Exploring big educational learner corpora for SLA research*

Perspectives on relative clauses

Theodora Alexopoulou[1], Jeroen Geertzen[1], Anna Korhonen[1] and Detmar Meurers[2]
[1] Department of Theoretical and Applied Linguistics, University of Cambridge / [2] Department of Linguistics, University of Tübingen

We consider the opportunities presented by big educational learner corpora for Second Language Acquisition (SLA). In particular, we focus on the *EF Cambridge Open Language Database* (EFCAMDAT), an open access database of student writings submitted to *Englishtown*, the online school of *EF Education First*. EFCAMDAT stands out for its size (33 million words, 85 thousand learners) and a range of 128 writing tasks covering all CEFR levels with data from learners from varying nationalities. We discuss methodological issues arising from analyzing big data resources generated in educational contexts and argue that Natural Language Processing (NLP) is essential for the automated processing of such datasets. As a study case, we follow the developmental trajectory of relative clauses, a construction that necessitates deeper syntactic analysis. We consider specific issues that can affect the developmental trajectory, including task effects, formulaic language and national language effects.

Keywords: big data, educational learner corpus, relative clauses, natural language processing for learner language, formulaic sequences

* We wish to thank Henriëtte Hendriks, Akira Murakami, Maria Kunevich and the audience of EUROSLA 2014 for comments and suggestions. We also thank three anonymous reviewers and the editors of this issue for insightful comments and their meticulous editing work. We are indebted to Rachel Baker from EF for valuable information and commentary and gratefully acknowledge the support of EF as sponsor of the EF Research Unit at the University of Cambridge and the Isaac Newton Trust, Trinity College, Cambridge for supporting building EFCAMDAT.

International Journal of Learner Corpus Research 1:1 (2015), **96–129**. DOI 10.1075/ijlcr.1.1.04ale
ISSN 2215–1478 / E-ISSN 2215–1486 © John Benjamins Publishing Company

1. Introduction

The collection and analysis of learner production data has served as an important empirical sounding board since the early days of Second Language Acquisition (SLA) research. Since Selinker (1972), learner language data has been particularly prominent in developmental research, i.e. the study of the transition from rudimentary second/foreign language (L2) linguistic systems to fully fledged L2 grammars. Longitudinal studies following the productions of individual learners have provided the empirical basis for important insights and a wealth of theoretical analyses and hypotheses (Ortega 2009, Perdue 1993).

With some notable exceptions (Feldweg 1991, Myles & Mitchell 2007), most available longitudinal corpora in the field of SLA consist of the productions of a few individuals gathered over a relatively short period of time. In other cases, important hypotheses are based on just one learner, as in the case of Patty, an L2 learner followed for a period of ten years (Lardiere 1998). Patty's data provided the basis for a new and strong hypothesis according to which syntax can develop despite lack of morphological marking. To confirm such a hypothesis, however, a much broader empirical base is required and it is also important to investigate how a wide range of influencing factors interact to shape morphological and syntactic development (Myles 2008).

At the same time, and rather independently from the SLA mainstream, a growing learner corpus community has built increasingly rich and large learner corpora (Granger 2008). Resources such as the *International Corpus of Learner English* (Granger et al. 2002, Granger et al. 2009) provide large samples of writings by university students. The drawback with such corpora, however, is that they typically lack a longitudinal dimension[1] and mostly consist of one text type, i.e. argumentative essays written by intermediate to advanced students at university level. Collecting rich individual learner data from early to advanced stages of acquisition for a variety of tasks is clearly identified as the next critical step, both in SLA and Learner Corpus Research (LCR) (de Bot et al. 2011, Vyatkina 2012). So far, however, the logistical cost of building such corpora has remained an important obstacle for scaling up empirical resources (Granger 1998, Myles 2008).

Against this backdrop, recent developments in online education open up important opportunities. The emergence of internet teaching platforms for a global audience has resulted in unprecedented amounts of electronically stored learner production data. The *EF Cambridge Open Language Database* (EFCAMDAT)

[1] See *The Longitudinal Database of Learner English* (LONGDALE, Meunier & Littré 2013, http://www.uclouvain.be/en-cecl-longdale.html, accessed 19 November 2014) for a recent project addressing this gap.

is, as far as we are aware, the first freely available large learner corpus exploiting online foreign language learning. EFCAMDAT was built at the University of Cambridge in collaboration with *EF Education First*, an international school teaching English as a foreign language (EFL). The data consists of writings submitted to *Englishtown*, the online school of EF. EF students using *Englishtown* sign up either for exclusively online self-study packages or packages where the online component is blended with (traditional) classroom teaching. The self-study package includes online classroom live lessons where small groups of students have a lesson with a teacher over the internet. Students may also include in their packages one-to-one lessons with a teacher over a video call.[2] The database is freely available to the research community through a web based interface.[3]

The global reach of EF has led to a resource of significant diversity, including 128 activities across sixteen teaching levels aligned to the *Common European Framework of Reference* (CEFR, Council of Europe 2001) as listed in Table 1. The corpus currently contains half a million scripts from 85 thousand learners from 172 nationalities, summing 33 million words, collected over approximately one year. A script is defined as a writing piece a learner submits as an answer to a writing task. An individual learner may have produced multiple scripts. Productions of individual learners can be tracked over time (see Geertzen et al. 2013a for more detail).

Table 1. Correspondence between EF teaching levels and CEFR levels

EF teaching levels	CEFR
1–3	A1
4–6	A2
7–9	B1
10–12	B2
13–15	C1
16	C2

EFCAMDAT displays two typical properties of big data: (i) significantly larger amounts of data than standard resources in the field and (ii) data generated through a real life activity (the teaching operations of an international school), rather than through collection for research purposes. These characteristics raise

[2] In 2010 EF sponsored the EF Research Unit at the Department of Theoretical and Applied Linguistics, University of Cambridge, to promote research in the area of second language learning of English. EFCAMDAT was built with joint funds by EF and the Isaac Newton Trust, Trinity College Cambridge.

[3] Available at http://corpus.mml.cam.ac.uk/efcamdat (accessed 19 November 2014).

some important methodological challenges. The first is the size of the database. Currently, learner corpora are often annotated manually for linguistic information or errors; research hypotheses are also often triggered by observations researchers make through manual inspection of the text files. Manual data processing needs to be complemented by automated data processing if the full size of a resource of half a million scripts is to be exploited. Automated Natural Language Processing (NLP) of big learner data is, therefore, vital (see Granger et al. 2007, Meurers 2009).

The second challenge relates to the nature of the EFCAMDAT data. A number of variables standardly controlled for by research design are not (easily) recoverable in a learner corpus compiled from *Englishtown* activities:

1. *Learners.* Obtaining all scripts produced by all *Englishtown* students in one year provides a rich record, but how do we draw appropriate samples from such a data set? Some students will have been very keen and progressed exceptionally fast in a year, while others may have taken a three-month break from their studies or dropped out altogether after a month.
2. *Tasks.* The writing tasks, topics, and prompts directly affect language production (Lozano & Mendikoetxea 2013, Perdue 1993, Sinclair 2005). The set of 128 topics across sixteen proficiency levels in EFCAMDAT provides a unique window into EFL learners' language development. However, they were developed for teaching, not with a view to creating a representative set of writing tasks for research purposes.
3. *Context.* EFCAMDAT data are produced in an EFL context with a strong Computer Assisted Language Learning (CALL) component, which sometimes affects language production in arbitrary ways (e.g. word limits have led to the predominance of short texts in EFCAMDAT).

All these factors need to be considered before a big learner corpus such as EFCAMDAT can be used to answer key SLA questions such as when a particular structure or feature is acquired. The potential of big learner data to validate or invalidate long standing SLA hypotheses has however already been demonstrated by Murakami (2013) who made use of the *Cambridge Learner Corpus* (CLC 2009) and EFCAMDAT to investigate the developmental paths of English grammatical morphemes. He reported clear native language (L1) effects, thus challenging the hypothesis of a universal L2 acquisition order of English morphemes (Dulay et al. 1982).

In this article, we consider issues specific to tasks, the data context and the learners' native language. Shifting the perspective from morphology to syntax, we focus on the acquisition of relative clauses (RCs). RCs have been extensively studied from a variety of perspectives since the early days of SLA research (Flynn et al. 2004, Schachter 1974, Shirai & Ozeki 2007) but they present a number of challenges for big learner data analysis. First, their identification requires the use of reliable parsing tools. Second, once identified, their interpretation in an EFL context is non-trivial. They may instantiate productive use but could also be examples of formulaic uses (Fillmore 1979, Myles 2012, Wray 2002). This distinction is critical for an accurate empirical documentation of their developmental trajectory. Third, there are no obligatory contexts for RC use and data sparsity may still be a real issue even in a corpus of half a million scripts. Fourth, the lack of RC obligatory contexts makes it hard to establish whether absence of RCs at a given developmental stage reflects absence from the learner grammar or lack of opportunity to produce an RC in a specific task.

In the sections that follow we focus on some challenging characteristics of RCs and explore computational techniques to tease apart factors affecting learner language and model the development of EFL learners' productive use of RCs. In Section 2, we introduce EFCAMDAT and the parser used for syntactic analysis and RC identification. We then focus on a number of methodological challenges specific to big educational corpora like EFCAMDAT. In particular, we investigate the impact of task types on formulaic language use (Section 3.3) and national language effects in RC production patterns (3.4). We also compare the trajectory of RCs with other subordinate clauses (3.5). The discussion remains largely exploratory, focusing on the nature of the methodological challenges and identifying methods for addressing them.

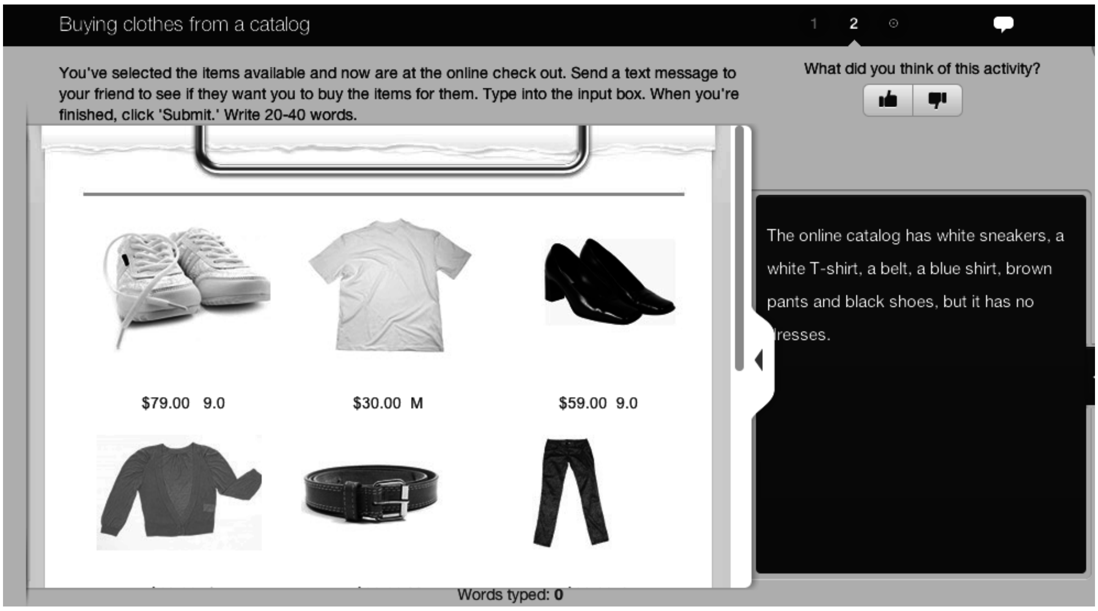
2. Background

2.1 Corpus data

A full course in *Englishtown* spans sixteen proficiency levels, each containing eight lessons with a variety of receptive and productive tasks. Students are allocated to proficiency levels after a placement test when they start a course at EF and are then expected to progress to the following stages. EFCAMDAT consists of all the scripts submitted by EFL learners as answers to the writing tasks at the end of each lesson (see Table 2 for examples of topics). A task prompt may consist of a short text from which learners are asked to extract information; it can also include a model answer. For example, the prompt for task 2.4 shown in Figure 1 contains an image with shopping items, the writing task and a model answer, all of which are available to students when writing their text.

Table 2. Examples of writing topics across teaching levels; level and unit number are separated by colon.

ID	Writing topic	ID	Writing topic
1:1	Introducing yourself by email	7:1	Giving instructions to play a game
1:3	Writing an online profile	8:2	Reviewing a song for a website
2:1	Describing your favourite day	9:7	Writing an apology email
2:6	Telling someone what you're doing	11:1	Writing a movie review
2:8	Describing your family's eating habits	12:1	Turning down an invitation
3:1	Replying to a new penpal	13:4	Giving advice about budgeting
4:1	Writing about what you do	15:1	Covering a news story
6:4	Writing a resume	16:8	Researching a legendary creature



Buying clothes from a catalog

You've selected the items available and now are at the online check out. Send a text message to your friend to see if they want you to buy the items for them. Type into the input box. When you're finished, click 'Submit.' Write 20-40 words.

