pupils, we felt that the contrast would have been too vast without the addition of a third, intermediate category.

For the subset consisting of English oral productions, this appeared to be less necessary, considering the more limited disparity between the groups initially distinguished (E1 and E6). The relative heterogeneity in terms of actual L2 competence among the pupils in group E1 meant that a corpus consisting of only these two age groups gave a solid enough impression of pupils' possible English L2 competence across secondary school.

The need for a multilingual corpus uniting these different datasets arose from two research projects at the Vrije Universiteit Brussel (Belgium), aiming to describe and compare the development of complex clauses and of linguistic complexity in general in L2 French and English. These projects required corpora which covered a sufficiently large time span and which allowed for the comparison of linguistic development in the two languages of interest.

3. Creating the corpus

3.1. Finding common ground

The most important requirement for a multilingual corpus is arguably the equivalence of the monolingual corpora. A considerable number of parameters can be expected to influence the participants' language learning process and linguistic performance, including the age of the participants, their L1, the L2 learning context and the elicitation task. While it may not be possible to control all these factors, it is nonetheless important to control a number of key variables in order to ensure a common ground for comparison.

Figure 1 summarizes to what extent the two corpora are matched in terms of some key variables (adapted from (Tono, 2003)). Language- and task-related variables were controlled for both corpora. In terms of learner-related variables, both corpora consist of texts by L1 speakers of Dutch who learned L2 French and English in identical educational settings, that is, mainstream FLT. The participants age is only partially matched across the corpora, to the extent that there is no corresponding 13-14 year group in the English corpus, even though the upper and lower age limits are identical. The students' L2 proficiency is unmatched, since it was the dependent variable of the original studies. The design of the corpora is again comparable, with the exception of the extra data collection point for French.

