# The phraseological errors of French-, German- and Spanish-speaking EFL learners: evidence from an error-tagged learner corpus

Jennifer Thewissen
Centre for English Corpus Linguistics, Université catholique de Louvain, Belgium

*Abstract*
*The value of learner corpora in the field of learner phraseology has been convincingly illustrated in a number of corpus studies (Granger 1998, Nesselhauf 2003, 2005, Paquot 2007). In addition, the growing attention paid to phraseological errors (Nesselhauf 2003, Osborne 2008, Wang and Shaw 2008) shows that phraseology still very much remains a linguistic "bête noire" even for the more advanced learners. In this study we look at several types of phraseological errors committed by three learner populations, viz. French- German- and Spanish- EFL learners. We do so by using a learner corpus which has been (a) fully error tagged, (b) divided into mother tongue backgrounds, (c) stratified into proficiency levels. This paper reports on two main analyses: (1) we provide an overview of several types of phraseological errors in the three learner populations by basing ourselves on the typology of phrasemes recently developed by Granger and Paquot (forthcoming 2008), (2) we then carry out an analysis of phraseological errors in terms of grammaticality vs acceptability errors (James 1998). The TaLC presentation itself will additionally look at the phraseological errors in the corpus (a) from the point of view of potential L1 influence, i.e. we determine how many phraseological errors in the three populations can be traced back to the learners' L1, and (b) from the point of view of language proficiency, i.e. we investigate whether the number and type of phraseological errors differ according to the proficiency level.*

Keywords: learner corpora, error tagging, phraseological errors, mother tongue backgrounds, proficiency levels

## Introduction

The current phraseological boom is evidenced by a series of new publications, especially the phraseology volumes by Granger and Meunier (forthcoming 2008) and Meunier and Granger (2008). These volumes are testimony to the upsurge of academic interest in the field of phraseology but also reflect the widening of the scope of phraseology itself which is now seen to encompass multi-word units that would previously not have been considered as phraseological. Learner phraseology in particular has been arousing keen

interest among researchers. To this day, many aspects of learner phraseological use have been studied: collocations involving high-frequency verbs have been the focus of studies such as that by Altenberg and Granger (2001) who investigated learners' phraseological use of *make*; phrasal verbs were, among others, studied by Hulstijn and Marchena (1989) and Laufer and Eliasson (1993); recurrent word combinations were the focus of a thorough analysis by De Cock (2003). While these studies nicely show learners' patterns of over- and underuse in phraseology, they do not yet provide us with a general overview of the wide range of phraseological errors committed by EFL learners. The present paper addresses this issue by drawing a larger picture of the types of phraseological errors committed by three EFL populations, viz. French-, German- and Spanish- speaking learners. This paper is written within the larger context of my PhD project, the aim of which is, among others, to analyse learner phraseological errors across mother tongue backgrounds and proficiency levels by using a fully-error tagged learner corpus. This paper thus constitutes a first exploratory investigation of learner phraseological errors in the wider sense. This study is subdivided into two main parts: (1) we give a general breakdown of the phraseological error types in the corpus by basing ourselves on Granger and Paquot's (forthcoming 2008) classification of phrasemes, and (2) we look at the phraseological errors in terms of grammaticality and acceptability (2008) errors. The TaLC presentation will, in addition, interpret the phraseological errors in terms of both potential L1 influence and proficiency levels.

## Data and methodology

The learner corpus used here is the *International Corpus of Learner English* (ICLE) which consists of essays by learners from as many as 16 mother tongue backgrounds (Granger et al. 2002). Three learner populations are the object of this study, viz. French-, German-, and Spanish-speaking EFL learners (henceforth FR, GE and SP). As shown in Table 1, a total number of 223 learner essays were used in our analysis. Each learner text was submitted to a rigorous rating procedure: the texts were given to two professional raters who were asked to give each essay a Common European Framework grade (CEF)

(Council of Europe 2001) ranging from threshold level B1 to mastery level C2. In cases where the first two raters disagreed by more than one band score, a third rater was called in to rerate the problematic texts. The mean CEF score was calculated for each L1 subcorpus and is presented in Table 1. These results show that while ICLE can generally be said to represent advanced learner writing, it also contains texts that represent lower proficiency levels. Our FR and GE samples were both rated at the advanced C1 level while the SP sample was found to display B1 proficiency overall.