What did you think of this activity?

The online catalog has white sneakers, a white T-shirt, a belt, a blue shirt, brown pants and black shoes, but it has no resses.

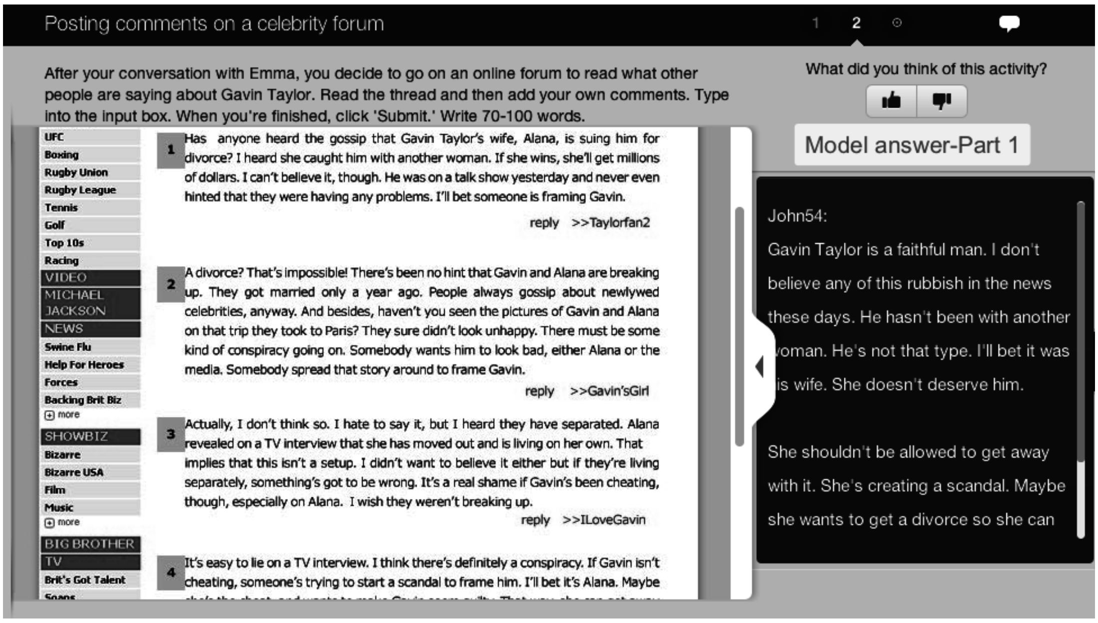
Words typed: 0

Figure 1. Task prompt Level 2, Unit 4.

Figure 2 illustrates another type of task in which learners are asked to contribute a comment to a blog and gossip about celebrities; the input includes three previous contributions to the blog and a rather detailed model answer.[4]

The majority of learners only complete portions of the program. Currently around 4,000 learners have completed a minimum of three teaching levels

[4] The availability of sample answers merits special attention as it is bound to affect learners' lexical choices as well as the discourse structure and organization of their writings. Here, we consider sample answers as part of the input provided by the task prompt but we acknowledge that a more systematic investigation of their role would be desirable.



Posting comments on a celebrity forum

After your conversation with Emma, you decide to go on an online forum to read what other people are saying about Gavin Taylor. Read the thread and then add your own comments. Type into the input box. When you're finished, click 'Submit.' Write 70-100 words.

1 Has anyone heard the gossip that Gavin Taylor's wife, Alana, is suing him for divorce? I heard she caught him with another woman. If she wins, she'll get millions of dollars. I can't believe it, though. He was on a talk show yesterday and never even hinted that they were having any problems. I'll bet someone is framing Gavin.

reply >>Taylorfan2

2 A divorce? That's impossible! There's been no hint that Gavin and Alana are breaking up. They got married only a year ago. People always gossip about newlywed celebrities, anyway. And besides, haven't you seen the pictures of Gavin and Alana on that trip they took to Paris? They sure didn't look unhappy. There must be some kind of conspiracy going on. Somebody wants him to look bad, either Alana or the media. Somebody spread that story around to frame Gavin.

reply >>Gavin'sGirl

3 Actually, I don't think so. I hate to say it, but I heard they have separated. Alana revealed on a TV interview that she has moved out and is living on her own. That implies that this isn't a setup. I didn't want to believe it either but if they're living separately, something's got to be wrong. It's a real shame if Gavin's been cheating, though, especially on Alana. I wish they weren't breaking up.

reply >>ILoveGavin

4 It's easy to lie on a TV interview. I think there's definitely a conspiracy. If Gavin isn't cheating, someone's trying to start a scandal to frame him. I'll bet it's Alana. Maybe

John54:

Gavin Taylor is a faithful man. I don't believe any of this rubbish in the news these days. He hasn't been with another oman. He's not that type. I'll bet it was is wife. She doesn't deserve him.

She shouldn't be allowed to get away with it. She's creating a scandal. Maybe she wants to get a divorce so she can

Figure 2. Task prompt Level 7, Unit 8.

(yielding 24 scripts per learner)[5] and 500 learners have completed every unit from level one to six (yielding 48 scripts per learner). Figure 3 shows average number of sentences per learner across EF teaching levels. Learners produce around 40 sentences on average in lower levels and around 70–80 in more advance levels (13–16). Only learners that have completed all eight units of a given level are included in our calculations.

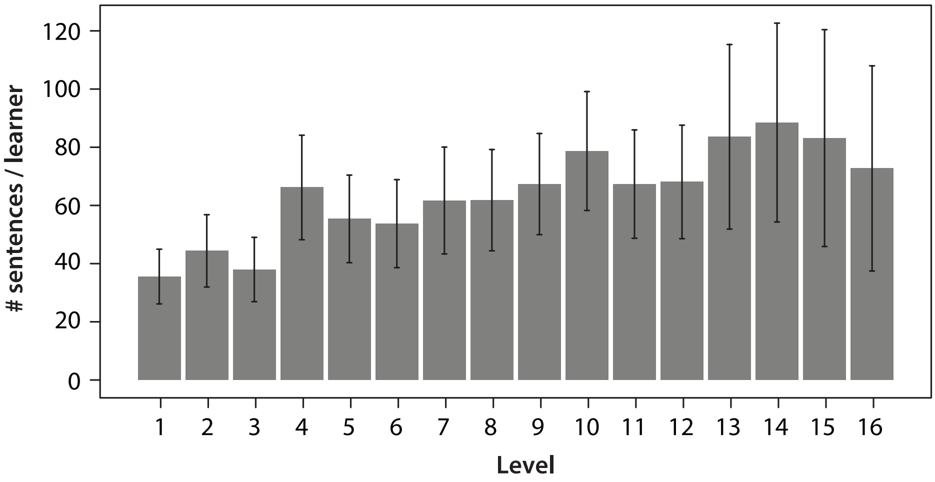


Figure 3. Average number of sentences produced per learner across EF teaching levels

[5] Writings are graded by language teachers, and learners may only proceed to the next level upon receiving a passing grade. If learners fail a writing task, they have to retake it, which means that EFCAMDAT may contain more than the required eight scripts per level for each student.

EFCAMDAT lacks information on the learners' L1 but we approximate L1 through matching nationality and country of residence. This combination provides a reasonable match when the L1 is the dominant or official language of the country. Admittedly, however, it cannot capture variety in L1 backgrounds or multilingualism, and there remains uncertainty on whether some of the sampled learners are indeed speakers of the national language even though they are residents/nationals of the country. Despite these uncontrolled variables, the overall numbers of students from specific countries are big enough to allow strong national language effects to emerge in the data as we will see in Section 3.4. We will use the term 'national language' (NL) to reflect the fact that our data could contain considerable noise with regard to the linguistic background of learners. In this paper, we focus on the language production of learners from the five most frequently represented nationalities in the corpus: Brazilians (36.9%), Chinese (18.7%), Russians (8.5%), Mexicans (7.9%), and Germans (5.6%). To discuss some specific hypotheses regarding particular L1s, we also sometimes include data from Japanese (2.1%) and Italian (4%) EFL learners.

2.2 Annotation and RCs

To automatically identify RCs, we need syntactic annotations. They can be provided by natural language parsers trained on native English which have been shown to perform with high accuracy on EFCAMDAT (Geertzen et al. 2013b). Here we use the C&C Combinatory Categorial Grammar (CCG) parser (Clark & Curran 2007), a rich grammar formalism storing much of the grammatical information in the lexicon (Steedman 2000). The C&C CCG parser performs very well in object and subject extraction from RCs (Rimell et al. 2009) and all 2,369,994 sentences produced by the learners from the five most frequent nationalities were parsed using its standard model and settings.

For our analysis, we targeted the relative pronouns *who* and *which* and the complementiser *that*. Relative pronouns in CCG correspond to two lexical categories defining a relative pronoun roughly as an element combining with a sentence 'missing' a noun phrase (NP), filled by the relative pronoun as a subject or object.[6] Unfortunately, we had to exclude zero RCs (*the girl I saw*) because they are particularly challenging for the parser. In addition, we excluded free relatives introduced by *whoever/whatever* (*whoever comes to the party*). We have, however, included RCs introduced by the interrogative pronoun *what*, to capture non-target patterns illustrated in examples like (1) (see Section 3.1).

[6] See Hockenmaier & Steedman (2007) for further details.

(1) a. The job what I have found for you
b. Something what we want.

3. The developmental trajectory of RCs

One fundamental goal of SLA research is to establish if and when a certain structure is acquired (Bley-Vroman 1989). Empirical documentation of the developmental trajectories of individual structures or phenomena is a prerequisite to this task but it is by no means straightforward. As discussed in Section 3.1, it is not always clear what counts as an RC in learner language given the relatively frequent production of non-target and erroneous patterns. Section 3.2 focuses on the parser performance to extract RCs. Section 3.3 considers the impact that formulaic language and task effects can have on empirical generalisations and Section 3.4 shows that developmental trajectories can differ according to the learners' NL. Finally, Section 3.5 compares the acquisition path of RCs with that of other structures and discusses this particular trajectory in the general context of L2 grammar acquisition.

3.1 What counts as an RC in learner language?

To trace RC development empirically, we first need to document all RC occurrences in learner writings across proficiency levels. As shown in (2), learner language includes many non-target-like RCs. Examples (2a) and (2b) are headless RCs. In (2c) *what* is used instead of *that* or *which*. In (2d) two items, *that* and *what*, are used instead of one. Finally, in (2e) an inanimate head, *the game*, is followed by the animate *who*.

(2) a. In the end who has the most points is the winner.
b. Wins who scores the most points.
c. This job is really the most suitable job what I have found for you.
d. If you want to know opinion that what you need.
e. This game who is called bowling alley takes place on an area of about 8 meters by 3 meters.

Large corpora allow us to establish which non-target-like patterns are systematic and which ones are restricted to a limited number of occurrences or just a few individuals. In the long term, sophisticated data-driven NLP techniques should help obtain regular patterns that may go unnoticed or cannot be predicted theoretically. To be successful, however, these techniques would necessitate richer annotations of learner data, for instance including semantic features to capture cases like (2e). Currently lacking such annotations, we evaluate how well a parser can identify learner RCs in the following section.

3.2 Evaluation of RC retrieval

To evaluate how well the C&C CCG parser identifies RCs, we annotated manually sampled scripts and compared the parser performance against our manual annotations. We sampled scripts primarily from elementary and intermediate levels (levels 4–7) as learners start producing more RCs at these levels (see Section 3.3). We also hypothesized that a large proportion of non-target-like patterns would appear at elementary and intermediate levels, thus making the identification task the most challenging for the parser. We sampled productions from three NL groups, i.e. Brazilian, Russian and Japanese EFL learners[7] and selected thirty random scripts for each NL from the ten tasks listed in Table 3. This resulted in a total of 900 scripts (ten tasks * thirty scripts * three NLs) summing 5,919 sentences.

The first author manually annotated the test data and tagged RCs introduced by *that*, *who*, *which* and non-target-like *what* (2c). In addition to non-target patterns, zero RCs were also systematically recorded to appreciate their frequency in learner language. A total of 204 sentences were found to contain a RC, representing 3.5% of the test set. These 204 sentences included 31 zero relatives (15%). No instance of non-target like structures introduced by both *that* and *what* was identified. By contrast, the remaining types of non-target patterns illustrated in (2) were all found in the sample scripts. Precision, recall and F1 measures were calculated for all RCs except zero relatives. Precision reflects parser success in correctly identifying RCs; a low precision score indicates frequent misanalysis of non-RC structures. Recall expresses parser success in finding all RCs in the data; a low recall score indicates that the parser misses many RCs. Finally, the F1-score is the harmonic mean of precision and recall. The extraction of the 173 RCs (excluding zero-relatives) shows good performance, with 80.7% precision, 87.3% recall, and an F1-score of 83.9%. Approximate randomization tests did not reveal a NL or proficiency (lower half vs upper half) effect on the measures.