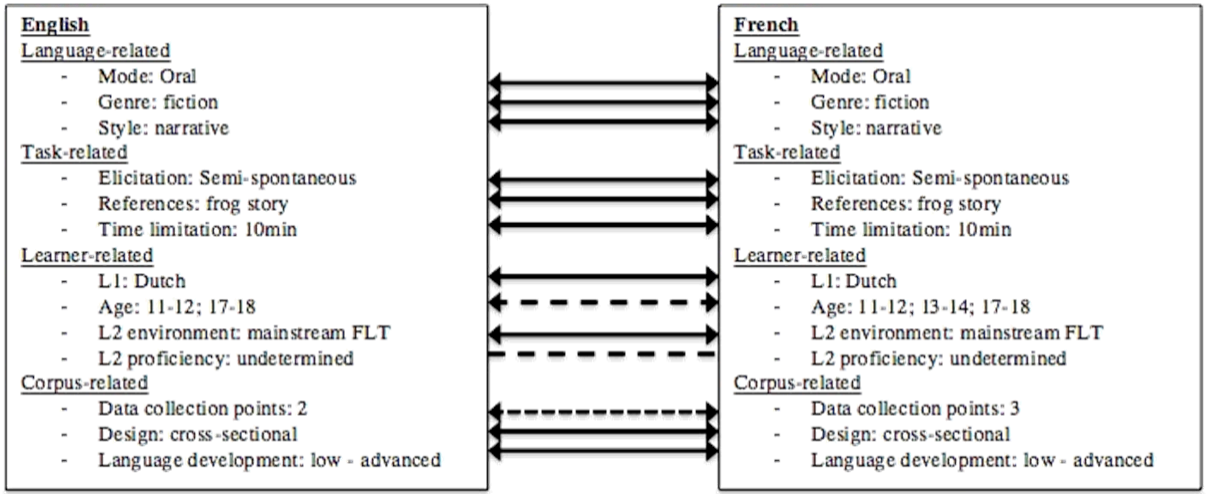

Figure 1. Comparison of the corpora for four types of variables

3.2. Quantifying proficiency

From the overview above, it transpires that the only unmatched variable in the corpora is the participants' L2 proficiency. The underlying reason is that, as the dependent variable of the original studies, linguistic proficiency could not be used a priori to match the corpora. Since the goal here is to arrive at a multilingual corpus of linguistic development, linguistic proficiency necessarily becomes an independent variable. In other words, the texts constituting a multilingual corpus need to be categorized in terms of the overall L2 proficiency level they reflect. The most straightforward way of determining proficiency post-hoc would be to subdivide the whole set of participants into groups based on their age (or, in cases where the subjects are school children or students, on their current grade). Of course, this assumes that L2 users' proficiency increases proportionally with their age or number of years of schooling, which is not always the case. In order to gauge a learner's general level of mastery in a given language, then, one could administer a standardized test (e.g. TOEFL, cloze tests (Oller, 1973)) developed for this purpose. Such tests have not been devised for all languages, however, and even if all participants contributing to a multilingual corpus can be subjected to a standardized test for any language concerned, this does not necessarily mean that the results are easily comparable between languages.

Alternatively, the corpus data, i.e. the participants' productions, can be analyzed in their own right. The past decades have seen the appearance of a number of computerized methods in this area. Despite the obvious advantages of automatic text analysis, however, the available methods making use of computerized tools have a number of drawbacks. Some were developed for the analysis of texts in one or more specific languages, excluding their use for the study of any others (such as Rapid Profile (Pienemann & Mackey, 1993), which can only be applied to English texts so far). Others, such as Coh-Metrix (Graesser et al., 2004), work only for the analysis of English written texts, whereas a large part of L2 development research is based on (transcriptions of) spoken data.

As a final option, custom proficiency measures can be applied to the multilingual corpus data at hand. The Complexity, Accuracy and Fluency framework (CAF ; Skehan, 1998 ; Ellis, 2003, 2008), for instance, combines concepts from various dimensions of language in order to estimate one's general level of proficiency in a given language. Previous research (Bulté & Housen, 2012 ; Wolfe-Quintero et al., 1998 ; Ellis & Barkhuizen, 2005 ; Verspoor et al., 2012) has shown the validity of some of these concepts as indicators of proficiency. Care should be taken, however, to represent the multidimensional nature of CAF : some measures of complexity may be more useful for early language development, but not for advanced learner language. Additionally, linguistic proficiency is a multidimensional construct as well and may not be accurately seized by a single measure. For instance, ever-increasing syntactic complexity may not be perceived as ever-increasing proficiency (Pallotti, 2009). For the purposes of our own research projects, we relied on a number of measures based on previous research in the CAF framework.

When deciding on the exact measures to include in our research, we had to take into account a number of limitations. The calculation of the fluency of L2 learners' oral productions would have been too time-consuming for our purposes. In addition, the measures were required to be independent of the variables of interest for the research projects. In consequence, generally accepted complexity indices could not be used either. As a result, we based ourselves on

measures of accuracy, which were found to be fairly solid indicators of general linguistic proficiency in earlier research (Wolfe-Quintero et al., 1998 ; Verspoor et al., 2012).

A further requirement for proficiency measures is that they are operationalizable in the different languages of interest, i.e. French and English. This excluded, for instance, correct choice of article or adjective gender as a measure, given it would be applicable to French only.

We eventually opted for a combination of two specific different accuracy measures, the proportion of correct prepositions and verb conjugations, and one general accuracy measure, the proportion of error-free clauses. These measures were calculated manually for each transcription. Texts with less than five verbal groups or prepositions were disregarded.

Table 2 below offers an overview of the average scores obtained within each age-based group for each of the three measures of accuracy applied to the data. As explained above, only French data were available for the third-year group. Comparing groups E1 and F1, we can observe that the accuracy for the English group is significantly higher in terms of correct preposition use and general accuracy ($p < 0.05$), but not for correct verbal morphology, which is significantly higher for French ($p < 0.05$). The results of groups E6 and F6 are more comparable, with percentage of error-free clauses as the only significant difference ($p < 0.05$).