In addition to being independently rated by two, and when necessary three, professional raters, each text was error tagged by a native-speaker linguist, i.e. each text was manually annotated for errors. A total of about 50 000 tokens per subcorpus were error tagged following the guidelines in the Louvain Error Tagging Manual 1.2. (Dagneaux et al. 2005). Following the manual, each error in the corpus is preceded by a descriptive tag which explains the nature of the error and is followed by a possible correction in between dollar signs, as in the following sentence where the error was tagged LP for lexical phrase: *(…) this type of evasion is **(LP) at everybody's hand** $at everybody's disposal$* (FR). Table 1 describes the number of error-tagged essays and tokens in each subcorpus as well as the mean CEF score for each mother tongue background.

| Subcorpora | Number of essays | Overall tokens | Mean CEF score |
|---|---|---|---|
| FR | 74 | 50 558 | C1 |
| GE | 71 | 49 945 | C1 |
| SP | 78 | 51 860 | B1 |
| **Total** | **223** | **152 363** | |

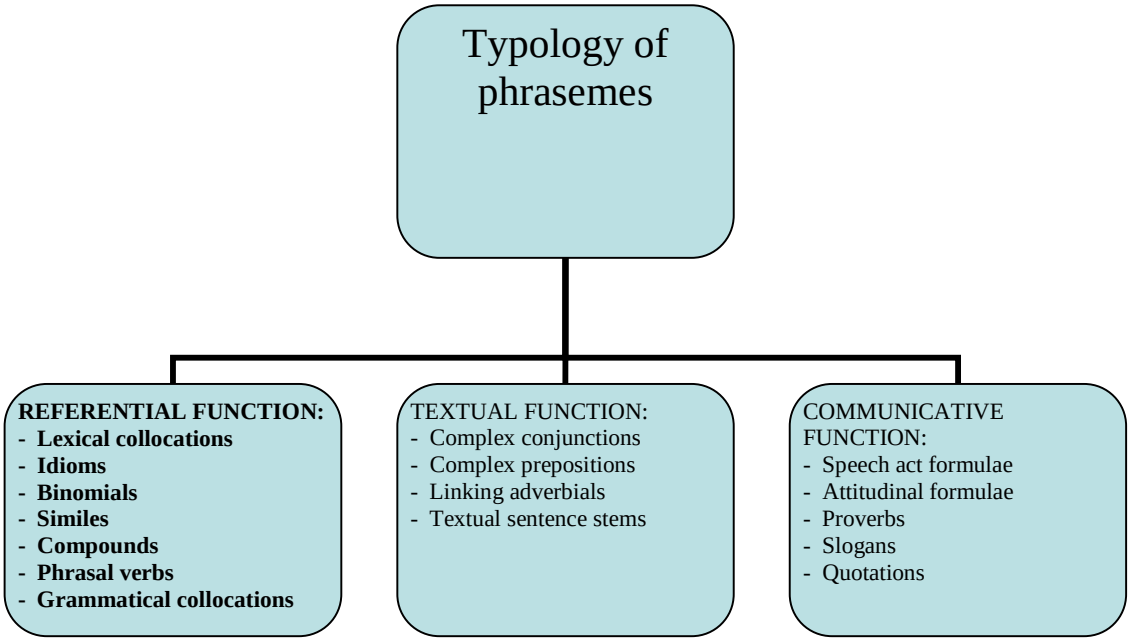
Table 1: Data description

Two of the 56 error tags of the Louvain Error Tagging Manual (Dagneaux et al 2005) will be analysed here: the LP tag which refers to lexical phrase errors and the X*PR tag, which refers to one specific subcategory of lexico-grammatical errors, viz. dependent preposition errors. Lexical phrase errors are lexical errors that affect word combinations, viz. compounds, idioms, phrasal verbs and some types of lexical collocations. The second tag, i.e. X*PR, targets dependent preposition errors. It is subcategorized according to the

grammatical category of the word the preposition is attached to: verb for XVPR, noun for XNPR and adjective for XADJPR.