The parser missed eighteen RCs and misanalysed twenty strings as RCs. False RCs fall in the following categories:[8]

[7] RCs have been argued to be particularly challenging for Japanese learners (Schachter 1974, Flynn et al. 2004).

[8] There were also two unclassified cases:

(i) a. Men can't wear jeans, baggy jeans, t-shirts and shorts.
b. Woment: can't wear miniskirts, dresses and top.

Table 3. Sampled tasks for manual annotation

Teaching Level & Unit	Task Title	Task Prompt
2.4*	Buying clothes from a catalogue	You've selected the items available and now are at the online check out. Send a text message to your friend to see if they want you to buy the items for them. Write 20–40 words.
3.8	Choosing a birthday present	You've received an email from your classmate about your teacher's birthday. Answer the email. …. Write 20–40 words.
4.4	Writing about what you like doing	Write an email reply to your friend and talk about the activities suggested. Say whether you can do them, like or don't like them and whether or not you wish to do them for your birthday. Write 50–70 words.
4.7	Complaining about chores	Look at the list of chores. As you see, Julia didn't do very much this week. In fact, YOU did most of her chores! Write her an email telling her all the chores you did. Write 50–70 words.
5.7	Writing a sick note	You and your friend, Mike, missed a good friend's wedding. Mike lets you see the email he sent to explain his illness. Now write your own email to your friends. Write 50–70 words.
5.8	Giving cultural tips to a visitor	Read the attached text about Canadian culture. Write an email to your friend who is coming to visit about what they should and shouldn't do in Canada. Start the letter with 'Dear friend'. Write 50–70 words.
6.3	Creating an office dress code	You receive another email from the manager. Read the instructions and write a new dress code for the office. Write 50–70 words.
6.6	Writing an email of advice	You are an online counselor. Write an email to Polaris. Give her a plan to help her fight her shopping addiction. Write 50–70 words.
6.7	Complaining about a meal	You just ate a very bad meal at a restaurant. Write a complaint in the complaints book. Follow the guidelines below. Write 50–70 words.
11.2	Helping a coworker deal with a phobia	You are Ian's friend and colleague. Read the leaflet and write an informal email to Ian encouraging him to keep his current job. Include a description of claustrophobia symptoms and some advice on how to cope with the phobia. Write 80–150 words.

*2.4 is a task from teaching level 2, Unit 4.

- Direct questions as RCs (six instances): *We can surfing and go swimming, **what do you think?***
- Indirect questions as RCs (four instances): *Tell me **what do you think**.*
- Demonstrative *that* as complementiser (four instances): *You shouldn't yell down the street to a friend, **that is viewed as inappropriate**.*

– Declarative *that* as RC complementiser (four instances): *I am writing this current email to try convincing you* ***that it is not necessary you leave your job…*.**

Misanalysis is often due to the ambiguity of the subordinator (*that*, *what*). To resolve such ambiguity, the parser would need to be more sensitive to the distinction between complements and modifiers.

The parser also missed eighteen RCs which involve a mixed set of cases with no evident pattern as evidenced in (3) to (6).

(3) Here are some tips about my country that you should know.

(4) Before I forget, here in Canada there are few things that not allowed.

(5) Mr. Smith, I will send to you the dress code that you've asked me.

(6) Another thing that we can suggest its cotton shirts for both women and men, it's natural and stylish.

In conclusion, existing NLP tools such as the C&C CCG parser perform reasonably well for the extraction of RCs from learner data, and, in particular from EFCAMDAT. A current shortcoming however concerns the identification of zero RCs, which represent 15% of RC production and have been shown to be an indicator of proficiency (Wulff et al. 2014, Tizón-Couto 2013).

3.3 Formulaic sequences and task effects in an EFL context

The occurrence of a particular structure in learner data can be interpreted as evidence for (the process of) its acquisition.[9] In varying forms, this assumption has underpinned SLA developmental studies which have focused on when particular linguistic forms are first used, and how often they are used across developmental stages (Bardovi-Harlig 2000, Ellis 2010, Wulff et al. 2009). Accordingly, Figure 4 shows the percentage of sentences including a RC out of the total number of sentences produced by EFL learners representing five different nationalities across twelve proficiency levels, and Figure 5 plots the percentage of learners who produced at least a single RC at each proficiency level. The two figures show that there are very few RCs before Level 4 when RCs increase until Level 6 after which they stabilise.[10] Further, there are no NL effects regarding the timing of RCs. From all

[9] Here we abstract away from the question of accuracy as a crucial component of operational definitions of acquisition, e.g. in the order of acquisition morpheme studies (Dulay et al. 1982). See Schachter (1974) for a discussion about the complex interaction between accuracy and rate of production of RCs.

[10] Except for the peak at level 11.

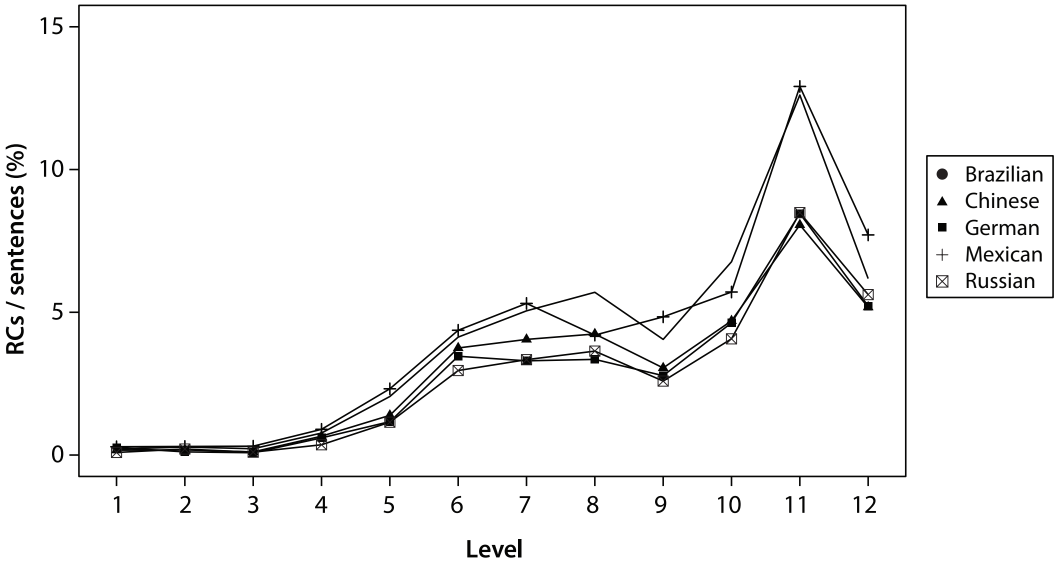


Figure 4. Percentage of RCs in sentences produced across twelve EF teaching levels for five nationalities.

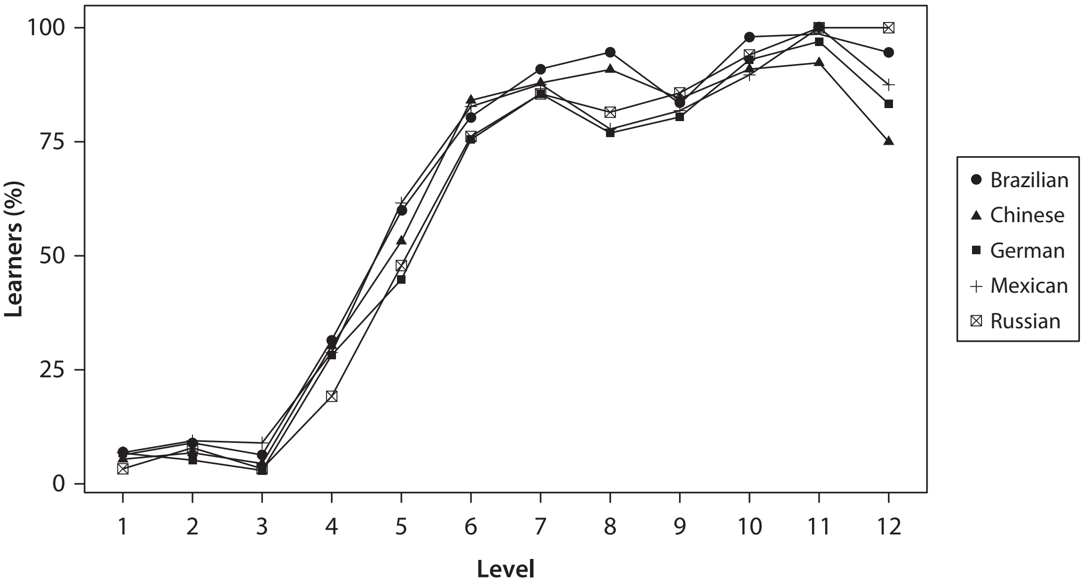


Figure 5. Percentage of learners producing at least a single RC across twelve EF teaching levels for five nationalities.

this, we could conclude that RCs emerge at Levels 4–6 with a good part of their acquisition (in terms of use) in place by Level 7.

The quantitative results in Figures 4 and 5, however, need to be interpreted with caution. Since the early days of SLA, it has been known that learners use formulaic sequences (FSs) to meet communicative needs (Fillmore 1979). A formulaic sequence is "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002: 465). FSs also play a crucial role in acquisition, not just as communicative 'crutches', but supporting the development of productive grammar (Fillmore 1979, Myles 2012).

RC occurrence therefore cannot be interpreted as direct evidence for the existence of grammatical knowledge, and productive uses of RCs must be distinguished from FSs involving those structures. It will also be important to establish whether production increases at levels 4–6 as a result of language development or whether this increase is induced by the specific writing tasks available at those levels. In an EFL context, task effects and FSs interact since FSs can be part of routines rehearsed in a lesson that learners are then asked to reproduce in writing.[11]

3.3.1 *Formulaic sequences in EFCAMDAT*

Examples (7a) and (7b) contain an RC and are from Level 1. The RC in (7a) is headed by a partitive (*those of you*) and includes a contracted negation agreeing in number with the partitive head. This contrasts with the remaining text which is simpler and also features 'basic' errors (*more* is wrongly preceded by the definite article, and *interesting* is used instead of *interested*). In (7b), on the other hand, the RC is headed by a simple indefinite noun phrase.

(7) a. **For those of you that don't know me**. My name is Liji Yuan. Here is an interesting fact. Do you know that the more companies are interesting in this product?

b. I had to married **a awful man that I don't love for some reasons**.

Let us assume for the sake of the argument that the RC in (7a) is an FS present in the writing task prompt. Compared to the simpler RC in (7b), its advanced complexity and accuracy shows how important it is to identify FSs and distinguish them from co-existing productive RCs.

3.3.2 *Task effects*

In EFCAMDAT the distinction between formulaic and productive use is further complicated by the very context of language production. Figure 6 shows the percentage of sentences containing a RC out of the total number of sentences produced for each writing task at the end of each lesson (thus providing eight data points per teaching level). The percentages of elicited RCs vary from task to task.

At the lowest levels (1–3), the production rate seems driven by one lesson; at the higher levels (6–8), there is more variation. A number of possible task effects can be considered to explain such a difference as described in the following scenarios:

[11] See Paquot (2013) on this issue regarding lexical bundles.

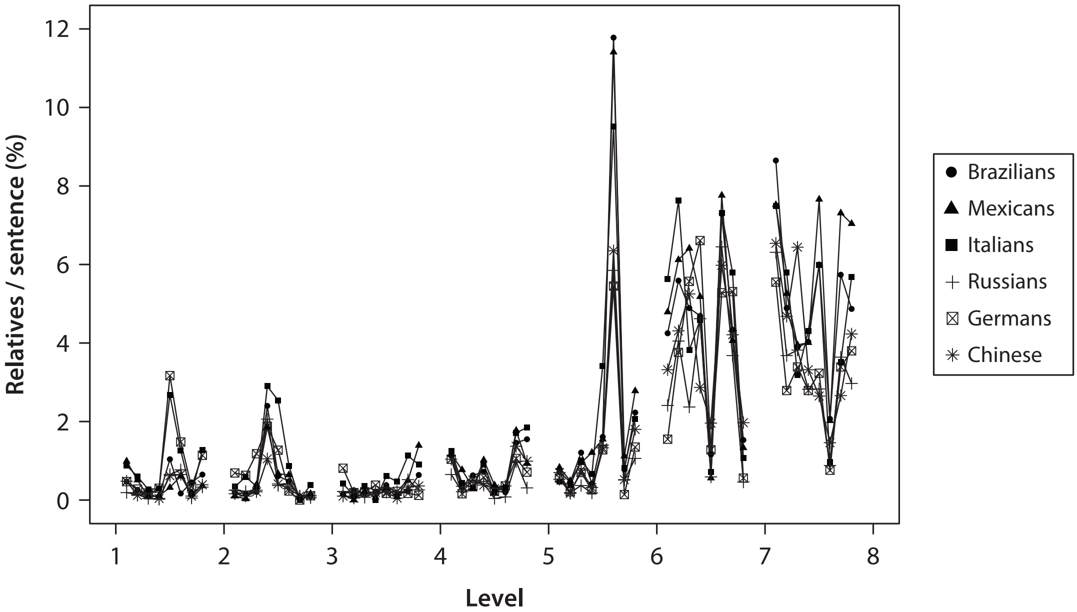


Figure 6. Percentage of RCs in sentences produced for each lesson across twelve EF teaching levels.

