# Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat)

**Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen**
**University of Cambridge**

## 1. Introduction

Naturalistic learner productions are an important empirical resource for SLA research. Some pioneering works have produced valuable second language (L2) resources supporting SLA research.[1] One common limitation of these resources is the absence of individual longitudinal data for numerous speakers with different backgrounds across the proficiency spectrum, which is vital for understanding the nature of individual variation in longitudinal development.[2]

A second limitation is the relatively restricted amounts of data annotated with linguistic information (e.g., lexical, morphosyntactic, semantic features, etc.) to support investigation of SLA hypotheses and obtain patterns of development for different linguistic phenomena. Where available, annotations tend to be manually obtained, a situation posing immediate limitations to the quantity of data that could be annotated with reasonable human resources and within reasonable time. Natural Language Processing (NLP) tools can provide automatic annotations for parts-of-speech (POS) and syntactic structure and are indeed increasingly applied to learner language in various contexts. Systems in computer-assisted language learning (CALL) have used a parser and other NLP tools to automatically detect learner errors and provide feedback accordingly.[3] Some work aimed at adapting annotations provided by parsing tools to accurately describe learner syntax (Dickinson & Lee, 2009) or evaluated parser performance on learner language and the effect of learner errors on the parser. Krivanek and Meurers (2011) compared two parsing methods, one using a hand-crafted lexicon and one trained on a corpus. They found that the former is more successful in recovering the main grammatical dependency relations whereas the latter is more successful in recovering optional, adjunction relations. Ott and Ziai (2010) evaluated the performance of a dependency parser trained on native German (MaltParser; Nivre et al., 2007) on 106 learner answers to a comprehension task in L2 German. Their study indicates that while some errors can be problematic for the parser (e.g., omission of finite verbs) many others (e.g., wrong word order) can be parsed robustly, resulting in overall high performance scores.

In this paper we have two goals. First, we introduce a new English L2 database, the EF Cambridge Open Language Database, henceforth EFCAMDAT. EFCAMDAT was developed by the Department of Theoretical and Applied Linguistics at the University of Cambridge in collaboration with EF Education First, an international educational organization. It contains writings submitted to *Englishtown*, the

* We are grateful to Detmar Meurers, Henriëtte Hendriks, and Rachel Baker and colleagues for comments and discussion throughout the development of the database and its evaluation. We would also like to thank Dimitris Michelioudakis and Toby Hudson for manual annotation/correction of parsed text, and Laura Rimell for her advice and comments. We also gratefully acknowledge support by the Isaac Newton Trust at Trinity College Cambridge and EF Education First for this research.

[1] See for instance the ESF database, ZISA corpus, FALKO, and The International Corpus of Learner English (ICLE) among others—a comprehensive list of learner corpora around the world is compiled by Sylviane Granger's team at the Centre for English Corpus Linguistics at the Université Catholique de Louvain and can be found at https://www.uclouvain.be/en-cecl-lcworld.html.

[2] There are of course exceptions to this: for instance, the ESF database (Feldweg, 1991) and FLLOC (Myles & Mitchell, 2007) both follow groups of learners over a period.

[3] Such works use early rule-based parsing as in Barchan, Woodmansee, and Yazdani, 1986; Jensen, Heidorn, Miller, and Ravin, 1983 or more modern approaches as in Amaral and Meurers, 2011; Menzel and Schröder, 1999.

online school of EF accessed daily by thousands of learners worldwide. EFCAMDAT stands out for two reasons: its size and rich individual longitudinal data from learners with a wide variety of L1 backgrounds. The magnitude of EF operations has allowed us to build a resource of considerable size, currently containing half a million scripts from 85K learners summing up 33 million words. Crucially, the progress of individual learners can be followed over time. As new data come in, the database will continue to grow allowing investigation of longitudinal development of larger numbers of learners. EFCAMDAT is an open access resource available to the research community via a web-based interface at http://corpus.mml.cam.ac.uk/efcamdat/, subject to a standard user agreement.

