LUBLIN STUDIES IN MODERN LANGUAGES AND
LITERATURE 37, 2013, HTTP://WWW.LSMLL.UMCS.LUBLIN.PL

Vita Kalnberzina, Vineta Rutenberga
University of Latvia,
Raiņa bulvāris 19, Rīga, LV-1050, Latvia

Syntactic indicators of language acquisition levels in English and French written language learner corpora

1. Introduction

Learner corpus research is based on collecting samples of student writing and examining them, which is not unlike writing assessment, where we develop tasks and collect the elicited learner texts to examine them using different criteria. These can range from orthographic, to lexical, syntactic and discourse level to ensure validity of assessment. Validation of writing assessment systems involves examination of the theoretical construct underlying the tasks and the assessment criteria, such as grammar, vocabulary, spelling and task achievement reliability of the marking and validity of the score interpretation (see for example Shaw & Weir's 2007 socio-cognitive writing test validation framework).

This kind of validation procedure should be sufficient for a monolingual examination system. However, when we are relating multilingual writing assessment systems, we are mostly advised to calibrate items, and relate the results statistically (see *e.g.* North *et al.* 2009). This works for dichotomous items, but is of little help when we

are dealing with comparing examinations based on essays produced in different languages, even if the writing tasks have been translated and moderated by experts and the marking scales for assessment are in the marker's native language.

One solution to the problem is to use learner corpora both for the training of the task developers and assessors as well as for benchmarking purposes. The data used are mainly informal collections of essays representing each level from every year to ensure the comparability of the levels of assessment from year to year. Sample papers are also used for relating different examinations across countries using the Common European Framework of Reference for Languages (the CEFR). The Manual for Relating Language Examinations to the CEFR[1] is often used together with sample scripts with comments on the Council of Europe website[2], which are highly appreciated by the teachers involved in task and assessment grid development. Other corpora that can be used are formal test taker corpora compiled by the examining bodies as research tools and databases to develop and validate language tests and provide evidence of spoken and written performance. For example, the Cambridge Learner Corpus (CLC)[3] containing 20 million words (58,000 exam scripts of the whole range of exams) serves as an archive of test formats and responses (learner corpora), and supports the existing statistical and other test validation procedures (Barker 2006). Although previous CLC research has mostly focused on lexical analysis, e.g. updating item writer and syllabus word lists for various examinations, analysing candidates' business lexis, comparing candidates' written and spoken vocabulary with the existing word lists, and investigating the influence of varieties of English on candidates' written vocabulary, in the latest publication of English profile (2011), grammatical criteria are included in language level

[1] www.coe.int/t/dg4/linguistic/manuel1_en.asp
[2] http://www.coe.int/t/dg4/portfolio/documents/exampleswriting.pdf
[3] www.cambridge.org/elt/corpus

description among other features of writing at the six proficiency levels (A1–C2) of the CEFR (Council of Europe 2001).

Recently, language corpora have been extensively used in contrastive linguistics as they facilitate the language acquisition research, which, according to Granger (2010:1), can meet our interlanguage and intercultural communication needs.

As corpus research has been expanding, so have its areas of application and the issues addressed by corpus researchers. According to Tono (2002), when we are building a new learner corpus, we need to take into account three different groups of criteria:

(a) language-related criteria (*e.g*. mode, medium, genre, topic),
(b) task-related criteria (*e.g.* longitudinal vs. cross-sectional; spontaneous vs. prepared),
(c) learner-related criteria (*e.g*. EFL or ESL, age, sex, mother tongue, overseas experience).

From this we can conclude that corpus researchers not only have to keep track of diverse criteria while developing their corpus, but can also answer questions regarding the three categories: not only dealing with linguistic parameters, but also concerning tasks and learners across languages. This allows us to suggest that corpora can be used as a test validation tool to provide evidence on reliability, validity and impact of the measurement of linguistic, task-related and learner-related criteria. The latter function, that of an additional validation procedure, is the focus of this article, as we will use English and French corpora to validate the examination levels in Latvian Year 12 examinations in English and French by comparing the frequency of use of complex sentences in different language performance levels and contrasting their use to the native speaker patterns of use reported in Cosme (2004).