A caveat of this study is that the LP and X*PR categories do not represent all the phraseological errors in the corpus. A number of other error categories also include phraseological errors and will be the subject of future research. Among the other tags that contain phraseological errors, we especially have LS, i.e. lexical single errors, which target errors in (a) isolated single words as in *he **(LS) affirms** $claims$ he is innocent* where the student confused two existing words, and in (b) lexical collocations where a single lexical word is erroneous as in ***(LS) high** $heavy$ responsibilities.* LS will thus need a considerable amount of weeding out to isolate the errors which are collocational in nature from those which affect words in isolation. In the meantime, because the LP and X*PR tags only represent part of the picture, the results presented here should be seen as exploratory.

All the LP and X*PR concordances in the FR, GE and SP subcorpora were extracted with WordSmith Tools 4 (Scott 2004) and were classified phraseologically following the typology of phrasemes recently developed by Granger and Paquot (forthcoming 2008). As the authors point out, this typology purposely adopts "a much wider perspective and includes many word combinations that would traditionally be considered to fall outside the scope of phraseology". Granger and Paquot's typology is presented in Graph 1 below, with the types of phrasemes found in the LP and X*PR categories highlighted in bold:




Graph 1: Typology of phrasemes (Granger & Paquot 2008)

It was nevertheless necessary to adapt the classification proposed in Graph 1 in order to classify the LP and X*PR errors into the different referential phraseme categories. The adaptation is explained below along with the resulting breakdown of the phraseological errors in the three subcorpora.

## Breakdown of phraseological errors

I analysed grammatical collocations, idiom-like phrases and phrasal verbs separately in so far as they constitute clearly identifiable entities:

1. **Grammatical collocation errors** correspond to X*PR errors and concern errors on dependent prepositions, i.e. cases where the dependent preposition in N/V/ADJ + preposition combinations is erroneous, e.g. *marriage may not **(XVPR) appeal** $appeal to$ people* (GE); *In my view there is no **(XNPR) justification in** $justification for$ capital punishment* (GE); she is ***(XADJPR) hard to** $hard on$ her son* (GE).
2. **Errors in idiom-like phrases** concern lexically opaque, i.e. non-compositional phrases, where the overall meaning cannot be deduced from the sum of the parts, e.g.

*This book gives you* ***(LP) food for the mind*** *$food for thought$* (FR); ***(LP) to turn over a new leave*** *$to turn over a new leaf$ (GE).*

3. **Phrasal verb errors** exclusively include errors on verb + adverbial particle combinations, e.g. *when I* ***(LP) stand up $get up$*** *at 9.30* (GE); *I hope to be able to* ***(LP) get my point through*** *$get my point across$* (SP); *people marry and* ***(LP) set up*** *$start$ a family* (GE)[1].

However, I grouped compound errors, binomials, similes and lexical collocation errors in the same category referred to broadly as the lexical collocation category. While the task of distinguishing between these types of word combination in the English of native speakers already constitutes a challenge to say the least (Cowie 1998, Howarth 1998, Nesselhauf 2005, Paquot 2007), it becomes even more arduous when dealing with learner errors. For instance, are the following examples instances of lexical collocation, compound or free combination errors?

- *The people who died in the war were* ***(LP) civil people*** *$civilians$* (SP)
- *I decided to make a last attempt* ***(LP) to get my stomache filled*** *$to satisfy my hunger$* (GE)

In their study of collocational errors by advanced EFL learners, Wang and Shaw (2008: 209) also emphasise the problem of distinguishing between errors that affect lexical collocations and free combinations: "when the collocations were produced or misused by the learners, it is very difficult to say which category, namely free ones or restricted ones, they belong to". The following examples illustrate the errors in the lexical collocation category: ***(LP) to run out of hand*** *$to get out of hand$* (FR); ***(LP) tendencies of consumption*** *$consumer habits$* (SP); *I am* ***(LP) the only child of my parents*** *$an only child$* (GE); *they are the* ***(LP) supporters of their families*** *$breadwinners$* (SP); *things like satellites or computers didn't even* ***(LP) come to their minds*** *$exist$ at that time* (GE); ***(LP) daughters and sons*** *$sons and daughters$* (GE).

[1]Errors on prepositional verbs, i.e. verb + dependent preposition combinations, are classified in the grammatical collocation category.