Our second goal is to evaluate parser performance on EFCAMDAT data. Our study provides users of EFCAMDAT with information on the accuracy of the automatically obtained morpho-syntactic annotations that accompany EFCAMDAT data. The parser performance is evaluated using a set of manually annotated EFCAMDAT data. We provide information on the effect of different types of errors on parsing as well as aspects of learner language that are challenging for automated linguistic analysis. This article is structured as follows: section 2 elaborates on the nature and characteristics of the learner data. The syntactic annotations are evaluated and discussed in section 3. We conclude in section 4.

## 2. Data characteristics

EFCAMDAT consists of essays submitted to *Englishtown*, the online school of EF Education First, by language learners all over the world (Education First, 2012). A full course in *Englishtown* spans 16 proficiency levels aligned with common standards such as TOEFL, IELTS, and the Common European Framework of Reference for languages (CEFR) as shown in Table 1.

**Table 1:** *Englishtown* skill levels in relation (indicative) to common standards.

| *Englishtown* | 1-3 | 4-6 | 7-9 | 10-12 | 13-15 | 16 |
|---|---|---|---|---|---|---|
| Cambridge ESOL | - | KET | PET | FCE | CAE | - |
| IELTS | - | <3 | 4-5 | 5-6 | 6-7 | >7 |
| TOEFL iBT | - | - | 57-86 | 87-109 | 110-120 | - |
| TOEIC Listening & Reading | 120-220 | 225-545 | 550-780 | 785-940 | 945 | - |
| TOEIC Speaking & Writing | 40-70 | 80-110 | 120-140 | 150-190 | 200 | - |
| CEFR | A1 | A2 | B1 | B2 | C1 | C2 |

Learners are allocated to proficiency levels after a placement test when they start a course at EF[4] or through successful progression through coursework. Each of the 16 levels contains eight lessons, offering a variety of receptive and productive tasks. EFCAMDAT consists of scripts of writing tasks at the end of each lesson on topics like those listed in Table 2.

**Table 2:** Examples of essay topics at various levels. Level and unit number are separated by a colon.

| ID | Essay topic | ID | Essay 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 |

Given 16 proficiency levels and eight units per level a learner who starts at the first level and completes all 16 proficiency levels would produce 128 different essays. Essays are graded by language teachers;

[4] Starting students are placed at the first level of each stage: one, four, seven, ten, thirteen, or sixteen.

learners may only proceed to the next level upon receiving a passing grade. Teachers provide feedback to learners using a basic set of error markup tags or through free comments on students' writing. Currently, EFCAMDAT contains teacher feedback for 36% of scripts.

The data collected for the first release of EFCAMDAT contain 551,036 scripts (with 2,897,788 sentences, and 32,980,407 word tokens) written by 84,864 learners. We currently have no information on the L1 backgrounds of learners.[5] Information on nationality is, thus, used as the closest approximation to L1 background. EFCAMDAT contains data from learners from 172 nationalities. Table 3 shows the spread of scripts across the nationalities with most learners.[6]

**Table 3:** Percentage and number of scripts per nationality of learners.

| Nationality | Percentage of scripts | Number of Scripts |
|---|---|---|
| Brazilians | 36.9% | 187,286 |
| Chinese | 18.7% | 96,843 |
| Russians | 8.5% | 44,187 |
| Mexicans | 7.9% | 41,115 |
| Germans | 5.6% | 29,192 |
| French | 4.3% | 22,146 |
| Italians | 4.0% | 20,934 |
| Saudi Arabians | 3.3% | 16,858 |
| Taiwanese | 2.6% | 13,596 |
| Japanese | 2.1% | 10,672 |

Few learners complete all of the proficiency levels. For many, their start or end of interacting with *Englishtown* fell outside the scope of the data collection period for the first release of EFCAMDAT. More generally, many learners only complete portions of the program. Nevertheless, around a third of learners (around 28K) have completed three full levels, corresponding to a minimum of 24 scripts.[7] Only 500 learners have completed every unit from level one to six (accounting for at least 48 scripts).