1. *Explicit encouragement to include a specific FS containing a RC.* Classroom teaching often involves rote-learning of FSs which learners are then encouraged to use. Learners may thus be encouraged to use FSs containing a RC in a specific writing task, leading to a higher rate of RCs in that particular lesson or task.
2. *Explicit encouragement to include a certain type of structure (e.g. RC).* A prompt may explicitly encourage students to use a particular type of expression (e.g. gerund, temporal clause).
3. *A task elicits RCs implicitly.* A prompt may elicit a high number of occurrences of a particular structure (e.g. temporal clauses, pronouns) as a natural consequence of the language required to meet the discourse/communicative requirements of the task.
4. *A task is neutral with regard to the elicitation of a specific structure.*
5. *Copying from input prompt.* Students often copy material from the writing prompt. Such copying ought to be random, not correlating with the nature of specific tasks, but it is noise worth identifying.

These five scenarios, by no means exhaustive, present some real challenges when it comes to analysing a big learner corpus such as EFCAMDAT.[12] It is impossible to predict which writing prompt will trigger any of these scenarios and we need automatic techniques to identify task-induced uses of RCs. In Section 3.3.3, we focus on scenario 1 and discuss the identification of task-induced FSs in more detail.

[12] There is a long literature documenting task effects in SLA, see for instance Tavakoli & Foster (2008).

3.3.3 *Identifying task-based formulaic sequences*

Wray's definition of a formulaic sequence captures the essence of what a FS is, but cannot help separate what is prefabricated from what is productive language use. One approach pursued by corpus linguists and recently applied to learner corpora (O'Donnell et al. 2013) is to define FSs through a range of parameters such as frequency of use, internal coherence and variability.

Building on O'Donnell et al. (2013), we operationalized FSs as repeated sequences of word *n*-grams and ranked them according to the following indicators: (a) frequency of *n*-grams in sentences containing RCs (b) number of learners who produced them, (c) length and (d) Mutual Information (MI) score (i.e. a statistical measure that captures word association strength, see Church & Hanks 1990). We ranked *n*-grams contained in sentences with a RC for each individual task of Levels 1–12 (96 writing tasks) in EFCAMDAT. Here, we restricted the investigation to the scripts submitted by one arbitrarily chosen NL group, Brazilian learners. We selected *n*-grams long enough to contain at least part of an RC and used by at least 10% of all learners. For each writing task, Table 4 reports the *n*-gram with the highest score according to a combined measure of MI with Sharedness (S, i.e. the percentage of learners who used the n-gram).[13] The highest scoring *n*-gram *for those of you that do not know me my name is* occurred 140 times (F = 140) in scripts written as an answer to the writing task of Lesson 5, Level 1, and was used by 51% of the learners (S = 51). It has a high MI score indicating the very low likelihood that those words would have appeared together by chance. The third column (RCs) of Table 4 gives the total number of RCs produced in a given lesson.[14]

To evaluate if the adopted procedure missed potential FSs we manually inspected the scripts used for the evaluation of RC extraction (Table 3). We found two candidate FSs: *few things you should know about Canada* in Task 5.8 and *clothes that are too short* in Task 6.3. The first *n*-gram instantiates a zero RC, an RC type excluded from our analysis. We inspected S*MI scores for Task 6.3, shown in Table 5. As we can see the candidate FSs picked by the manual inspection are indeed ranked high in terms of their S*MI score. Recall that in Table 4, we included only the highest scoring *n*-gram in each task, which is why these *n*-grams were excluded. The shorter *n*-gram *clothes that are too short* has a lower MI score than the longer *clothes that are too short too tight too baggy*, a higher scoring *n*-gram, which is, nevertheless produced by a smaller number of learners.

We also inspected RCs extracted from tasks 2.4, 4.7, 5.6 and 5.8. These tasks have higher RC production rates (see Figure 6) and we were interested to see

[13] We report the highest scoring *n*-gram for 5.8 although it was only used by 9% of learners.

[14] The *n*-gram for 4.7 *let me tell you what I did* is an indirect question, misanalysed by the parser as an RC.

Table 4. Top ranked *n*-grams in sentences with RCs produced by Brazilian learners*

Lesson	n-gram	RCs	F	RCs:F	S	MI	S*MI
1.5	For those of you that do not know me my name is	370	140	2.6	51	54	2730
4.7	Let me tell you what I did	237	114	2	52	26	1363
5.8	There are things that we should remember	160	14	11.5	9	39	371
6.3	Anyone who does not follow the dress code will lose their job	259	31	8.3	14	65	926
6.4	A job that allows me to use my	197	73	2.9	37	34	128
6.6	Here is a plan that might work for you	300	45	6.6	19	44	810
7.1	For each pin that is knocked down	228	554	4.11	30	29	866
8.3	Things that I would like to do	244	27	10	13	28	347
8.6	Will be assigned an instructor who will	210	8	26.5	12	33	411
10.3	That I will have to pay off the loan	263	28	9.3	13	53	698
11.7	Allows the company to refuse to pay me for something that is not	91	10	9.1	16	64	998
12.3	The sand painting that you	104	21	4.95	28	17	473

* The RCs column shows the total number of RCs, Frequency (F) shows the frequency of the *n*-gram, Sharedness (S) shows the % of learners who have produced the *n*-gram and Mutual Information (MI) shows the MI score for the *n*-gram.

whether this could be due to a higher rate of missed FSs. For example, Table 6 shows that the repeated sequence *story that began* in Task 5.6 is partially contained in two highly ranked *n*-grams. The third *n*-gram has a low MI and S*MI score but was used by a quarter of the learners.

To summarise, the manual evaluation showed that the S*MI scores can identify potential FSs but additional inspection is necessary for cases with lower MI scores which are nevertheless used by many learners. Importantly, all additional *n*-grams picked by the manual inspection are ranked high based on their S*MI score.

Searching for the source of our candidate FSs we found that eight out of the twelve *n*-grams listed in Table 4 were included either in the task prompt or the model answer; we checked this for tasks 4.7, 6.3, 6.4, 6.6, 7.1, 10.3, 11.7 and 12.3

Table 5. Top ranked *n*-grams in sentences with RCs produced by Brazilian learners for task 6.3.

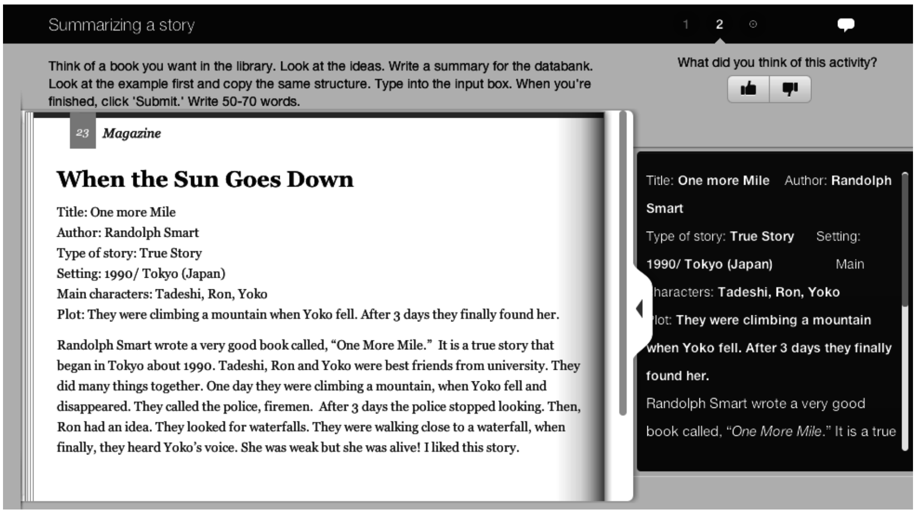
Level	n-gram	F	S	MI	S*MI
6.3	Anyone who does not follow the dress code will lose their job	31	14	65	926
6.3	Clothes that are too short too tight too baggy	39	18	38	685
6.3	Clothes that are too short	65	30	19	562

Table 6. Top ranked *n*-grams in sentences with RCs produced by Brazilian learners for task 5.6.

Level	n-gram	F	S	MI	S*MI
5.6	It is a	155	32	9	298
	Story that	87	17.5	15	262
	That began in	117	24.5	10	247

and the additional FSs manually identified for tasks 5.6 and 6.3. The task prompts for the latter are shown in Figures 7 and 8.

Reconsidering Figure 6, some peaks in the lower levels occur in tasks for which FSs were identified (1.5, 4.7, 5.8). It is however not easy to determine how



Summarizing a story

Think of a book you want in the library. Look at the ideas. Write a summary for the databank. Look at the example first and copy the same structure. Type into the input box. When you're finished, click 'Submit.' Write 50-70 words.

23 *Magazine*

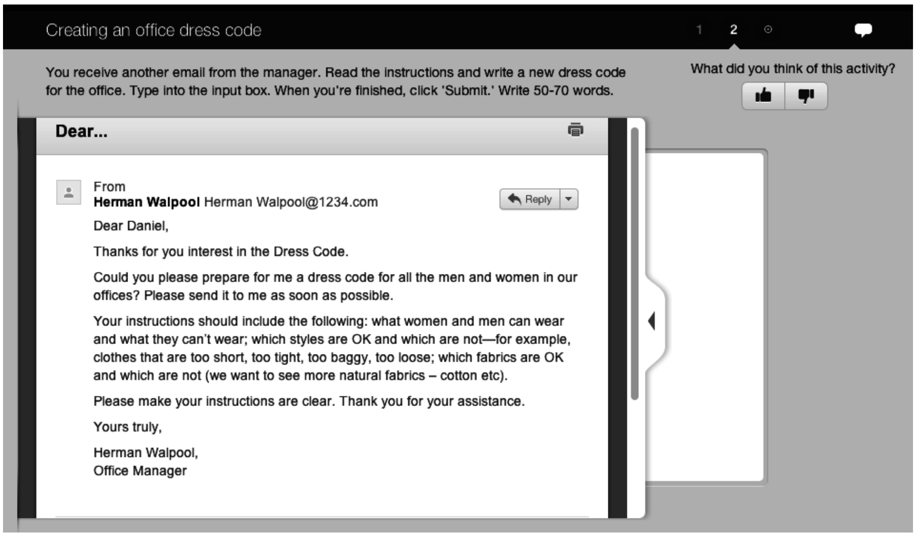
When the Sun Goes Down

Title: One more Mile
Author: Randolph Smart
Type of story: True Story
Setting: 1990/ Tokyo (Japan)
Main characters: Tadeshi, Ron, Yoko
Plot: They were climbing a mountain when Yoko fell. After 3 days they finally found her.

Randolph Smart wrote a very good book called, "One More Mile." It is a true story that began in Tokyo about 1990. Tadeshi, Ron and Yoko were best friends from university. They did many things together. One day they were climbing a mountain, when Yoko fell and disappeared. They called the police, firemen. After 3 days the police stopped looking. Then, Ron had an idea. They looked for waterfalls. They were walking close to a waterfall, when finally, they heard Yoko's voice. She was weak but she was alive! I liked this story.

Title: **One more Mile** Author: **Randolph Smart** Type of story: **True Story** Setting: **1990/ Tokyo (Japan)** Main haracters: **Tadeshi, Ron, Yoko** lot: **They were climbing a mountain when Yoko fell. After 3 days they finally found her.** Randolph Smart wrote a very good book called, "*One More Mile.*" It is a true

Figure 7. Task prompt Level 5, Unit 6.



Creating an office dress code

You receive another email from the manager. Read the instructions and write a new dress code for the office. Type into the input box. When you're finished, click 'Submit.' Write 50-70 words.

Dear...

From
Herman Walpool Herman Walpool@1234.com

Dear Daniel,

Thanks for you interest in the Dress Code.

Could you please prepare for me a dress code for all the men and women in our offices? Please send it to me as soon as possible.

Your instructions should include the following: what women and men can wear and what they can't wear; which styles are OK and which are not—for example, clothes that are too short, too tight, too baggy, too loose; which fabrics are OK and which are not (we want to see more natural fabrics – cotton etc).

Please make your instructions are clear. Thank you for your assistance.