		Preposition		**Morphology**		**Error-free**	
	n	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
E1	120	66.73	18.51	47.12	29.23	38.05	22.44
F1	153	53.00	27.17	54.81	23.00	19.01	14.84
F3	27	66.17	18.74	74.79	21.60	34.4	18.78
E6	85	88.51	8.46	92.79	12.03	82.06	8.77
F6	106	88.15	9.38	93.64	6.22	64.33	14.58

Table 2. Average scores for all three measures and all five original groups

3.3. Determining proficiency levels

If CAF measures are used to triangulate linguistic proficiency, it should be kept in mind that CAF measures essentially yield descriptive, continuous scores which cannot directly be transformed into proficiency levels. With the purpose of deriving a classification of data into proficiency levels from a number of CAF measures, a couple of considerations are in order. Firstly, it should be verified whether the chosen measures function in the same way for each language. Even if CAF measures can be operationalized in different languages, they may function differently. In our case, the lower scores for the error-free clause measure in French do not necessarily indicate lower proficiency levels. Instead, the measure is more sensitive to some of the particularities of French. For instance, if a learner uses the wrong article gender for a common word (e.g. “un grenouille” instead of “une”), each sentence containing that word will diminish the error-free clause score.

Secondly, the scores obviously need to display sufficient variation in order to distinguish between proficiency levels. If the distribution of CAF scores is too narrow, the measure may

not be a good discriminator of proficiency. Alternatively, low variability may simply indicate that the proficiency levels in the datasets are more homogeneous than anticipated.

Thirdly, it must be verified whether the supposed relationship between the measures and general language proficiency is valid. Ideally, the more different dimensions of CAF analyzed, the more precise the triangulation of proficiency. If the number of usable measures is restricted due to an element of the CAF triad being a variable of interest (e.g. complexity, as is the case in our research), a manual verification of the transcripts may be in order. In the case of accuracy measures, for instance, a learner's transcript may receive high scores, not because s/he is very proficient, but because s/he only uses simple repetitive utterances.

These considerations should inform the choice of cut-off points in the proficiency scores. If the measures do not function similarly in the different languages, it may be more interesting to determine cut-off points for each language separately (e.g. on the basis of the distribution of scores within each language). If the measures do not always accurately reflect the participants' linguistic proficiency, an additional criterion could be used to determine proficiency levels. In our case, it could be assumed, for instance, that first grade pupils will unlikely achieve the same levels of mastery as sixth grade pupils, and would thus be excluded from higher proficiency levels.

By means of these principles, participants were given a rank for their scores on each measure. On the basis of these ranks, the texts were categorized in one of four proficiency levels. The first two proficiency levels can be described as beginner and low intermediate levels respectively, and only contain texts by pupils from the groups E1, F1 and F3. The two higher proficiency levels were tentatively labeled high-intermediate and advanced respectively and consist of narratives from the groups E6 and F6.

Table 3 below provides the average scores obtained within each new group for the three accuracy measures established earlier, and adds the minimum and maximum values and standard deviation for each combination of group and measure.

			Preposition		Morphology		Error-free	
	Level	**n**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
French	1	74	35.7	23.8	38.2	18.6	9.2	7.0
	2	92	66.9	19.1	68.9	16.0	27.6	14.9
	3	62	81.6	9.6	89.3	6.7	51.6	11.7
	4	58	92.9	7.2	96.9	3.3	73.0	12.0
English	1	55	55.4	17.9	31.0	25.5	18.6	11.4
	2	65	76.3	12.8	60.7	25.1	54.5	15.1
	3	31	83.8	9.0	84.6	16.2	74.5	8.0
	4	54	91.2	6.9	97.5	4.3	86.4	5.7

Table 3. Average scores for all three measures and all four new groups per language

The means for each of the three measures increase without exception from one level to the next, both in the French and in the English L2 data.