Characterizing scripts quantitatively is difficult, because of the variation across topics and proficiency levels. Texts range from a list of words or a few short sentences to short narratives or articles. As learners become more proficient they tend to produce longer scripts. On average, scripts count seven sentences ($SD$ = 3.8). Sample scripts are shown in Figure 1.

## 3. Syntactic annotations

### *3.1. Background and motivation*

Many learner corpora have been annotated for errors and more recently for lemmas, parts of speech (POS), and grammatical relations (Díaz-Negrillo, Meurers, Valera, & Wunsch, 2010; Granger, 2003; Granger, Kraif, Ponton, Antoniadis, & Zampa, 2007; Lüdeling, Walter, Kroymann, & Adolphs, 2005; Meurers, 2009; Nicholls, 2003). Information on POS and grammatical relations allows the investigation of morpho-syntactic and also some semantic patterns. Such annotation typically involves two distinct levels, one providing category information (POS) and a second level providing syntactic structure, typically represented by phrase structure trees or grammatical dependencies between pairs of words.

As mentioned in the introduction, NLP tools can provide annotations automatically. One critical question is how learner errors and untypical learner production patterns affect parsing performance. POS taggers rely on a combination of lexical, morphological, and distributional information to provide the most likely POS annotation for a given word. But in learner language these three sources of information may be pointing to different outcomes (Díaz-Negrillo et al., 2010; Meurers, 2009; Ragheb & Dickinson,

[5] Metadata on the L1 background of learners is being collected for the second release of the database.

[6] Of the 172 nationalities, 28 have over 100 learners, and 38 nationalities over 50 learners.

[7] If learners don't receive a satisfying score on their writing, they may repeat the task, which means that a learner will have a minimum of eight scripts per level but may have more if they repeat the task.

1. LEARNER 18445817, LEVEL 1, UNIT 1, CHINESE

Hi! Anna,How are you? Thank you to sendmail to me. My name's Anfeng.I'm 24 years old.Nice to meet you !I think we are friends already,I hope we can learn english toghter! Bye! Anfeng.

2. LEARNER 19054879, LEVEL 2, UNIT 1, FRENCH

Hi, my name's Xavier. My favorite days is saturday. I get up at 9 o'clock. I have a breakfast, I have a shower... Then, I goes to the market. In the afternoon, I play music or go by bicycle. I like sunday. And you ?

3. LEARNER 19054879, LEVEL 8, UNIT 2, BRAZILIAN

Home Improvement is a pleasant protest song sung by Josh Woodward. It's a simple but realistic song that analyzes how rapid changes in a town affects the lives of many people in the name of progress. The high bitter-sweet voice of the singer, the smooth guitar along with the high pitched resonant drum sound like a moan recalling the past or an ode to the previous town lifestyle and a protest to the negative aspects this new prosperous city brought. I really enjoyed this song.

**Figure 1:** Three typical scripts, in which learners are asked to introduce themselves (1), describe their favourite day (2), and review a song for a website (3).

2011). In particular, Díaz-Negrillo et al. (2010) discuss a range of mismatches, as for instance between form and distribution (1-a) and stem and distribution (1-b). In (1-a) the verb *want* lacks 3rd person agreement. Thus, the distributional information corresponds to a 3rd-person verbal form tag (VBZ) while the morphological information is consistent with a bare verb form tag (VB). Similarly, the stem *choice* is of category noun, but in (1-b), the morphological and distributional information would indicate verbal tags.

(1) a. ...if he **want** to know this...
b. ... to be **choiced** for a job...

Díaz-Negrillo et al., 2010, ex.9,16

Turning to syntax, learners often make syntactic mistakes, e.g., word order mistakes involving misplaced adverbs as in *it brings rarely such connotations* (Granger et al., 2007).