2. Research context

The situation of language examination validation in Latvia differs in case of English and French language examinations. The process of Year 12 English language examination validation in Latvia was started as soon as the system was developed: see *e.g.* Kalnberzina

(2002) for the qualitative validation of Year 12 examination, Kalnberzina (2007) for the qualitative relation of Year 12 writing examination to the CEFR and Kunda (2011) for the quantitative relation. As a result, a tentative relationship was established with the CEFR levels and the Latvian Year 12 foreign language examination levels are aiming at the CEFR level B2, with the top performances being related to C1 (level A in Latvia). The task developers and markers used the system developed by the English language examination to establish comparability with other language examinations (French, German, Russian and Latvian) via the school curriculum, test specifications and assessment scales which are all based on the CEFR levels. An additional means of standardisation across language examinations are statistical procedures for grade awarding: all the examination results are routinely processed to calculate the mean, the standard deviation and grade the students' performances using the distribution curve. However, there have been no formal studies on French examination validity. Therefore, the present research can be considered as the first attempt to use linguistic features to validate the French examination levels.

The lack of formal validation for the French examination has led the examination centre to doubt the reliability of the assessment levels, the hypothesis being that the uniform statistical grading procedure has possibly created a discrepancy between English and French language acquisition levels. This is due to the differences in the population of the examination: English language examinations are taken by the whole population (19,169 students in 2012), while French is taken only by the students studying in specialised language schools (49 students in 2012). Although the distribution curves of the writing test in both languages are normal, the standard deviations and means differ. In the French examination the standard deviation is 11% and in the English examination - 24%, whereas the mean in the English examination is 50%, while in the French examination it is 66%, suggesting that the French examination is easier and the test developers have been pressurized to make the examination more demanding to correspond to the English language examination

statistics. The contrastive analysis of the French and English learner corpora is an attempt to examine the claim that the French examination is easier than the English examination based on a contrastive analysis of the two learner corpora.

Granger's (2010:3) typology of corpora distinguishes between monolingual and multilingual corpora. In our case the corpora are multilingual, as we set out to compare the texts produced by Latvian and/or Russian students writing test essays in English and French and assessed by Latvian and/or Russian markers. We examine the syntax of the language learners of French and English, because in contrast to lexical and morphological structures sentence structures are comparable across languages.

The CEFR, whose levels serve as the basis of the secondary school foreign language curriculum and test specifications, identifies the linguistic structures that foreign language learners should know at a certain level of language proficiency. For example, at the Threshold level[4] learners: 1) should be able to understand and produce simple and compound sentences; 2) should be *expected* to produce complex sentences which are straightforward in character, e.g. limited to one subordinate clause of fairly simple structure with a main clause frame of a basic character; 3) should be able to understand embedded clauses. At Vantage level[5] learners should be able to understand and produce simple, compound and complex sentences.

The question that we are addressing is whether the levels of language performance in French and English are comparable. According to Pienemann's Processability theory (1999), the first stage in language acquisition is attributed to a word, which is followed by the processes related to the word category. After that the learner builds phrases that form sentences with their morphology, and, finally, subordinate clauses are produced at the very last stage of language acquisition. What is more, each procedure has its time boundaries, i.e.

[4] www.coe.int/t/dg4/linguistic/dnr_EN.asp
[5] http://www.coe.int/t/dg4/linguistic/Vantage_CUP.pdf

no other procedure in the hierarchy can take place if the previous one has not been accomplished.

Hence, out of the five levels, we have decided to focus our attention on subordinate clauses as a preliminary analysis suggests that their number increases at the higher levels of language proficiency not only in foreign language, but also in second language use in both primary and secondary language examination (see Kunda 2011).

3. Research procedure

The present study is a corpus-based research of syntactic structures. When compiling the corpora, the written essays of year 2009 centralised examination in English and French were chosen as a sampling unit, since essays are defined similarly in all foreign languages test specifications, which allows us to ensure the comparability of the texts produced by the English and French test-takers. In 2009 the English language test-takers had to write an essay about 'Reasons for leaving Latvia':

> One of the main reasons why people have left Latvia during the last few years is that they say they are better paid in other countries. Add two other reasons and discuss all of them in an essay, giving your own opinion.

In French the theme of the essay was:

> Pensez vous qu'il soit encore utile d'apprendre des langues étrangères alors que l'anglais est actuellement la langue de communication mondiale (échanges commerciaux, économiques, politiques...)? Présentez votre réflection de façon argumentée. (Do you think that it is still useful to learn foreign languages as nowadays English is the language of communication (in business, economics, politics…) in the world? Give your point of view by providing arguments.)