The breakdown of errors in the lexical collocation, grammatical collocation, idiom-like phrases, and phrasal verb categories is presented below for each mother tongue background.

| **Type of phraseological error** | **FR** | **GE** | **SP** | **Total** |
|---|---|---|---|---|
| *Lexical collocation category*<br>Free combinations<br>Lexical collocations<br>Compounds | 97 (44%) | 108 (41,5%) | **213 (44,5%)*** ** | 418 (43,5%) |
| Grammatical collocations | 69 (31%) | 82 (31,5%) | **208 (43,5%)*** ** | 359 (37,5) |
| Idiom-like phrases | 20 (9%) | 21 (8%) | 18 (4%) | 59 (6%) |
| Phrasal verbs | 35 (16%) | 48 (18,5%) | 37 (8%) | 120 (12,5%) |
| **Total** | **221 (100%)** | **259 (100%)** | **476 (100%)*** ** | **956 (100%)** |

Table 2: Breakdown of LP and X*PR phraseological errors

No significant difference was highlighted in the total number of phraseological errors between the FR and GE groups[2]. In fact, Table 2 shows that the phraseological profiles for the FR and GE groups are very similar, with no significant difference in the number of errors across the four phraseological error subcategories. However, a highly significant difference was found between the total number of errors in the SP and the FR and the SP and GE data ($p \leq 0.0001$ each time). The difference between the SP group and its FR and GE counterparts is mainly due to the significantly higher number of errors in the lexical and grammatical collocation categories in the SP data (with $p \leq 0.0001$ for FR and SP and GE and SP). Concerning grammatical collocations, the majority of X*PR errors affect verb + dependent preposition combinations. This applied to the three subcorpora: for the FR group 69,5% of all X*PR errors affect verbs; for the GE and SP groups the proportions reach 62% and 72%, respectively. Examples of XVPR errors include:

- *We use the word "religion" and say "television is the opium of the masses" in order to **(XVPR) refer** $refer to$ society at the end of the 20th century* (SP)
- *They are **(XVPR) dying for** $dying from$ starvation or lack of medicines* (SP)

[2] This study uses the chi-square test to detect statistically significant differences.

The higher number of lexical and grammatical collocation errors in the SP subcorpus is to be related to the lower level of proficiency displayed by the SP sample.

## Grammaticality vs acceptability errors

It is generally agreed that distinguishing between correct and erroneous instances is more straightforward for certain linguistic categories than for others. This is the case for article errors, for instance, (Tomiyana 1980, Ekiert 2004, Díez Bedmar and Papp 2005) as well as certain syntactic and morphological errors (Bardovi-Harlig & Bofman 1989) where errors usually "occur in a patterned, rule-governed way" (Ferris 1999: 6) and are therefore more easily detectable. Lexis, however, is a domain where the distinction between right and wrong becomes much more blurred. Ferris (1999: 6) calls lexical errors "untreatable" errors, i.e. errors for which "there is no handbook or set of rules students can consult to avoid or fix those types of problems". A multi-word unit may indeed be erroneous in one of two ways: it may be (1) formally wrong or (2) formally correct but inappropriately used in context. To make this distinction James (1998) differentiates between grammaticality errors which correspond to formally inexistent multi-word units, and acceptability errors which are formally existing sequences but which are inappropriately used in context. Formally inexistent multi-word units include cases such as *they should* ***(LP) keep close watch*** *$keep a close watch$ on them* (FR), which are near-hits, viz. the combination is a very close approximation of an existing multi-word unit, and cases such as *I would like to* ***(LP) make a vindication of*** *$stress$ the importance of literature* (SP), which bear no resemblance to any existing multi-word unit. On the other hand, acceptability errors include cases such as (in reference to mental hospitals) *such a method could not work with those who are* ***(LP) out of their mind*** *$mentally ill$* (FR), which contain existing but contextually inappropriate combinations.

The British National Corpus (http://corpus.byu.edu/bnc/) as well as a number of native and learner dictionaries and my own native speaker intuition were used to establish the existing vs inexistent nature of the errors in the three subcorpora. The results are presented below:

| | FR | GE | SP | Total |
|---|---|---|---|---|
| Acceptability errors | 125 (56,5%) | 131 (50,5%) | 190 (40%) | 446 (46,5%) |
| Grammaticality errors | 96 (43,5%) | 128 (49,5%) | 286 (60%) | 510 (53,5%) |
| Total | 221 (100%) | 259 (100%) | 476 (100%) | 956 (100%) |

Table 3: Proportion of grammaticality vs acceptability word combinations

No significant difference was found in the number of grammaticality and acceptability errors between the FR and GE subcorpora. The SP subcorpus was found to include significantly more acceptability ($p \leq 0.01$) and grammaticality ($p \leq 0.02$) errors than the FR subcorpus, but no significant difference was highlighted between the GE and the SP data ($p \leq 0.08$ for acceptability and $p \leq 0.1$ for grammaticality errors).