Results discussed in Díaz-Negrillo et al. (2010) and Meurers (2009) indicate that taggers show robustness to many such mismatches and that good accuracy scores can be obtained for POS tagging. There is, though, the question of whether native annotations are suitable descriptions of learner language. Ragheb and Dickinson (2011) rightly argue that annotating learner language with existing tools is an instance of Bley-Vroman's *comparative fallacy*. Bley-Vroman (1989) has argued for the need to analyse learner language in its own right rather than with categories from the target language. In a series of papers, Ragheb and Dickinson (2011) and Dickinson and Ragheb (2009, 2011) propose an annotation model that seeks to capture the properties of learner language rather than to superimpose native tags. They propose a multilayered annotation scheme where distributional and morphological information is annotated separately. For example, the annotation for (1-a) would just record the mismatch. Under this view (a good part of) errors are mismatches between stem, morphological, and distributional information. Errors affect not only POS tagging but can also reduce the probability of parses (Wagner & Foster, 2009). At the same time, parsers show robustness to many errors, e.g., word order (Ott & Ziai, 2010).

In the following sections we present results on parser performance for EFCAMDAT data and discuss the effect of learner errors on parsing.

### *3.2. Methodology*

To evaluate how parsers for well-formed English perform on learner language, a sample from EFCAMDAT data was annotated with POS tags and syntactic structure. The Penn Treebank Tagset (Marcus, Marcinkiewicz, & Santorini, 1993) was used for POS annotation because it is widely used and offers a relatively simple and straightforward set of tags. It comprises 36 POS tags, as listed in the Appendix, and features a further 12 tags to mark punctuation and currency symbols. Syntactic structure has been annotated in the form of *dependency relations* between a pair of words, where one word is analyzed as a head (governor) and the other as a dependent (Tesnière & Fourquet, 1959). The main feature of this annotation scheme is the absence of constituency and phrasal nodes. Because of its less hierarchical structure, this framework is particularly suitable for manual annotation. An example is illustrated in Figure 2 and compared with a more standard phrase structure annotation.[8]

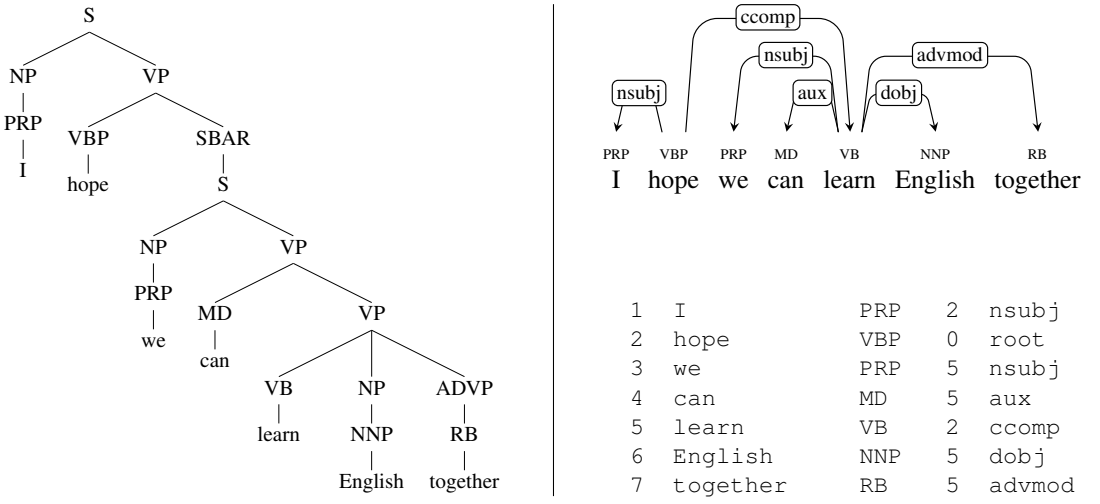


**Figure 2:** A sentence in phrase structure (left) and dependency relations (right) with the dependency graph (top right) and corresponding column-based coding (bottom right).