The essays, whose length ranged from 404 tokens to 13 tokens, were classified according to the level obtained at the local examination (see Table 1 below). It should be specified that there were no texts of levels E/A1 and F in French as the number of test-takers per year does not exceed 100 (in 2010 it was 71; in 2011 - 77; in 2012 - 49) and they are mainly pupils from language schools. Moreover, the lowest level F does not correspond to any of the CEFR

proficiency level descriptions, as the produced pieces of writing are very poor.

Consequently, the compiled English learner corpus consists of 44,387 tokens, while the French learner language corpus contains 28,378 tokens.

Table 1. Nr of tokens per language performance level in English and French learner corpora.

	Total	A/C1	B/B2	C/B1	D/A2	E/A1	F
Nr of tokens in English learner corpus per level	44,387	5,193	11,526	10,277	9,446	6,266	1,679
Nr of tokens in French learner corpus per level	28,378	5,279	11,908	10,311	880		

Furthermore, all the essays were transcribed and all the sentences were classified into simple, compound and complex sentences. According to Jackson (2007), a simple sentence is composed of a single main clause (e.g. *He was very happy about the results.)*; a compound sentence contains at least two main clauses in a relation of coordination (e.g. *Robert went to the cinema and his sister watched television.*) and a complex sentence consists of a main clause and at least one subordinate clause (e.g. *The number of people who have left Latvia has increased*).

Subsequently, the focus was attributed to the finite embedded constructions taking into consideration Dik's (1997) taxonomy of embedded constructions. According to Dik, we distinguish between finite and non-finite embedded constructions (Figure 1). The finite constructions are the ones in which "the predicate can be specified for the distinctions which are also characteristic of main clause predicates" (Dik 1997:144). Moreover, only finite embedded constructions make subordination.

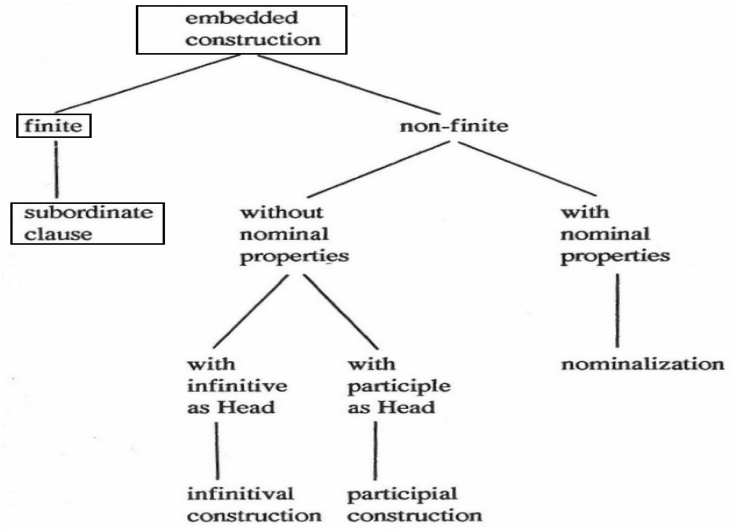

Figure 1. Taxonomy of embedded constructions (Dik, 1997).

The obtained results were compared with Cosme's (2004) research data on the native speaker use of finite complex sentences, as she developed a cross-linguistic corpus to equate various clause-linking patterns in comparable (authentic) corpora to collate the cases of subordination and coordination in three languages – English, French and Dutch.

Finally, complex sentences were classified into three groups according to the first subordination, which followed directly the main clause. Thus, we distinguish: 1) a nominal clause - a type of subordinate clause that functions in sentence structure where noun phrases usually occur (My intuition says *that the government will soon fall*.); 2) an adjectival clause – a type of subordinate clause that functions like an adjective, i.e. 'describes' a noun (It is our duty to help those *who are in trouble*.) and 3) an adverbial clause – a type of subordinate clause which functions as an adverbial in sentences (*When*

I was a little girl, I lived in the countryside with my grandparents.) (Jackson 2007).

3. Research results and discussions

The research data of different types of sentences show that the frequency of complex sentences in both languages differs across levels of language proficiency.

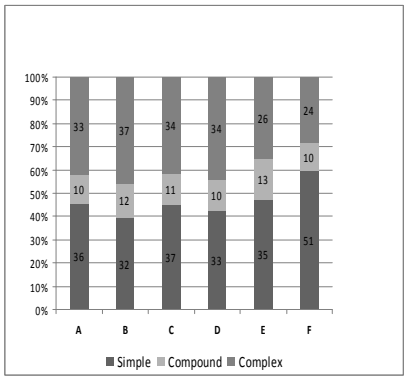

Figure 2. The frequency of clause types in English learner corpus.