The error proportions within each sample show that the FR subcorpus includes slightly more acceptability than grammaticality errors while the GE group almost displays a 50-50 distribution between the number of grammaticality and acceptability errors; the SP group, on the other hand, clearly committed a much higher proportion of grammaticality than acceptability errors (60% vs 40%). This may again perhaps be related to the fact that the SP population displays a lower level of language proficiency than its FR and GE counterparts. Thus, whereas the higher level of proficiency in GE and SP samples allows these learners to use a high number of existing multi-word units but inappropriately in context, the SP group, because of more limited language command, tends to use more inexistent combinations in the first place.

## Potentially transfer-related errors

The TaLC presentation will give the results of the number of potentially transfer-related errors in the three language groups. The process used here to detect potential transfer errors is back translation, a method advocated, among others, by Granger (2008) as one way of assessing the potential influence of the learners' L1 phrasicon on L2

performance[3]. Word combination errors were translated back into the learners' L1 and in cases where an L1 equivalent to the error could be found, the word combination was categorised as a potential transfer error. An example of this is *everyone has to work hard at school, at university, and later on in **(LP) the active life** $at work$ in order to find a job* (FR) where "the active life" is a direct calque of the French "la vie active", meaning "at work". As will be shown during the TaLC presentation, the results for the more advanced samples, i.e. FR and GE, seem to confirm Kellerman's (1984: 121) claims that the "hoary old chestnut" according to which transfer does not affect the more advanced learner "should finally be squashed underfoot as an unwarranted generalization based on very limited evidence" (for more on the link between transfer and language proficiency see also Kellerman 1977, 1978, 1979, Wode 1976, Zobl 1980).

Our in-depth analysis of phraseological errors will also show that the causes of errors at the more advanced levels are by no means clear-cut but rather, as Granger (2004: 135) puts it, "advanced interlanguage is the result of a very complex interplay of factors: developmental, teaching-induced, and transfer-related". This complex interplay will be illustrated across the three languages.

## Conclusion

Although our analysis only takes two error tags into account and therefore yields a relatively patchy picture of the phraseological errors found in the learner corpus investigated here, it nevertheless goes a long way to showing the value of adopting a computer-aided-error-analysis approach to the study of learner phraseological errors. Our analysis of the LP and X*PR errors in the French, German and Spanish components of ICLE has revealed a number of interesting findings: first, the Spanish subcorpus contains significantly more phraseological errors than its French and German counterparts. This is mainly due to the significantly higher number of errors in the lexical and grammatical collocation categories. Second, in terms of grammaticality and acceptability errors, the

[3] We refer to « potentially » transfer-related errors in so far as, although the error bears trace of possible L1influence, it is not certain that the learner did indeed resort to the L1 when the error was committed.

Spanish subcorpus also stands out as including more grammaticality than acceptability errors. As suggested, both these findings should be related to the Spanish subcorpus displaying a lower level of language proficiency, which leads these learners to make (a) more overall phraseological errors and (b) of a different kind (more grammaticality errors).

## References

**Altenberg B.** and **Granger S.** 2001. The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics* 22 (2), 173-194.

**Bardovi-Harlig K.** and **Bofman T.** 1989. Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11: 17-34.

**Council of Europe.** 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

**Cowie A.P.** (ed.) 1998. *Phraseolgy: Theory, Analysis and Applications*. Oxford: Oxford University Press.

**Dagneaux E., Denness S., Granger S., Meunier F., Neff J.** and **Thewissen J.** 2005. *Error Tagging Manual Version 1.2*. Centre for English Corpus Linguistics. Université Catholique de Louvain, Louvain-la-Neuve. Unpublished manual.

**De Cock S.** 2003. *Recurrent sequences of Words in Native Speaker and Advanced Learner Spoken and Written English.* Unpublished doctoral dissertation. Université catholique de Louvain. Centre for English Corpus Linguistics.