A set of 1,000 sentences (11,067 word tokens) from EFCAMDAT was pseudo-randomly sampled with equal representation from all 16 proficiency levels and five of the best represented nationalities (i.e., Chinese, Russian, Brazilian, German, and Italian). These sentences were then tagged and parsed using Penn Treebank POS tags and a freely available state-of-the-art parser (Stanford parser; Klein & Manning, 2003). Two trained linguists manually corrected the automatically annotated evaluation set (500 sentences each). They used the grammatical relations of the Stanford Dependency scheme (De Marneffe & Manning, 2008) and worked with collapsed representations.[9] The annotators marked learner errors and tagging/parsing errors manually on words. They used two error tags, L for learner error and P for POS tagging and parsing errors, and provided a corrected version. Each annotator corrected 500 sentences. Prior to the annotation, they both corrected a set of 100 additional sentences with corresponding parses in order to measure inter-annotator agreement. Inter-annotator agreement turned out to be 95.3% on the full combination of POS, attachment to the head, and relation type. All disagreements were subsequently discussed with a third linguist to reduce any incidental mistakes, which increased the final inter-annotator agreement to 97.1%.

One of our goals was to evaluate the effect of learner errors on automated annotation, since errors are rather frequent. In our sample of 1,000 sentences, 33.8% contains at least one learner error. These learner errors are mostly spelling and capitalization ones, but also involve morphosyntactic and semantic

[8] Our choice, therefore, does not reflect a theoretical preference but is guided by the need to adopt an intuitive and flexible annotation system that manual coders can manage easily.

[9] To simplify patterns for extraction, dependencies involving prepositions, conjuncts, as well as information about the referent of relative clauses are collapsed to get direct dependencies between content words.

irregularities or missing words. Some characteristic examples are illustrated in (2) and (3), which show learner errors together with corrections suggested by the annotators.

(2) a. I often have meetings **ar** go on business trips
correction: *and*
b. I finally **ger** an offer as **an** secretary in a company
correction: *get*, *a*
c. At night **i** go to bed
correction: **I**
d. Because to **whome** could **i** talk English in Germany?
correction: *whom*, *I*
e. I can't swim, sometimes I like **take** a walk
correction: *to take*
f. If you would like to work, like to volunteer, call **to** number 80000
correction: *delete 'to'*

(3) a. To help me with my vocabulary I note down all the words that I cannot recognize in a notebook **which enclosed** with a sentence for reference
correction: *delete 'which' and replace 'enclosed' with 'along'*
b. It must be by far the **exhilarating** experience
correction: *the most exhilarating*
c. you must **can** my house soon
correction: *be able to*
d. to motivate people to **reaching** aims they would not reach **theirselves**
correction: *reach*, *themselves*
e. I'll think about **change** my career
correction: *changing*

Examples in (2) and (3) give a sense of the range of difficulties. Some errors are more likely to affect the parser than others. For instance, the spelling error in (2-b) where we have *ger* instead of *get* may not be as damaging as the spelling error in (2-a) where instead of *and* we have *ar*. This is because it is easier to recover a verbal category than a conjunction based on distributional evidence. Similarly *whome* instead of *whom* in (2-d) may also have an effect on the parser as it may miss a wh-category. The errors in (2-e-f) could be viewed as subcategorization ones and it is conceivable that the parser can still recover the correct dependencies.[10]

The effect of an error like (3-a) is rather unpredictable. Such cases illustrate how challenging it can be to describe some errors. Under one analysis, this sentence may just have a missing *'is'* (*which is enclosed....*). However, our annotator preferred to rephrase. Cases like this require a systematic analysis of the learner's productions to establish whether such errors are an instance of a more general pattern (e.g., auxiliary BE omission) or not. It is hard to see how such issues can be addressed by error annotation schemes.

The errors in (3-b-c) are semantic. Though interesting from an SLA perspective, they are unlikely to affect the parser. The pair in (3-d-e) both involve an erroneous verbal form. But the effect on the parser can be different. The parser may establish the correct dependency between *reaching* and *motivate* in (3-d). But in (3-e) it may analyze *change* as a noun rather than a verbal category.

Our data also contain many cases of the mismatches discussed by Díaz-Negrillo et al. (2010) as illustrated in (4).

(4) a. but the young woman was **catched at** the hair by a passer-by
correction: *caught*, *by*
b. please find **enclose**
correction: *enclosed*

[10] See Dickinson and Ragheb (2009) for a discussion of annotation issues arising in relation to subcategorization errors in learner language.

c. I am also **certificated** in cardio kickboxing
correction: *certified*

Figure 3 illustrates a case where a learner error results in a parsing error. The sentence contains what looks like a preposition in place of an article, and the parser analyzes the item as a preposition and postulates the wrong dependency, where the noun *sweater* is the object of the preposition. The annotator corrects the direct object interpretation for the noun, and provides corrections as shown in Figure 3.