Yours truly,

Herman Walpool,
Office Manager

Figure 8. Task prompt Level 6, Unit 3.

much of each peak is due to the use of a task-induced FS. A rough estimation can be obtained by comparing the number of (RC-containing) *n*-grams with the number of all RCs in a task, as reflected in the ratio RCs:F provided in Table 4. This ratio tends to grow larger as proficiency increases, indicating that the task-based *n*-grams/FSs progressively account for fewer instances of RC production. On the other hand, there are also peaks for tasks where no FSs were identified (see Figure 6). We manually inspected scripts from a number of lessons with such peaks. As shown in Figure 7, the prompt in Task 5.6, for example, asks for a summary of a book, a task that appears to have naturally elicited many RCs like the ones in (8) (RCs are given in bold).

(8) a. After many days came to **a land we know today as the Americas** brought with him an old row from **gold plates that had salvaged** from **a distant relative who said** that were descendants of Joseph of Egypt.
b. The book said about **the strategic military who can be** in real life.
c. The Olimpians is **a book that tell a story** about semi-Gods.

Task 5.6 thus seems to instantiate scenario 3, i.e. a task eliciting a high number of RCs because of its communicative/discourse requirements. We also identified a FS in task 5.6, *story that began in*, which is included in the task prompt. The use of this FS exemplifies scenario 1, where the FS is explicit in the input, therefore, priming the learners to use it. Interestingly, what data from task 5.6 show is that scenarios 1 and 3 may co-exist. There was no instance of scenario 2, i.e. a prompt explicitly encouraging RC use (note that RCs are not explicitly taught before Level 9).

In summary, it has been shown that tasks can have a strong effect to elicit certain language structures and FSs. However, it is also the case that if we were to subtract FSs from Figures 4 and 5, the overall picture would not change dramatically. RCs would be even fewer at beginner levels while production peaks would still appear at higher levels (e.g. task 5.6 or task 7.1).

3.3.4 *Formulaic sequences and productivity*

In the previous section we explored the usefulness of the S*MI scores for the identification of task-induced FSs, an important step in separating formulaic from productive language use. However, the fact that learners used a FS does not necessarily mean that they could not analyse or produce it on the basis of their own linguistic knowledge. After all, native speakers use FSs when their grammatical knowledge would allow them to produce more complex language. A FS is also a gradient concept in SLA. FSs develop over time as learners break them down over the course of acquisition, a process that often results in less accurate or fluent language and sometimes produces the impression of regression. For this reason, a second approach to identifying FSs pays attention to the relation of candidate

FSs to the rest of a learner's production, often employing the following qualitative criteria (Myles 2012):

1. General length and complexity: longer and more complex than the rest of a learner's production
2. Greater phonological coherence: fluent and non-hesitant
3. Inappropriate use and overgeneralisation
4. Non-substitutability
5. Grammatical accuracy
6. Comparison with productive grammar.

As EFCAMDAT is a parsed written corpus, it is possible to evaluate these characteristics except for 2. For example, we can analyse the FS *for those of you that do not know me* and check whether (i) partitive constructions are to be found outside of FSs and (ii) there are other instances of complex RC heads. Such analyses are likely to reveal a discrepancy between this RC and other learner productions. A further element is the amount of variation in the *n*-gram. Example (9) shows signs of analysis: a simplified head consisting of just the demonstrative rather than the full partitive, *who* replacing *that* and an agreement error (*those — doesn't*).

(9) Good evening Ladies and Gentlement, for those who doesn't know me, may name's Olga S.

At more advanced stages of acquisition, however, it is more difficult to identify FSs on the basis of complexity. Consider *let me tell you what I did* in task 4.7. This FS is contained in the model answer provided to learners but many RCs and indirect questions of similar complexity and structure are also found in learner language (examples 10a–d).

(10) a. I am sending you all that I did last week.
b. See what I had to do.
c. Check what I did beyond my own obligations.
d. Please look at the list of chores which I did this week.

In addition, we find sentences with the formula *let me +verb* (examples 11a-d).

(11) a. Let me remind you of one Russian saying.
b. Let me list to you what I did last week.
c. Let me show you what you should have done.
d. Let me remember you what I did.

What this case illustrates is that it is still not straightforward how to interpret this use in relation to the learner's productive language knowledge, even when we can locate an FS in a writing prompt and reasonably assume that the reason it was produced by around half of the learners in this case is probably because it was part of the input. This is because the distinction between formulaic and productive use is not a dichotomous one.[15] FSs can also function as templates which learners use to scaffold their production. Examples (11b–d) can be viewed as variants of *let me tell you what I did* but one could also argue that they instantiate the more open template *let me*+ verb +*you*+ *what*-indirect question.

Another example is *for each pin that is knocked down* (task 7.1). This *n*-gram occurred 554 times in a task which produced 2282 RCs (Table 4). At first glance, there is nothing particularly complex about this RC. However, if we compare it with other RCs from the same task we find that many involve non-target patterns of headless RCs introduced by *who* as in (12a), (13a) and (13b). Such headless RCs could be argued to be less complex than the headed RC of the prompt. In addition, these RCs involve systematic errors (word order in 12a, 12c, 13a, 13b; lack of agreement in 12b, 12d; absence of auxiliary for passive form in 13c). Thus, while learners use RCs productively, the RC contained in the *n*-gram may well be beyond their productive abilities. This can be further investigated by quantifying the frequency of RCs with a quantifier head (e.g. 16c, 16d), passivisation, overall accuracy etc. to estimate how likely it is that a learner's grammar could produce the RC.

(12) a. Wins who have the most points.
b. The team that get the flag of the other team and bring it to his field wins the game.
c. Wins the team that keep the major number of people in the end of this game.
d. The player that make most points is the winner.

(13) a. Win who make the most points.
b. Win the game who is the more fast and more intelligent.
c. Give a Frisbee to each player who allowed to take two shots on each turns.
d. and put ten bottles that can be used as the bowling pins.

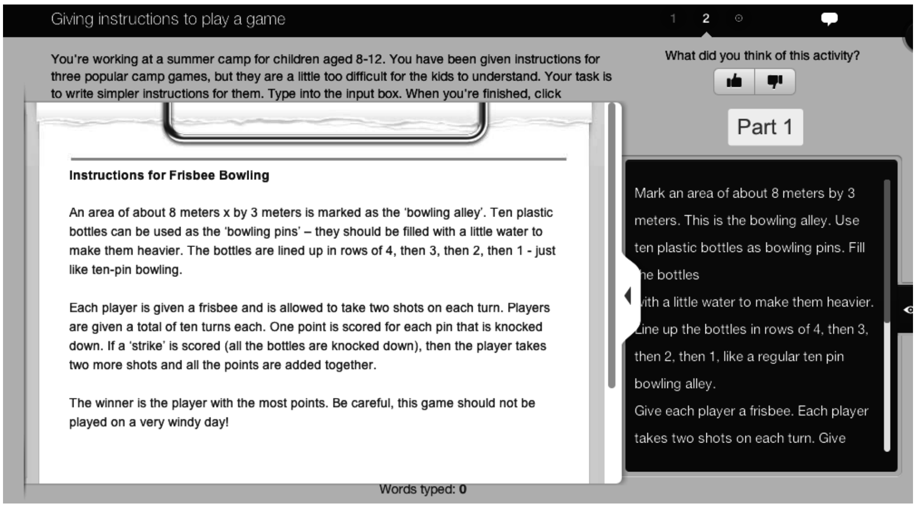
In Task 7.1, students have to simplify a text, a task with both predictable and constrained content (Figures 9 and 10). The FS *for each pin that is knocked down* is contained in both the original text and the model answer. Looking at the ways in which learners reproduced this FS sheds some light on how well learners comprehend and manipulate it. Learners for instance may change the original structure through preposing the RC (*for each pin that is knocked down, one point is scored*) or by simplifying the non RC part (*score one point for each pin that is knocked down*). Analysis of attachment similarity shows that only 25% of the Brazilian

[15] We thank an anonymous reviewer for this point.

learners who used the *n*-gram kept intact the structure in the input. So, 75% of those who used it 'extracted' it from its original structure and embedded it in a (slightly) different structure.

Looking at how learners who did not use the FS expressed the required meaning can also be revealing. Examples (14a–f) are some alternatives used; the phrasings corresponding to the FS are in italics.

(14) a. Temporal clause: score one point *when you knock down one pin*.
b. Conditional: Score one point for each pin *if it is knocked down.*
c. Non-finite sequence: *score a point: knock down each pin.*
d. Non-tense sequence: *knock down a pin, score one point.*
e. Prepositional phrase: score one point *for knocking down a pin.*
f. Infinitive + purpose infinitive: *knock down a pin to score a point.*


Giving instructions to play a game
1
2
What did you think of this activity?
Part 1
Words typed: 0

You're working at a summer camp for children aged 8-12. You have been given instructions for three popular camp games, but they are a little too difficult for the kids to understand. Your task is to write simpler instructions for them. Type into the input box. When you're finished, click

Instructions for Frisbee Bowling

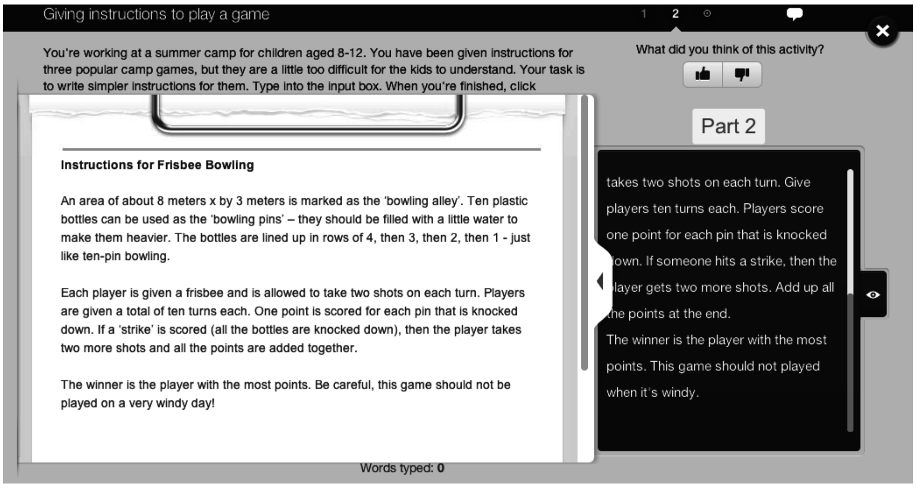
An area of about 8 meters x by 3 meters is marked as the 'bowling alley'. Ten plastic bottles can be used as the 'bowling pins' – they should be filled with a little water to make them heavier. The bottles are lined up in rows of 4, then 3, then 2, then 1 - just like ten-pin bowling.

Each player is given a frisbee and is allowed to take two shots on each turn. Players are given a total of ten turns each. One point is scored for each pin that is knocked down. If a 'strike' is scored (all the bottles are knocked down), then the player takes two more shots and all the points are added together.

The winner is the player with the most points. Be careful, this game should not be played on a very windy day!

Mark an area of about 8 meters by 3 meters. This is the bowling alley. Use ten plastic bottles as bowling pins. Fill he bottles
ith a little water to make them heavier. Line up the bottles in rows of 4, then 3, then 2, then 1, like a regular ten pin bowling alley.
Give each player a frisbee. Each player takes two shots on each turn. Give

Figure 9. Task prompt Level 7, Unit 1, page 1 of model answer.


Giving instructions to play a game
1
2
What did you think of this activity?
Part 2
Words typed: 0

You're working at a summer camp for children aged 8-12. You have been given instructions for three popular camp games, but they are a little too difficult for the kids to understand. Your task is to write simpler instructions for them. Type into the input box. When you're finished, click

Instructions for Frisbee Bowling

An area of about 8 meters x by 3 meters is marked as the 'bowling alley'. Ten plastic bottles can be used as the 'bowling pins' – they should be filled with a little water to make them heavier. The bottles are lined up in rows of 4, then 3, then 2, then 1 - just like ten-pin bowling.

Each player is given a frisbee and is allowed to take two shots on each turn. Players are given a total of ten turns each. One point is scored for each pin that is knocked down. If a 'strike' is scored (all the bottles are knocked down), then the player takes two more shots and all the points are added together.

The winner is the player with the most points. Be careful, this game should not be played on a very windy day!

takes two shots on each turn. Give players ten turns each. Players score one point for each pin that is knocked own. If someone hits a strike, then the layer gets two more shots. Add up all he points at the end.
The winner is the player with the most points. This game should not played when it's windy.

Figure 10. Task prompt Level 7, Unit 1, page 2 of model answer.

Some phrasings indicate reanalysis of the RC into alternative clauses (conditionals, temporal) while others show simpler language (e.g. 14d, 14e).

3.4 The developmental trajectory of RCs and national language effects

The interplay between FSs and task effects discussed above already sheds some light on the relation between input and language development in an EFL context. Another major factor shaping L2 acquisition is L1 effects — the very fact that the acquisition of the second language is preceded by the acquisition of the first language. Resources like EFCAMDAT allow us to study the written production of learners from diverse backgrounds in reasonably similar conditions — at least the 'immediate' input and overall proficiency level are largely matched. The limitation, however, is that we can only rely on NL rather than L1 information.