**Díez Bedmar M.B.** and **Papp S.** 2005. The usage of central articles by Spanish and Chinese learners of English at University level. Paper presented at the workshop held in

conjunction with the 4th International Contrastive Linguistics Conference, September 20-23 2005, Santiago de Compostela, Spain.

**Ekiert M.** 2004. Acquisition of the English Article System by Speakers of Polish in EFL and ESL settings. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, 4 (1), 1-23.

**Ellis R.** 2003. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

**Ferris D.** 1999. The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing* 8, 1-10.

**Granger S.** 1998. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In Cowie, A. (ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 145-160.

**Granger S.** 2004. Computer learner corpus research: current status and future prospects. In Connor U. and Upton T. (eds.) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam & Atlanta: Rodopi. 123-145.

**Granger S.** 2008. Some major challenges for theoretical and applied phraseological research. Paper to be presented at the 3rd International Postgraduate Conference on Formulaic Language (FLaRN), Nottingham, 19-20 June 2008.

**Granger S., Dagneaux E.** and **Meunier F.** 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain. Available from http://www.i6doc.com.

**Granger S.** and **Meunier F.** (eds). 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: Benjamins.

**Granger S.** and **Paquot M.** 2008. Disentangling the phraseological web. In Granger S. and F. Meunier (eds) *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia : Benjamins.

**Howarth P.** 1998. The Phraseology of Learners' Academic Writing. In Cowie A.P. (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: OUP, 161-186.

**Hulstijn J.** and **Marchena E.** 1989. Avoidance: grammatical or semantic causes. *Studies in Second Language Acquisition* 11: 242-255.

**James C.** 1998. *Errors in Language Learning and Use*. London and New York: Longman.

**Kellerman E.** 1977. Towards a characterization of the strategy of transfer in second language learning. *Interlanguage Studies Bulletin* 2 (1): 58-145.

**Kellerman E.** 1978. Giving learners a break: native language intuitions as a source of predictions about transferability. *Working Papers on Bilingualism* 15, 59-92.

**Kellerman E**. 1979. Transfer and non-transfer: where are we now? *Studies in Second Language Acquisition*, 37-57.

**Kellerman E**. 1984. The empirical evidence for the influence of the L1 in interlanguage. In Davies A., Criper C. and Howatt A. (eds.) *Interlanguage*. Edinburgh: Edinburgh University Press, 98-122.

**Laufer B.** and **Eliasson S.** 1993. What causes avoidance in L2 learning: L1/L2 difference, L1/L2 similarity, or L2 complexity? *Studies in Second Language Acquisition* 15 (1), 35-48.

**Meunier F.** and **Granger S**. (eds). 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam and Philadelphia: Benjamins.

**Nesselhauf N.** 2003. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics* 24 (2), 223-242.

**Nesselhauf N.** 2005. *Collocations in a Learner Corpus*. Amsterdam: Benjamins.

**Osborne J.** 2008. Phraseology effects as a trigger for errors in L2 English: The case of more advanced learners. In Meunier F. and Granger S. (eds) *Phraseology in Foreign Language Learning and Teaching*, Amsterdam/ Philadelphia: Benjamins, 67–83.

**Paquot M**. 2007. *EAP vocabulary in EFL learner writing: from extraction to analysis: A phraseology-oriented approach.* Unpublished doctoral dissertation. Université catholique de Louvain, Centre for English Corpus Linguistics.

**Scott M.** 2004. *WordSmith Tools 4*. Oxford: Oxford University Press.

**Tomiyana M.** 1980. Grammatical errors and communication breakdown. *TESOL Quarterly* 14, 71-79.

**Wang Y.** and **Shaw P.** 2008. Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME Journal* 32, 201-228.

**Wode H.** 1976. Developmental sequences in naturalistic L2 acquisition. *Working Papers on Bilingualism* 11, 1-13.

**Zobl H.** 1980. The formal and developmental selectivity of L1 influence on L2 acquisition. *Language Learning* 30, 43-57.

## The author

*Jennifer Thewissen is an English language assistant at the Université catholique de Louvain, Belgium. She works at the Centre for English Corpus Linguistics where she is*

*currently working on her PhD project. Her doctoral research focuses on the importance of the construct of linguistic accuracy both in SLA and in the field of language testing. Her research is based on an error-tagged sample of the International Corpus of Learner English.*