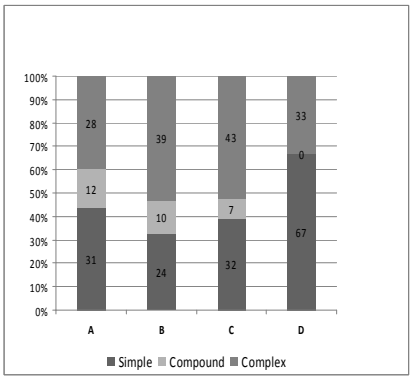

Figure 3. The frequency of clause types in French learner corpus.

In English (Figure 2) they constitute 24% at level F, and then gradually rise to 34% at levels C and D attaining their peak at level B (37%). In French (Figure 3) the complex sentences are unevenly distributed. The majority of them appear at levels C (43%) and B (39%).

If we examine the frequency of complex sentences containing a finite subordinate clause, the data reveal (Figure 4) that there is a different pattern for the raw frequency of the use of subordinate clauses in English and French. We can observe an increase towards the highest levels of language proficiency, i.e. A - C in the use of complex sentences in both languages. However, in French there is a peak already at level C, which corresponds to the Threshold level

descriptors. This, according to the CEFR, is the level where students only start producing the simplest type of complex sentences, in which the relative pronoun functions as subject, e.g. *A lot of them are young people who are getting education abroad.* Therefore, the French examination markers in Latvia have not given them the top mark.

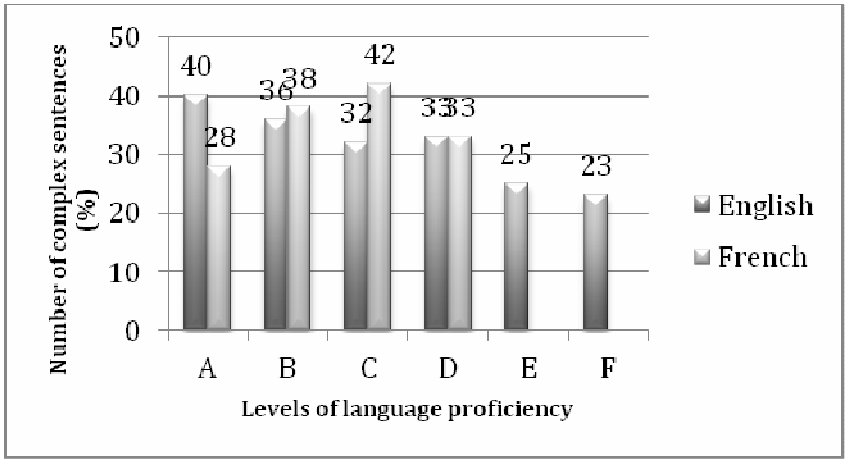

Figure 4. Comparison of the frequency of complex sentences containing finite sub-clauses in English and French.

The number of complex sentences diminishes at lower levels of language proficiency in both languages. At these levels the students do not use appropriate subordinate conjunctions, they start the sentence with a coordinating conjunction, although it is irrelevant and inappropriate, or avoid the conjunctions at all. They do not discriminate between restrictive and non-restrictive relative clauses, which is of utmost importance in English. The difficulties in discriminating among different clause types could be observed already at level C/B1, though the tendency is not as visible as at levels D/A2 and E/A1. At levels D/A2 and E/A1 many students of English use the adjectival clause in which they state the reasons for leaving Latvia

(*e.g. And there I have come to the second reason why people have left Latvia*.). Such adjectival clauses do not reveal their level of proficiency, as this clause type is included in the task rubric which they have just copied.

Although the number of complex sentences differs across levels and languages, the tendency of complex sentence frequency of use agrees with both the CEFR and Pienemann's Processability theory, i.e. their frequency of use increases towards the highest levels of language proficiency.

If we compare our research data with Cosme's (2004) findings on native speaker use of finite complex sentences (Table 2), we see that the native speakers (NS) use subordination more than the learners of English and French. Thus, according to Cosme, 46% of the complex sentences marked in the French native speaker corpus contain finite subordinate clauses versus 70% in the English native speaker corpus, whereas the learners of English produced on average 31.5% of sub-clauses and the learners of French – 35% of sub-clauses.