In Section 3.3, no NL effects were observed for the timing of RCs. However, Schachter (1974) showed that the learners' L1 can affect production rates. Comparing RC production rates of Persian, Arab, Chinese and Japanese learners of English, she argued that learners from L1s with RCs structurally different from English tend to avoid producing RCs in comparison with learners from L1s with RCs structurally similar to English. Thus, Japanese learners whose L1 contains prenominal modifiers that are not marked by a distinct relativiser face the biggest challenge and, thus, produce the fewest RCs. They are followed by Chinese learners. Chinese marks RCs with a specific element *da* which, according to Schachter, is comparable to English *that*, but places RCs prenominally. In Persian and Arabic, RCs are postnominal and introduced by a complementiser-like element as in English. Thus, Persian and Arab learners have less difficulty in comparison to Chinese and Japanese learners and, thus, produce RCs at a rate matching native controls. As for the NLs under study here, Schachter would accordingly have predicted the following order of difficulty for the acquisition of English RCs depending on L1:

Japanese > Chinese > Russians, Germans, Romance (Brazilians, Italians, Mexicans)

English RCs are easiest for Russians, Germans and speakers of Romance languages (Brazilian, Mexican and Italian): RCs are postnominal in these languages and introduced either by a pronoun (Russian, German) or a *that*-style complementiser (Romance). They would be harder for Chinese learners and even more so for Japanese learners.[16]

[16] In addition to the top five nationalities considered so far, we add here Japanese EFL learners since they are central to Schachter's predictions.

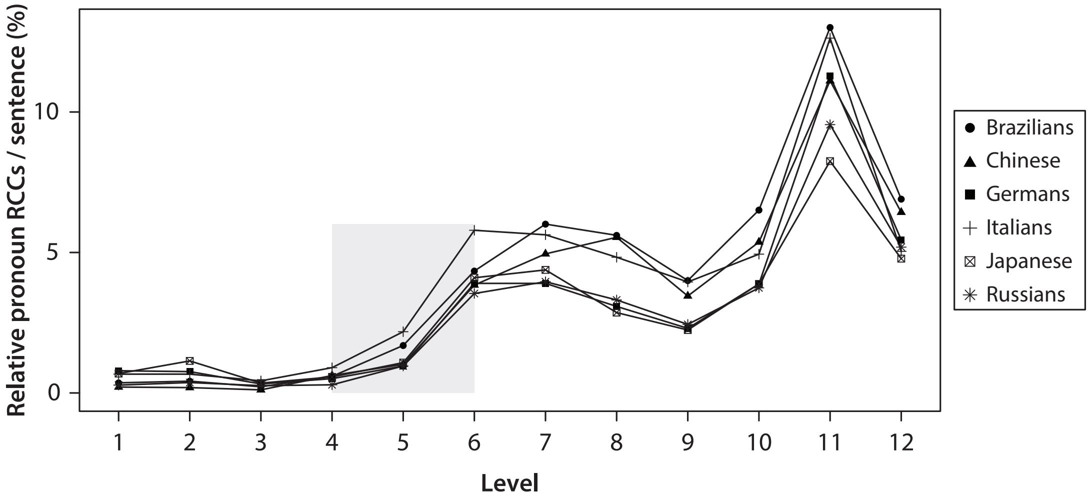


Figure 11. Percentage of RCs in sentences produced in EF teaching levels 1–12 for six nationalities.

Schachter's predicted order is however not confirmed by our data. A one-way between subjects ANOVA was conducted at each level to test whether RC production rates differ across nationalities. There was a significant effect at all levels except at Level 12. Generally, production rates for Brazilian and Mexican learners tend to be higher, but post-hoc analysis using Tukey's HSD revealed only a statistically significant difference between those two nationalities with any other at Level 5 (with $F(4) = 17.58$, $p < 0.001$) and Level 7 (with $F(4) = 14.56$, $p < 0.001$). Brazilians also showed a significantly higher production rate as compared to Chinese, German, and Russian learners at Levels 8, 10, and 11.

Some clear NL effects nevertheless emerge when the types of RCs used by different learners are investigated. Figure 12 shows percentages of object RCs (examples 15a-d) introduced by different types of relativisers: *that* (15a), *who* (15b), *which* (15c) and the ungrammatical *what* (15d).

(15) a. Overall it is a good travelling experience that we will remember for a long time.
b. In the picture you can see very nice couple who I met in hotel.
c. Absolutely the worst mean which I ate in all my life.
d. The e-ticket is a receipt what you paid your ticket.

The ungrammatical *what* RCs represent a small percentage across all NLs. But a pattern emerges: the Mexicans, Italians and Brazilians have a strong preference for *that*-RCs. By contrast, Russians, Germans and Chinese produce as many *that*- as *wh*-RCs. The same tendencies are observed for subject RCs introduced by *who* (16a), *which* (16b) and *that* (16c).[17] Speakers of Romance languages (Mexican

[17] As expected, learners produce many more subject *wh*-RCs than object *wh*-RCs.

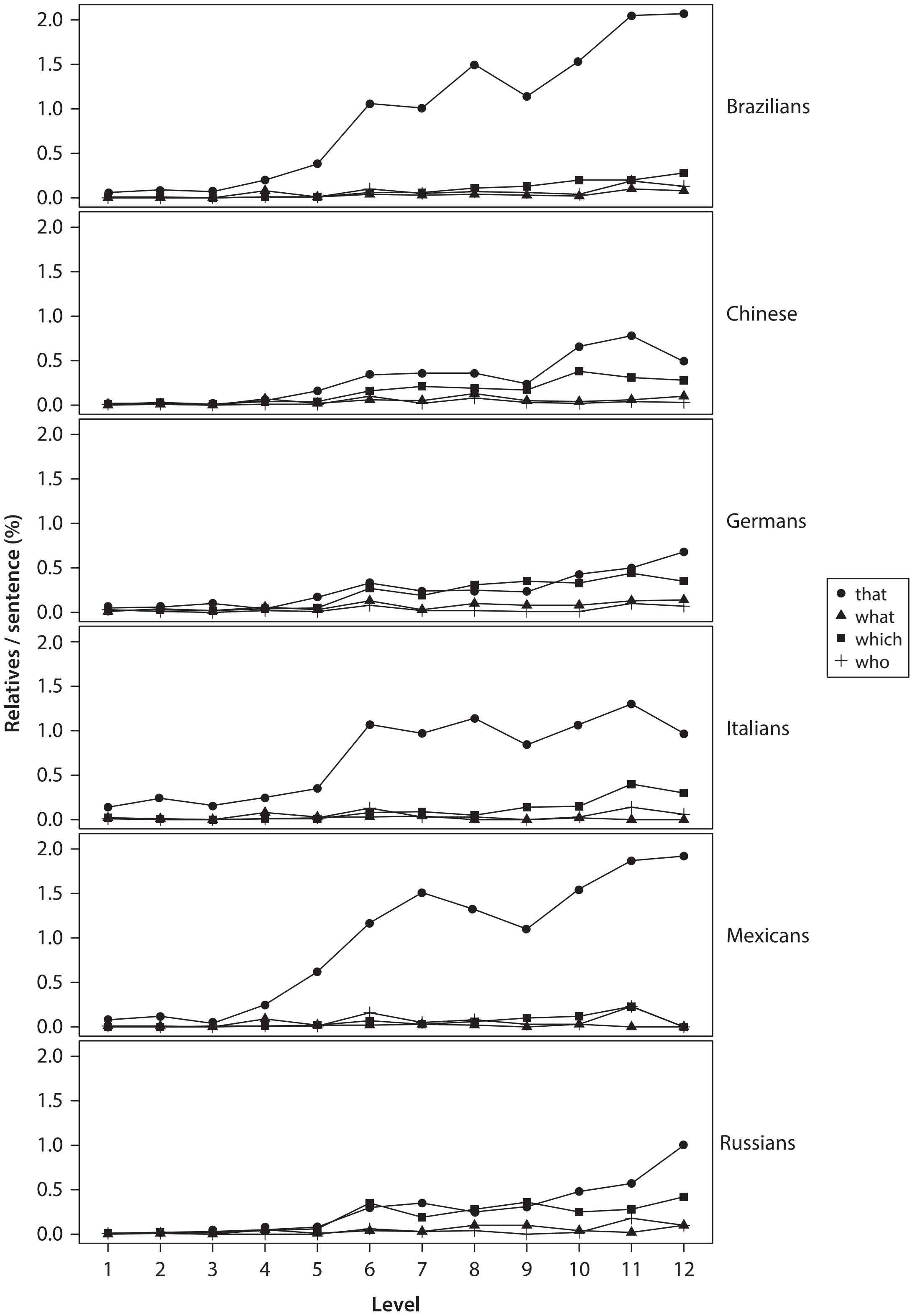


Figure 12. Object RCs split by type of relativiser (*who*, *which*, *that*, *what*) across EF teaching levels 1–12 for six nationalities.

Spanish, Italian and Brazilian Portuguese) prefer *that*-RCs, absent from the productions of Russians, Germans and Chinese.

(16) a. On the left is my friend Paul who he works with me and very beautiful.
b. I work as an engineer in a company which works with thermal insulation.
c. John Woodward is a singer-songwriter that brought your feelings on this music.

We counted the number of occurrences of relativisers (*who*, *which*, *that*, *what*) in the first fourteen levels. Due to the strongly unbalanced sample sizes, we conducted a fixed effects Poisson regression (see Agresti 2002) with nationality, relativiser, and RC type (subject/object) as independent variables, followed by Tukey HSD pairwise comparisons of least-square means to assess interactions between nationality and relativisers. We used the GLM procedure and the *lsmeans* package in the statistical software *R* (Team 2008) and checked for lack of overdispersion and a satisfying deviance goodness-of-fit. Pairwise comparisons between levels of nationality and relativiser confirm the preference for *that*-RCs ($p < .0001$) for Brazilians, Mexicans, and Italians compared to other NLs.

Manual data inspection suggested that Russian, German and Chinese learners avoid *that* when the RC head is animate ('a person that lies all the time'). To confirm this hypothesis, we used a rule-based approach that relies on the lexical database *WordNet* (Miller 1995) to automatically determine noun phrase animacy with an accuracy rate of about 90% (Orasan & Evans 2007).

Figure 14 shows the production of different types of RCs with an animate head noun. Russian, German and Chinese learners use *who* to introduce all the RCs headed by an animate noun. By contrast, speakers of Romance languages use both *who* and *that* for such RCs. A comparison with 'inanimate' RCs shows that for Romance learners the complementiser *that* is by far the preferred option, while Germans, Chinese and Russians use *which* as well as *that* (Figure 15).

To test whether there is an interaction between nationality and relativiser in predicting animacy of the head noun, a logistic regression model was fit with animacy of the head noun (animate/inanimate) as a dependent variable and with nationality, relativiser, and their interaction as independent variables. We targeted Level 7, where RC production is relatively high. To test whether the interaction is significant, a second model was fit without the interaction terms, and a likelihood ratio test of the model with interaction against the model without interaction showed a difference ($\chi 2$ (12) = 230.1, $p < .0001$). Tukey HSD pairwise comparisons of least-square means revealed that Brazilians use *who* less often with animate heads than Russians ($p < .0001$), Germans ($p = 0.0013$), or Chinese ($p = 0.0004$). The same trend was found for Italians.