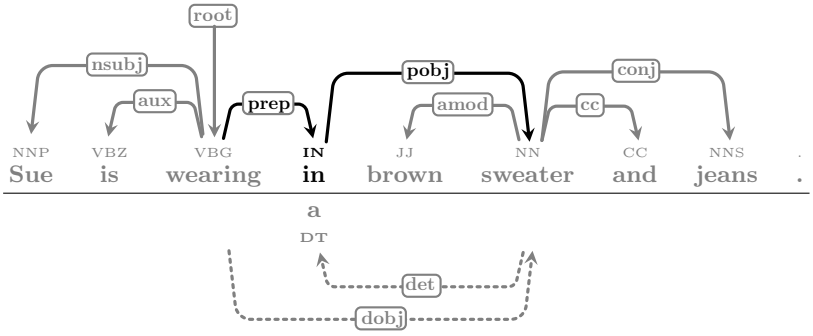


**Figure 3:** Example of a parsed sentence with learner and parsing errors (above), and manual corrections (below).

To assess the parser's performance on L2 English and the effect of errors, it is crucial to know if parsing errors occur in the context of well-formed or ungrammatical sentences. Manual annotation of learner errors and parsing errors allows us to identify the following three scenarios:

1. learner error without a POS-tagging/parsing error;
2. learner error with a POS-tagging/parsing error;
3. a POS-tagging/parsing error without a learner error.

The first scenario may arise with semantic or morphological errors at word level that are not strong enough to affect the grammatical rule selection during parsing. Such a situation is depicted by Figure 4, in which the surface form of what appears to be intended as the verb *get* is correctly identified as such.

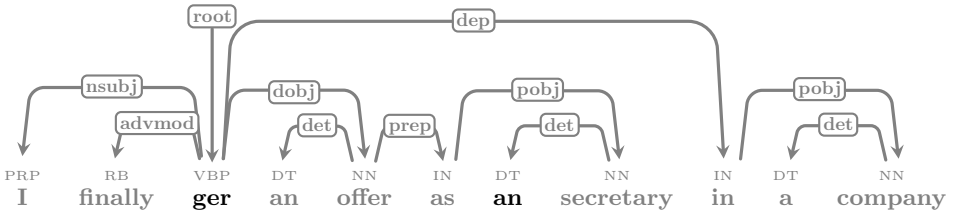


**Figure 4:** Learner errors (in black) without a parsing error.

The second scenario is one in which a learner error causes a parsing error, as exemplified in Figure 5. Both 'mop' and 'de' are identified as foreign words and interpreted to be compound nouns (see also Figure 3).

Finally, parsers are not always accurate even with grammatically correct native English. The third scenario, therefore, involves a parsing error without a learner error as illustrated in Figure 6 where the word *laundry* is misanalysed as a subject of *afternoon* rather than an object of the verb *does* and *afternoon* as a clausal complement rather than a modifier of the verb. Our goal here is to evaluate how often each of these scenarios is realized.