In the French part of the corpus, the proportion of correctly used prepositions according to the target language rules is 35.7% at level 1 and increases significantly ($p < 0.05$) with each level, with a value of 66.9% for level 2, 81.6% for level 3, and 92.9% correctly used prepositions at level 4. A similar tendency can be observed for the measure gauging target-like use of verbal morphology, increasing from 38.2% (level 1) over 68.9% (level 2) and 89.3% (level 3) to 96.9% (level 4). For this measure as well, all differences between groups are significant ($p <$

0.05). The third index, measuring the proportion of error-free clauses in the targeted L2, yields only 9.2% error-free clauses at level 1, and then rises remarkably to 27.6% (level 2), 51.6% (level 3) and finally 73.0% of clauses containing no errors whatsoever (level 4). Once more, all four levels differ significantly from one another ($p < 0.05$).

As for the data produced by the learners in English as a second language, similar observations to those for French can be seen, though from the first level onwards, scores are usually noticeably higher for English than they were for the same measure in French as an L2. As explained before, this is probably due both to the difference in status of and exposure to the languages in question, and to the different applicability of some of the measures involved (particularly the proportion of error-free clauses). For the first measure, the proportion of correctly used prepositions, the average score for level 1 amounts to 55.4% and gradually increases to 76.3% (level 2), 83.8% (level 3) and finally 91.2% (level 4). Statistical analysis shows, however, that levels 2 and 3 do not differ in a statistically significant way ($p = 0.124$). The other groups' scores are significantly different from one another ($p < 0.05$). Verb morphology, in this case the correct conjugation of the third person singular present, is accurate in only 31.0% of cases at level 1, yet rises quickly to 60.7% at level 2, then 84.6% (level 3) and no less than 97.5% (level 4). With respect to this measure, all levels differ statistically significantly from one another. The same development was observed for the proportion of error-free clauses, reaching a score of only 18.6% at level 1, yet increasing over 54.5% (level 2) and 74.5% (level 3) to a final score of 86.4% (level 4). In this case as well, all contrasts between the four groups are statistically significant ($p < 0.05$).

The number of texts per level was then limited to 25 randomly selected narratives for each language involved, meaning that 100 texts were included per language and that the full sample under scrutiny consists of 200 texts.

3.4. Using the corpus

In the final version of our corpus, the participants' linguistic proficiency has been turned into an independent variable for each language, so that the resulting corpus is also partially matched for proficiency. Importantly, there is no guarantee that the proficiency levels in each language are equal. This is due to a number of reasons. Firstly, previous studies on the individual corpora have indicated that proficiency levels tend to be slightly higher for English than for French (De Clercq & Simoens, forthcoming). Secondly, because of linguistic differences in the target languages, the accuracy measures may have been more sensitive for one language than for the other. If a direct comparison of the proficiency levels is thus not advisable, it is nonetheless possible to compare developmental trends across the two L2s, since both monolingual corpora represent a similar development of proficiency, from low to advanced.

4. Conclusion

As has been shown in this paper, complete equivalence in a multilingual corpus of language development may be impossible to attain in practice. Due to a number of factors, including participant characteristics, time limitations, and the nature of the different languages under investigation, it may be more sensible to strive for sufficient comparability between the languages involved, instead of total equivalence in a multilingual corpus.

Based on our own research experience, we have attempted to describe the difficulties researchers may and will encounter when trying to establish a multilingual L2 corpus and

have proposed a number of tools and operations which can contribute to greater comparability between corpus languages. More specifically, we have combined insights from SLA research and contrastive linguistics in order to match similarly designed corpora on a number of key variables, with special attention to the quantification of linguistic proficiency as an independent variable.

Where research in the field of comparative L2 acquisition is concerned, this results in the comparison of tendencies in the development of linguistic structures, rather than a direct comparison of the object of interest. Considering the difficulties in developing a fully equivalent multilingual corpus, this remark probably also holds true for other types of corpora as well. We hope that these considerations will help researchers to unlock the potential of existing, similarly designed datasets through contrastive research.

References

Aarts J. (1998). Introduction. In Johansson, S. and Oksefjell, S. editors, *Corpora and Cross-linguistic Research* (pp. ix-xiv). Amsterdam: Rodopi.