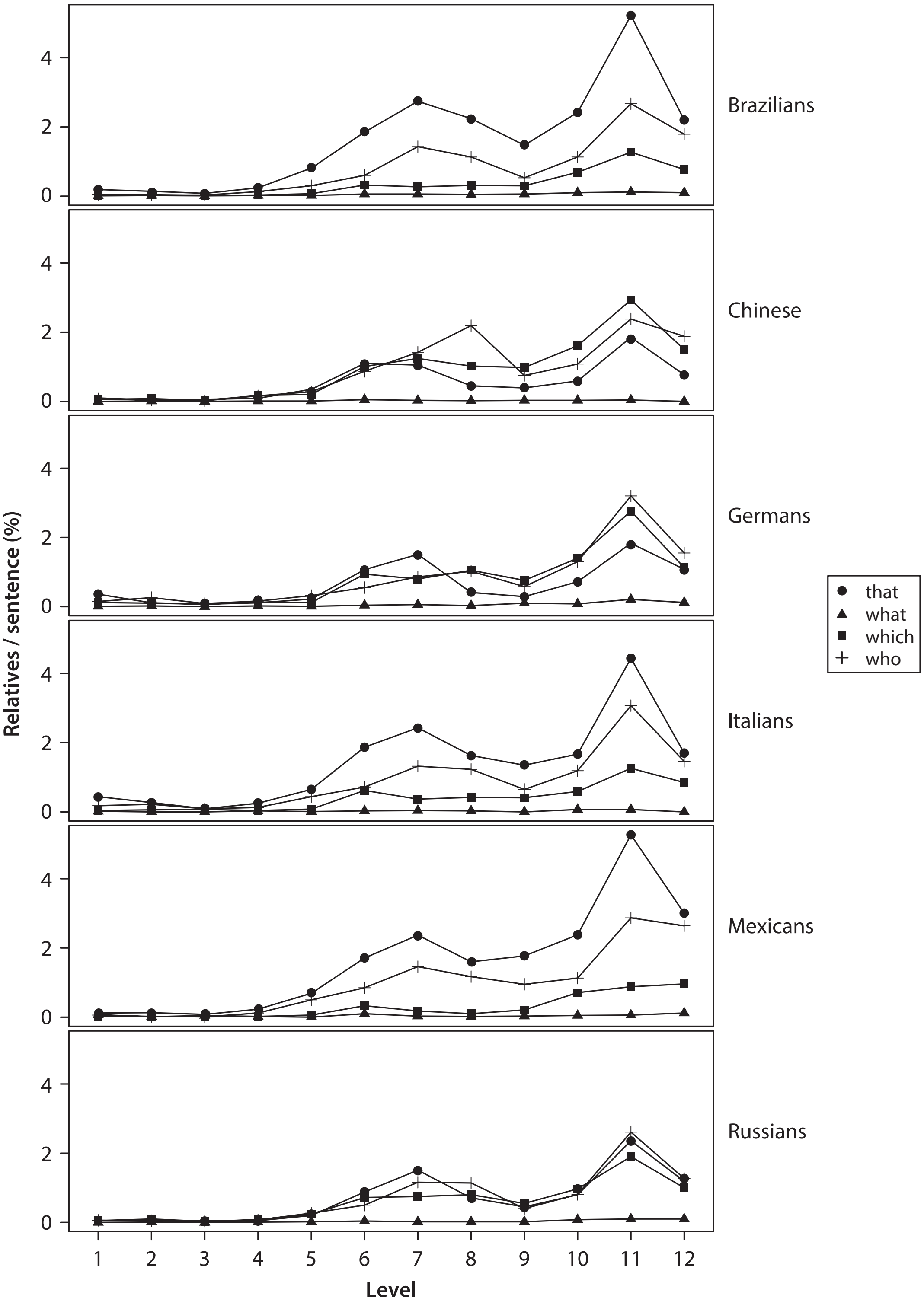


Figure 13. Subject RCs split by the type of relativiser across EF teaching levels 1–12 for six nationalities.

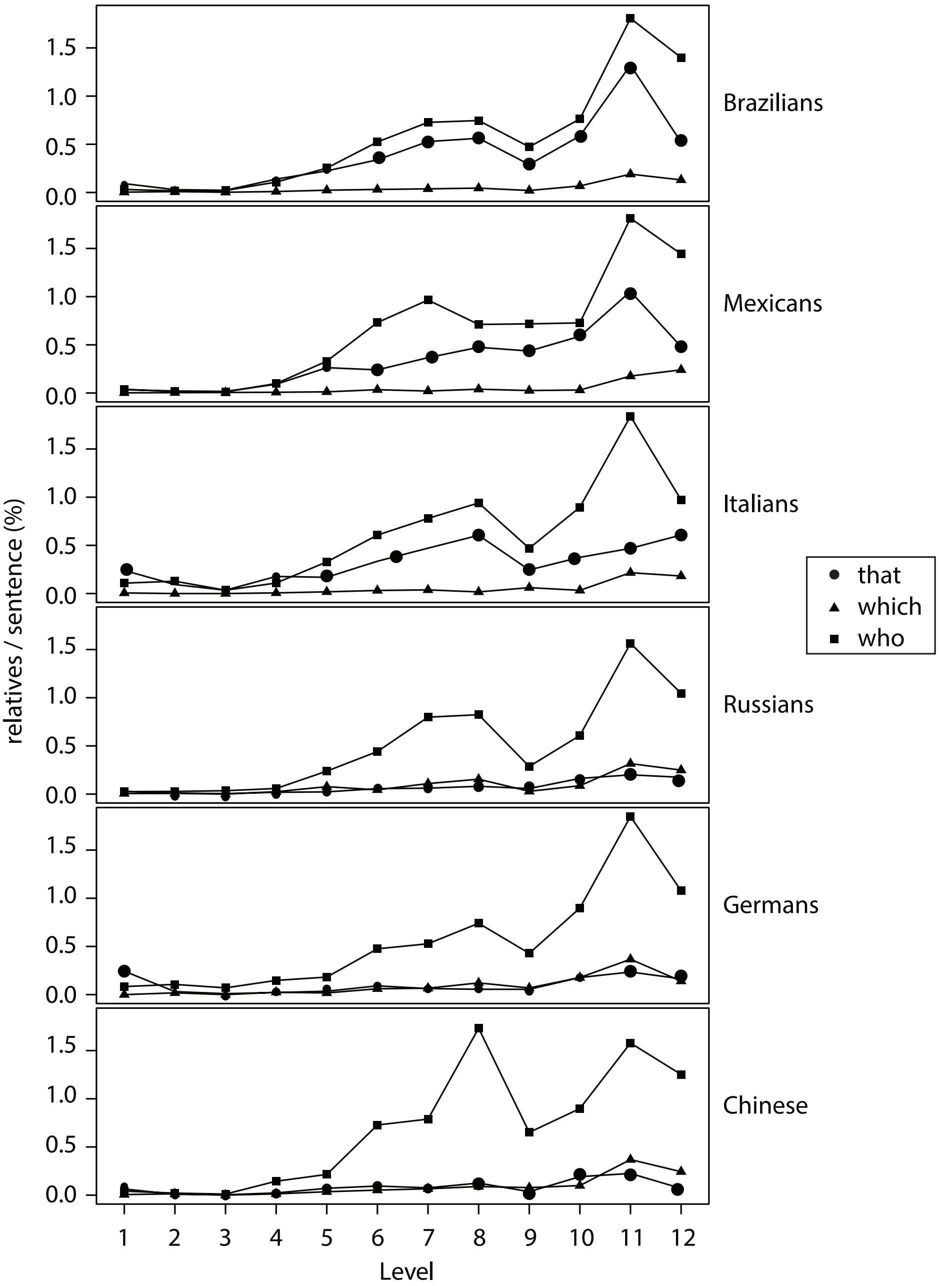


Figure 14. RCs with an animate head noun, split by relativiser type (*who*, *which*, *that*, *what*) across EF teaching levels for six nationalities.

In sum, despite the lack of information on L1, we can study NL effects and identify some new patterns regarding the types of RCs different learners produce.

3.5 Beyond RCs

The question of whether the acquisition of RCs is intrinsically linked to other structural phenomena or proceeds independently is important for theoretical debates

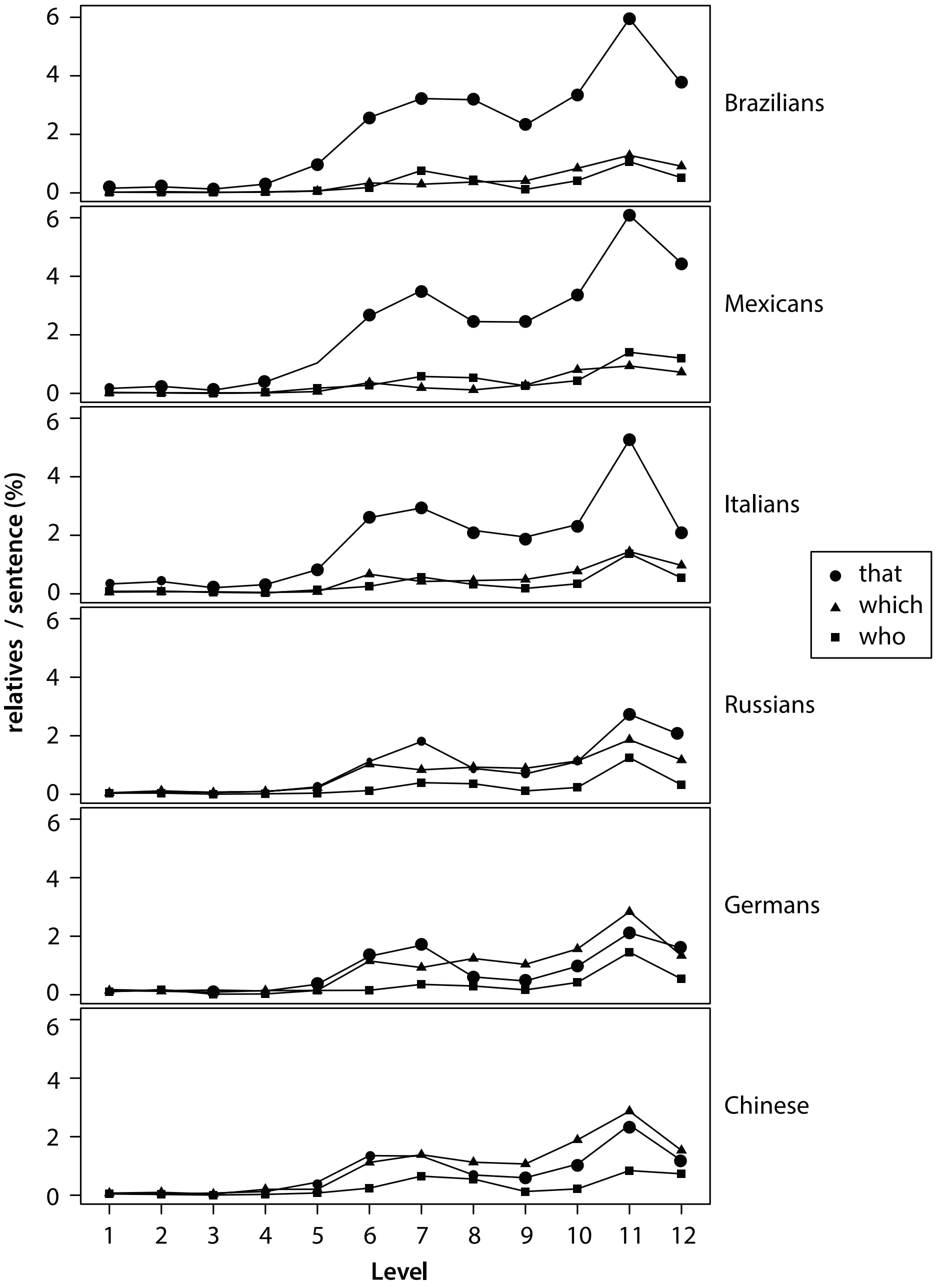


Figure 15. RCs with an inanimate head, split by type of relativiser (*who*, *which*, *that*, *what*) across EF teaching levels 1–12 for six nationalities.

on whether acquisition is item/construction driven or whether deeper structural properties are acquired together (Robinson & Ellis 2008, White 1989). Consider, for example, the trajectory of subordinate clauses in Figure 16. Interestingly, the numbers of subordinate clauses also increase at Level 4, at the same time as RCs.

A more targeted comparison between RCs and *that*-declaratives reveals a parallel developmental path as shown in Figure 17. Declaratives are produced at

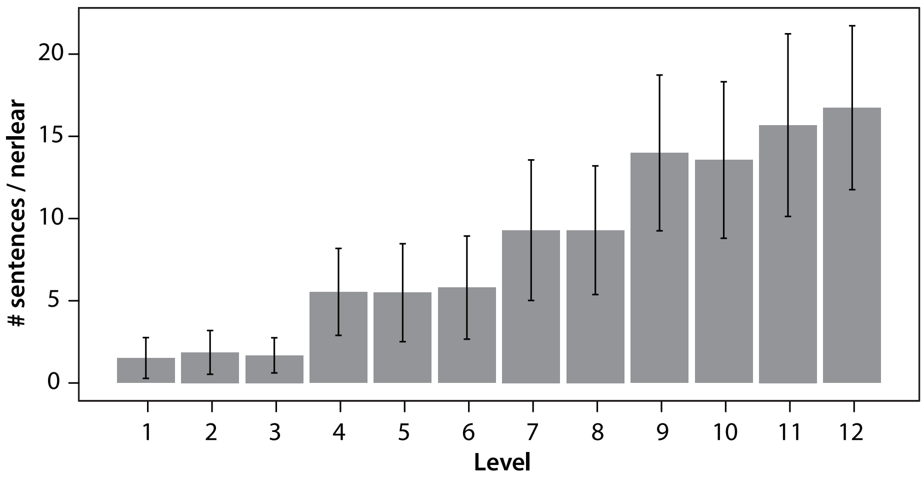


Figure 16. Sentences with subordinate clauses across EF teaching levels.

higher rates than RCs and continue their upward trend well after the intermediate levels, but the two structures emerge together at Levels 4–7.