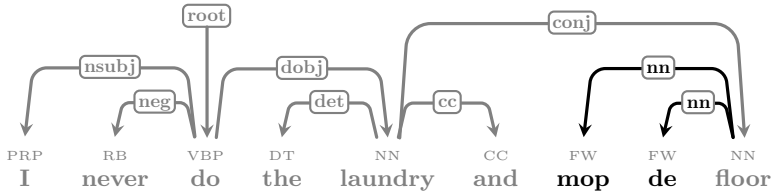


**Figure 5:** Learner error with a parsing error (both in black).

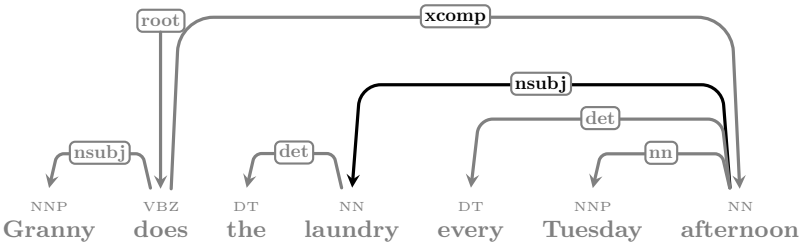


**Figure 6:** Parsing error (in black) without learner error.

### *3.3. Results*

Parser errors can be measured in various ways. Measurement may focus on errors per sentence or per word. Evaluation can also consider different aspects of annotation, POS tagging, dependency attachment, and dependency relation label. For instance Figure 5 contains POS errors for *mop* and *de* which are tagged as foreign words. This tagging error results in an attachment and labelling error since these two words are misanalysed as dependents of *floor* in a noun compound dependency. This is a case where a POS error has a knock-on effect on the syntactic annotation. But parsing errors may also appear without a POS error. For instance, in Figure 6 the dependency between the verb and *afternoon* is mislabelled as XCOMP (clausal complement) instead of modification, but this misanalysis is not triggered by a tagging error.

Parsing performance involving dependency relations is typically evaluated using *labelled attachment score* (LAS) and *unlabelled attachment score* (UAS). The former metric is the proportion of word tokens that are assigned both the correct head and the correct dependency relation label. The latter is the proportion of word tokens that are assigned the correct head regardless of what dependency relation label is assigned. A correct assignment of a dependency relation has to match the (manually) annotated relation exactly.[11] Over all our sentences, the Stanford parser scored 89.6% LAS and 92.1% UAS as shown in Table 4. This slightly exceeds results on Wall Street Journal (WSJ; Marcus et al., 1993) news text parsing at 84.2% LAS, 87.2% UAS (Cer, Marneffe, Jurafsky, & Manning, 2010). This measure does not include POS tagging. POS tagging also reaches a high accuracy of 96.1%. When POS tagging errors were included in the scoring this yielded 88.6% LAS and 90.3% UAS.

To get an idea of how parsing errors are distributed across sentences, accuracies were also computed at sentence level. In our sample of 1,000 sentences, 54.1% was without any labelled attachment error and 63.2% was without any unlabelled attachment error. As for POS tagging, 73.1% of sentences was error free.

The high accuracy scores indicate that the parser can recover the correct syntactic structure despite learner errors as long as the likelihood of the overall dependency tree is sufficiently high (see Figure 4). To assess how the presence of learner errors affected parsing performance, we split the 1,000 sentences into two sets, the 33.8% that contain at least one learner error and the remaining 66.2% without learner

[11] A more lenient way of evaluating parsing performance could take the hierarchical ordering of the dependency relations in the Stanford typed dependency scheme into account. Under- or over-specification could then be treated as correct or partially correct. For instance, a passive nominal subject (`nsubjpass`) that gets labelled as a subject (`subj`) can be considered less of an error than a direct object (`dobj`) and a subject.

errors, and parsed and evaluated them separately. For convenience, we will refer to these sets as the learner error absent (LA) set and learner error present (LP) set, respectively. Scores are shown in Table 4. As expected, the difference in parsing accuracy between sentences with and without learner errors is considerable.

**Table 4:** Accuracy scores for sentences with and without learner errors.

| Sentences | LAS | UAS |
|---|---|---|
| All | 89.6% | 92.1% |
| Without learner error (LA) | 92.6% | 95.0% |
| With at least one learner error (LP) | 83.8% | 87.4% |

In the LP set, 593 words were directly associated to a learner error. Of those words, 49.2% received an incorrect POS or syntactic dependency assigned. Note that a small percentage of these 49.2% words would also have received erroneous assignments in absence of the learner error. To obtain more detailed information on the effect of learner errors, we assessed the nature of assignment errors to words. To this purpose, we selected in each set of sentences the words with assignment errors, and looked at the error type, which can relate to POS, head attachment, and the relation. Table 5 shows percentages of assignment error types for individual words, with and without learner errors. POS tagging errors represent 21.6% for the LP set. Not surprisingly, in 53% of erroneous assignments, a POS error causes also a head and a relation error, which on themselves occur considerably less frequent. The picture is rather different for the LA set. POS tagging errors in isolation account for only 7.3% while head+relation errors in absence of POS tagging errors are the most challenging category for this set, accounting for 41.4% of all assignment errors. In sum, learner errors affect primarily category assignment (POS tagging), and as a consequence, assignment of a dependency head and labelling of the relation. Nonetheless, automatic assignment is still accurate for 50.8% words in the LP set, indicating robustness to a significant number of learner errors.