Bulté B. and Housen A. (2012). Defining and operationalising L2 complexity. In Housen A., Kuiken F. and Vedder I. editors, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy, and Fluency in SLA*. Amsterdam: John Benjamins, pp. 21-46.

De Clercq B. and Simoens H. (forthcoming). Two at the time: The parallel learning of French and English as foreign languages.

Ellis R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis R. (2008). *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis R. and Barkhuizen G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Graesser A. C., McNamara D. S., Louwerse M. M. and Zhiqiang C. (2004). Coh-Metrix: Analysis of text on cohesion and language. In *Behavior Research Methods, Instruments, & Computers*, 36, pp. 193-202.

Granger S. (2002). A Bird's eye view of learner corpus research. In Granger S., Hung J., and Petch-Tyson S. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, pp. 3-33.

Granger S. (2003). The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies. In Granger S., Lerot J. and Petch-Tyson S. editors, *Corpus-based approaches to contrastive linguistics and translation studies*. Amsterdam: Rodopi, pp. 17-29.

Housen A. and Kuiken F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. In *Applied Linguistics*, 30(4), pp. 461-473.

Larsen-Freeman D. (2009). Adjusting Expectations: The Study of Complexity, Accuracy, and Fluency in Second Language Acquisition. In *Applied Linguistics*, 30(4), pp. 579-589.

Laviosa S. (1997). How comparable can comparable corpora be? In *Target, 9*(2), 289-319.

MacWhinney B. (2001). From CHILDES to TalkBank. In: Almgren M., Barreña A., Ezeizaberrena M., Idiazabal I. and MacWhinney B. editors, *Research on Child Language Acquisition*. Somerville: Cascadilla, pp. 17-34.

Mayer M. (1969). *Frog, where are you?* New York: Dial Books.

McEnery A. and Xiao R. (2007). Parallel and comparable corpora: What is happening. In Anderman G. and Rogers M. editors, *Incorporating Corpora. The Linguist and the Translator*. Clevedon: Multilingual Matters, pp. 18-31.

Myles F. (2005). Interlanguage corpora and second language acquisition research. In *Second Language Research*, 21(4), pp. 373-391.

Oller J. W. Jr. (1973). Cloze tests of second language proficiency and what they measure. In *Language Learning*, 23(1), pp. 105-118.

Pallotti G. (2009). CAF: Defining, refining and differentiating constructs. In *Applied Linguistics, 30*(4), 590-601.

Pienemann M. and Mackey A. (1993). An empirical study of children's ESL development and Rapid Profile. In McKay P. editor, *ESL Development. Language and Literacy in Schools*, Vol. 2, Commonwealth of Australia and National Languages and Literacy Institute of Australia, pp. 125-259.

Skehan P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Spoelders M. (1997). Frans op de basisschool. Evaluatie van de onderwijspraktijk en voorstellen tot optimalisering (OBPWO project 95.05). Brussels: Ministerie van Onderwijs van de Vlaamse Gemeenschap.

Teubert W. (1996). Comparable or parallel corpora? In *International journal of lexicography, 9*(3), 238-264.

Tono Y. (2003). *Learner corpora: design, development and applications*. Paper presented at the 2003 Corpus Linguistics Conference.

Verspoor M., De Bot K. and Lowie W. editors. (2011). *A dynamic approach to second language development: Methods and techniques*, Vol. 29. Amsterdam: John Benjamins.

Verspoor M., Schmid M. S. and Xu X. (2012). A dynamic usage based perspective on L2 writing. In *Journal of Second Language Writing* 21(3), pp. 239-263.

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

Wolfe-Quintero K., Inagaki S. and Kim H.-Y. (1998). *Second language development in writing: Measures of Fluency, Accuracy, and Complexity*. University of Hawai'i, Second Language Teaching and Curriculum Center.

Xiao R. (2007). What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. In *Indonesian Journal of English Language Teaching, 3*(2), 1-19.