Table 2. Proportion of complex sentences containing finite sub-clauses across examination levels in English and French, and in Cosme's (2004) native speaker corpora.

	NS (Cosme)	A/C1	B/B2	C/B1	D/A2	E/A1	F
Complex sentences in English (%)	**70**	40	36	32	33	25	23
Complex sentences in French (%)	**46**	28	38	42	33	-	-

The subsequent analysis of different clause types demonstrates (Figure 5) that the distribution of *nominal clauses* in English is rather uneven, ranging from 39% at level D; 38% at levels A and F to 35% at level E; 33% at level C and then slightly falling at level B to 29%. However, this clause type has been used at all levels of language proficiency only with a small fluctuation. *The number of adverbial*

clauses increases towards the lowest levels of language proficiency, attaining the highest number at level F – 52%. Yet, there is a considerable fall at level D – 24%. As for *adjectival clauses*, their distribution across levels is diametrically opposed to the distribution of adverbial clauses. In English the number of adjectival clauses constitutes 40% at levels A and B, then considerably falls at level C reaching only 26%. The numbers do not vary greatly from level C to level E. Then again there is a noticeable decrease at level F, where the numbers reach only 10%.

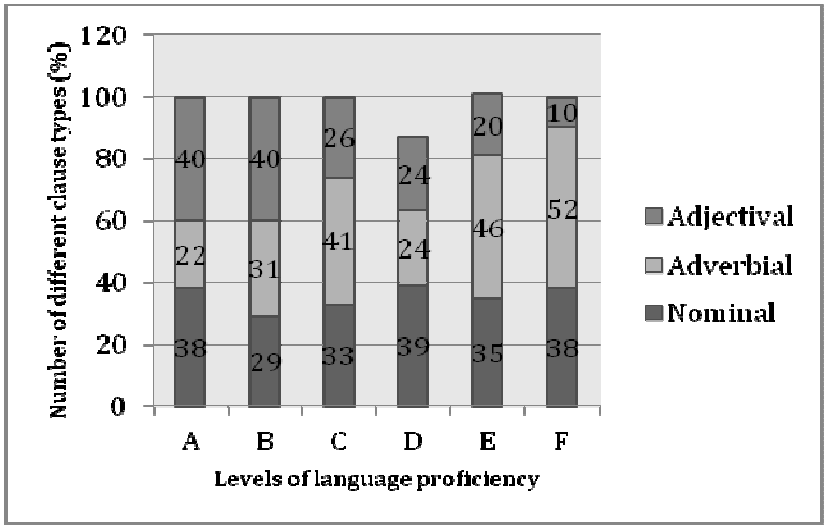

Figure 5. Frequency of subordinate clauses in English.

In French (Figure 6) the frequency of *nominal clauses* is rather similar to English (ranging from 29% at level A to 32% at level C and 34% at level B). At level D the numbers reach 100% as at this level of language proficiency there are just 4 complex sentences and all of them contain a nominal clause. The frequency of *adverbial clauses* is rather stable at all levels comprising on average 45%. Adjectival

clauses have been used the least effectively at all levels (their numbers vary from 20% to 26%).

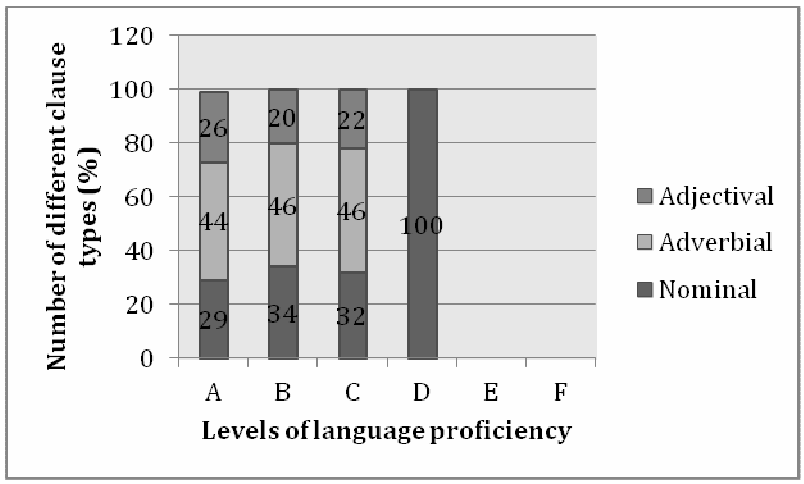

Figure 6. Frequency of subordinate clauses in French.