Such comparisons can reveal important correlations between learners' use of different structures. Of course, all the issues raised earlier regarding the effect of input and tasks would also need to be carefully factored in. It is however worth highlighting two points. First, the presence of a linguistic item in the input or its priming in some other ways does not guarantee that learners will use it. Indeed, some of the most challenging features of L2 acquisition like the definite article are abundant in the input and taught explicitly but article omission persists even at advanced levels (DeKeyser 2005). The second point is that by establishing ordering patterns and correlations between different structures, we should be able to make comparisons with other learner corpora representing language learning in foreign language contexts or naturalistic environments, or L1 acquisition.

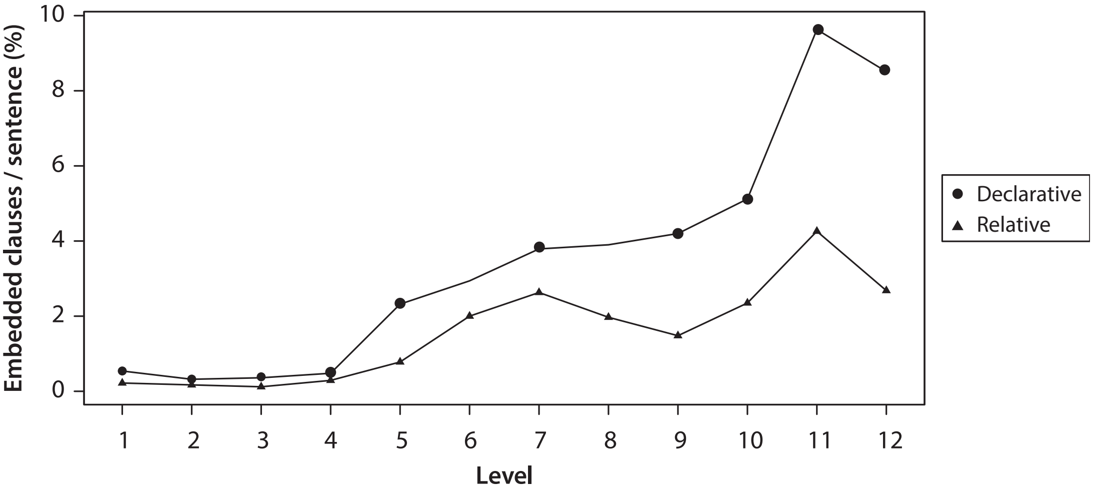


Figure 17. Percentage of sentences with *that*-RCs and *that*-declaratives across EF teaching levels 1–12.

4. Conclusion and open issues

A key empirical task of SLA research is to document the developmental trajectories of individual structures. Big data resources like EFCAMDAT can provide rich datasets across a range of proficiency levels to support this empirical task. However, a number of methodological challenges need to be tackled beforehand: evaluating task effects, identifying FSs, identifying non-target-like patterns of production. Using RCs as a study case, we showed that existing NLP tools can already enable us to make significant progress in tackling these issues. The case study presented here has also shown that, despite its relatively lack of metadata, EFCAMDAT can be used to produce results worthy of consideration for SLA theory and practice. For example, findings indicate that RCs emerge at Levels 4–6, which correspond to the CEFR A2 Level. In addition, they do not confirm Schachter's (1974) predictions about acquisition order and L1 transfer.

A number of issues remain open, however, in particular the identification and modelling of non-target-like patterns of production. Another challenge relates to the need to capture longitudinal aspects of FSs as they become part of the learner's inventory across different tasks and have been shown to be central to the process of acquisition. As mentioned in Section 2.1, EFCAMDAT contains longitudinal data for individual learners. Although the focus of this study has been on cross-sectional analyses, combining the cross-sectional perspective with an analysis of individual learner variation is a necessary next step.

References

Agresti, A. 2002. *An Introduction to Categorical Data Analysis 2*. New York: John Wiley & Sons. DOI: 10.1002/0471249688

Bardovi-Harlig, K. 2000. *Tense and Aspect in Second Language Acquisition: Form, Meaning and Use*. Oxford: Blackwell.

Bley-Vroman, R. 1989. "What is the logical problem of foreign language learning?". In S. M. Gass and J. Schachter (Eds.), *Linguistic Perspectives on Second Language Acquisition*. New York: Cambridge University Press, 41–68. DOI: 10.1017/CBO9781139524544.005

Cambridge Learner Corpus. 2009. Cambridge ESOL and Cambridge University Press. Available at http://cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/.

Church, K. W. & Hanks, P. 1990. "Word association norms, mutual information, and lexicography", *Computational Linguistics* 16(1), 22–29.

Clark, S. & Curran, J. R. 2007. "Wide-coverage efficient statistical parsing with CCG and log-linear models", *Computational Linguistics* 33(4), 493–552. DOI: 10.1162/coli.2007.33.4.493

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

de Bot, K., Lowie, W. & Verspoor, M. H. (Eds.). 2011. *A Dynamic Approach to Second Language Development. Methods and Techniques*. Amsterdam: John Benjamins. DOI: 10.1075/lllt.29.website

DeKeyser, R. M. 2005. "What makes learning second language grammar difficult? A review of issues", *Language Learning* 55, S1, 1–25. DOI: 10.1111/j.0023-8333.2005.00294.x

Dulay, H., Burt, M. & Krashen, S. 1982. *Language Two*. New York: Oxford University Press.

Ellis, N. C. 2010. "Construction learning as category learning". In M. Pütz & L. Sicola (Eds.), *Cognitive Processing and Second Language Acquisition: Inside the Learner's Mind*. John Benjamins, 27–48. DOI: 10.1075/celcr.13.05ell

Feldweg, H. 1991. *The European Science Foundation Second Language Database*. Nijmegen: Max Planck Institute for Psycholinguistics.

Fillmore, L. W. 1979. "Individual differences in second language acquisition". In C. Fillmore, D. Kempler & W. S.-Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior*. New York: Academic Press, 203–228. DOI: 10.1016/B978-0-12-255950-1.50017-2

Flynn, S., Foley, C. & Vinnitskaya, I. 2004. "The cumulative enhancement model for language acquisition: comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses", *The International Journal of Multilingualism* 1(1), 3–16. DOI: 10.1080/14790710408668175

Geertzen, J., Alexopoulou, T., Baker, R., Hendriks, H., Jiang, S. & Korhonen, A. 2013a. *The EF Cambridge Open Language Database (EFCAMDAT): User Manual Part I: Writtings*. Available at http://corpus.mml.cam.ac.uk/EFCAMDAT/ EFCAMDAT User Manual v02. pdf. (accessed 19 November 2014).

Geertzen, J., Alexopoulou, T. & Korhonen, A. 2013b. "Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT)". In R. T. Miller, K. I. Martin, C. M. Eddingon, A. Henery, N. Marcos Miguel, A. M. Tseng, A. Tuninetti & D. Walter (Eds.), *Proceedings of the 31st Second Language Research Forum (SLRF), Carnegie Mellon*. Cascadilla Proceedings Project, 240–254.

Granger, S. 1998. *Learner English on Computer*. London: Longman.

Granger, S. 2008. "Learner corpora". In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin and New York: Walter de Gruyter, 259–275.

Granger, S., Dagneaux, E. & Meunier, F. 2002. *International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English*. Version 2 (Handbook + CD-ROM). Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G. & Zampa, V. 2007. "Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness", *ReCaLL* 19(3), 252–268. DOI: 10.1017/S0958344007000237

Hockenmaier, J. & Steedman, M. 2007. "CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank", *Computational Linguistics* 33(3), 355–396. DOI: 10.1162/coli.2007.33.3.355

Lardiere, D. 1998. "Dissociating syntax from morphology in a divergent L2 end-state grammar", *Second Language Research* 14(4), 359–375. DOI: 10.1191/026765898674105303

Lozano, C. & Mendikoetxea, A. 2013. "Learner corpora and second language acquisition: The design and collection of CEDEL2". In N. Ballier, A. Díaz-Negrillo & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, 65–100. DOI: 10.1075/scl.59.06loz

Meunier, F. and Littré, D. 2013. "Tracking learners' progress: adopting a dual corpus cum experimental data approach", *Modern Language Journal* 97, 61–76. DOI: 10.1111/j.1540-4781.2012.01424.x

Meurers, D. 2009. "On the automatic analysis of learner language", *CALICO Journal* 26(3), 469–473.

Miller, G. A. 1995. "WordNet: a lexical database for English", *Communications of the ACM* 38(11), 39–41. DOI: 10.1145/219717.219748

Murakami, A. 2013. *L1 Influence and Individual Variation in the L2 Accuracy Development of Grammatical Morphemes: Insights from Learner Corpora*. Unpublished doctoral dissertation, University of Cambridge, UK.

Myles, F. 2008. "Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues". In L. Ortega and H. Byrnes (Eds.), *The longitudinal Study of Advanced L2 Capacities*. New York and London: Routledge, 58–72.

Myles, F. 2012. "Complexity, accuracy and fluency; the role played by formulaic sequencies in early interlanguage development". In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA, Language Learning & Language Teaching*. Amsterdam & Philadelphia: John Benjamins, 71–94. DOI: 10.1075/lllt.32.04myl

Myles, F. & Mitchell, R. 2007. *French learner language oral corpora (FLLOC)*. Available at http://www.flloc.soton.ac.uk/ (accessed 19 November 2014).

O'Donnell, M. B., Römer, U. & Ellis, N. C. 2013. "The development of formulaic sequences in first and second language writing: investigating effects of frequency, association and native form", *International Journal of Corpus Linguistics* 18(1), 83–108. DOI: 10.1075/ijcl.18.1.07odo

Orasan, C. & Evans, R. 2007. "NP animacy identification for anaphora resolution", *Journal of Artificial Intelligence Research* 29, 79–103.

Ortega, L. 2009. *Understanding Second Language Acquisition*. London: Hodder Education/Routledge.

Paquot, M. 2013. "Lexical bundles and L1 transfer effects", *International Journal of Corpus Linguistics* 18(13), 391–417. DOI: 10.1075/ijcl.18.3.06paq

Perdue, C. 1993. *Adult Language Acquisition: Volume I: Field Methods*. Cambridge University Press.

Rimell, L., Clark, S. & Steedman, M. 2009. "Unbounded dependency recovery for parser evaluation". In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*. Association for Computational Linguistics, 813–821.

Robinson, P. and Ellis, N. C. 2008. *Handbook of Cognitive Linguistics and Second Language Acquisition*. London and New York: Routledge.

Schachter, J. 1974. "An error in error analysis", *Language Learning* 24, 205–214. DOI: 10.1111/j.1467-1770.1974.tb00502.x

Selinker, L. 1972. "Interlanguage", *International Review of Applied Linguistics in Language Teaching* 10(1–4), 209–232. DOI: 10.1515/iral.1972.10.1-4.209

Shirai, Y. & Ozeki, H. 2007. "Introduction to the special issue: The acquisition of relative clauses and the noun phrase accessibility hierarchy: a universal in SLA?", *Studies in Second Language Acquisition* 29, 55–167. DOI: 10.1017/S027226310707009X

Sinclair, J. 2005. "How to build a corpus". In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford: Oxbow Books, 79–83.

Steedman, M. 2000. *The Syntactic Process*. Cambridge: MIT Press.

Tavakoli, P. & Foster, P. 2008. "Task design and second language performance: the effect of narrative type on learner output", *Language Learning* 58(2), 439–473. DOI: 10.1111/j.1467-9922.2008.00446.x

Team, R. C. 2008. *R: a language and environment for statistical computing*. Vienna: Foundation for Statistical Computing.

Tizón-Couto, B. 2013. *Clausal Complements in Native and Learner Spoken English. A Corpus-based Study with LINDSEI and VICOLSE*. Bern: Peter Lang.

Vyatkina, N. 2012. "The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study", *The Modern Language Journal* 96, 576–598. DOI: 10.1111/j.1540-4781.2012.01401.x

White, L. 1989. *Universal Grammar and Second Language Acquisition*. Amsterdam and Philadelphia: John Benjamins. DOI: 10.1075/lald.1

Wray, A. 2002. *Formulaic Language and the Lexicon*. New York: Cambridge University Press. DOI: 10.1017/CBO9780511519772

Wulff, S., Ellis, N. C., Römer, U., Bardovi-Harlig, K. & LeBlanc, C. 2009. "The acquisition of tense-aspect: Converging evidence from corpora and telicity readings", *Modern Language Journal* 93, 354–369. DOI: 10.1111/j.1540-4781.2009.00895.x

Wulff, S., Lester, N. & Martinez-Garcia, M. T. 2014. "*That*-variation in German and Spanish L2 English", *Language and Cognition* 6, 271–299. DOI: 10.1017/langcog.2014.5

Author's address

Theodora Alexopoulou
University of Cambridge
Department of Theoretical and Applied Linguistics
English Faculty Building
9 West Road
Cambridge CB3 9DA
UK

ta259@cam.ac.uk