The results of the present contrastive analysis support the assumption based on Pienemann's Processability theory that syntax is one of the parameters signalling a certain level of language acquisition. We also see that subordinate clauses can serve as a criterial feature for attributing higher marks in language examinations. However, the number of subordinate clauses used by the test-takers differs in English and French essays of the same level, which may indicate either the misinterpretation of the assessment criteria or the problem of reliability of the assessors.

4. Conclusion

The focus of the study was the comparison of syntactic features in English and French examination corpora as a means of validation of

Year 12 writing examinations. Our main findings concerning the frequency of syntactic patterns are:

1. the number of complex sentences rises in both English and French learner language with the increase of the language acquisition levels, thus suggesting that the examinations are comparable;
2. the pattern of use of complex sentences agrees with Pienemann's Processability theory and the Common European Framework of Reference level description, which suggests construct validity of the English and French writing examinations;
3. the native speaker frequency of use of subordinate clauses in Cosme's (2004) corpus is higher than that of learner corpora, which could be expected, but further research is necessary to compare our findings to larger native speaker corpora;
4. the peak of the frequency of use in both English and French language learner corpora were at level B2 , which suggests the need for deeper analysis of the corpora as well as further test validation procedures to examine the causes;
5. the patterns of use of the nominal, adverbial and adjectival subordinate clauses differ in English and French learner corpora, which suggests a need for further research in both native speaker corpora and/or other language learner corpora.

As regards the methodology of corpus linguistics and contrastive analysis, manual transcription and tagging is an incredibly meticulous and time-consuming approach, especially at the lower language acquisition levels, where it is difficult to tell apart not only the syntactic patterns, but even words and letters. However, when the texts have been transcribed and tagged, it is possible to compare the syntactic patterns across language acquisition levels as well across languages, and even small learner corpora can offer new insights into test data.

References

Barker, F. (2006): Corpora and language assessment: Trends and prospects. *Research Notes* 26, 2-4.

Cosme, C. (2004): Towards a corpus-based cross-linguistic study of clause combining. Methodological framework and preliminary results. *Belgian Journal of English Language and Literatures* (BELL). New Series 2, 2004, p. 199-224.

Council of Europe (2001): Common European Framework of Reference for Languages. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf (last accessed on 14 January, 2012)

Dik, S.C., ed. Hengeveld, K. (1997): *The Theory of Functional Grammar, Part 2: Complex and Derived*. Berlin: Mouton de Gruyter.

English Profile. Introducing the CEFR for English. (2011). Cambridge: Cambridge University Press. http://www.EnglishProfile.org (last accessed on January 14, 2012).

Granger, S. (2010). *Comparable and Translation Corpora in Cross-linguistic Research: Design, Analysis and Applications*. Centre for English Corpus Linguistics, Université catholique de Louvain. http://sites.uclouvain.be/cecl/archives/Granger_Crosslinguistic_research.pdf (last accessed on January 14, 2012).

Jackson, H. (2007): *Key terms in linguistics*. London: Continuum.

Kalnberzina, V. (2002): *Interaction between meta-cognitive and affective variables in language performance: The case of test anxiety.* Doctoral dissertation, Lancaster University.

Kalnberzina, V. (2007): *Impact of Relation of Year 12 English Language Examination to CEFR on the Year 12 Writing Test*. Paper presented at the International Conference of the FIPLV Nordic-Baltic Region "Innovations in language teaching and learning in the multicultural context" LVASA.

Kunda, T. (2011): *Relating Latvian Year 12 Examination in English to the CEFR.* http://visc.gov.lv/eksameni/vispizgl/dokumenti/20110920_petijums_en.pdf (last accessed on 14 January, 2012).

North, B., Figueras, N., Takal, S., Van Avermaet, P. Verhelst, N. (2009): *Manual for Language test development and examining for use with the CEFR – produced by ALTE on behalf of the Language Policy Division.* Strasbourg: Council of Europe.

Pienemann, M. (1999): *Language Processing and Second Language Development: Processability Theory.* Amsterdam/Philadelphia: John Benjamins.

Shaw, S. & Weir, C.J. (2007): *Examining Writing in a Second Language*. *Studies in Language Testing* 26. Cambridge: Cambridge University Press and Cambridge ESO.

Tono, Y. (2002). *Learner corpora: Design, development and applications*. Graduate School of Applied Linguistics, Meikai University, JAPAN. Paper presented at the Corpus Linguistics 2003 Conference (CL 2003), Lancaster.