**Table 5:** Error types of assignment errors in the presence of a learner error (based on 292 words of the LP set) and in the absence of a learner error (based on 572 words of the LA set), where 'only' means that all other aspects are correctly assigned.

| Error type | LA set | LP set |
|---|---|---|
| POS only | 7.3% | 21.6% |
| Head only | 18.7% | 2.4% |
| Relation only | 16.3% | 3.4% |
| Head + Relation only | 41.4% | 8.6% |
| POS + Head + Relation | 9.3% | 53.1% |

Our next question was whether proficiency interacts with parser performance. Our sample was balanced for proficiency, but it could have been possible that at lower proficiency levels accuracy scores were significantly lower.[12] We, thus, calculated scores for levels 1-6 and 7-16 and 10-16 separately, shown in Table 6. Accuracy scores are higher at higher proficiency levels but the effect seems small.

**Table 6:** Accuracy scores for different levels of proficiency.

| Proficiency level | LAS | UAS |
|---|---|---|
| L1-6 | 89.0% | 91.5% |
| L7-16 | 90.2% | 92.4% |
| L10-16 | 91.4% | 93.2% |

[12] We would like to thank two anonymous reviewers for raising this point.

| *human* \ *machine* | acomp | advmod | amod | appos | auxpass | ccomp | conj-and | dep | dobj | nn | nsubj | nsubjpass | poss | prep-in | tmod | xcomp | # |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ROOT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 |
| advcl | 0 | 0 | 0 | 0 | 0 | 18 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 |
| advmod | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 26 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 53 |
| amod | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 29 | 14 | 0 | 0 | 0 | 0 | 0 | 34 |
| ccomp | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |
| conj-and | 0 | 0 | 11 | 11 | 0 | 9 | 0 | 18 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 53 |
| cop | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 |
| dep | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 | 0 | 0 | 0 | 0 | 0 | 13 |
| det | 0 | 0 | 17 | 0 | 0 | 0 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 |
| dobj | 6 | 4 | 0 | 0 | 0 | 0 | 3 | 7 | 0 | 3 | 37 | 0 | 0 | 0 | 4 | 3 | 107 |
| iobj | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 50 | 0 | 0 | 0 | 0 | 0 | 8 |
| mark | 0 | 45 | 0 | 0 | 0 | 0 | 0 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 |
| mwe | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 |
| nn | 0 | 0 | 29 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 31 |
| nsubj | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 8 | 0 | 12 | 8 | 0 | 0 | 0 | 49 |
| prep-to | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 10 |
| rcmod | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| ref | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 7 |
| tmod | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 47 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 |

**Figure 7:** Error matrix containing percentages of the most frequent errors in assigning dependency relation labels.

As some dependencies are more challenging for the parser to recover than others, we also measured accuracy for individual grammatical dependencies. The evaluation measures used were Precision, Recall, and $F_1$. Recall measures how many of the annotated instances of a specific grammatical relation are correctly recovered by the parser. Precision measures how many of the instances labelled by the parser as being a specific grammatical relation are actually correct. The $F_1$-score is the harmonic mean of precision and recall.[13] The grammatical relations that occur at least five times in the evaluation data are listed in Table 7, together with the evaluation metrics and the absolute frequency.

As can be seen in Table 7, only a few grammatical relations scored below 90% $F_1$. Adjectival complements (`acomp`, 73.3% $F_1$) suffered from a rather low recall, as did clausal subjects (`csubj`, 57.1% $F_1$) and appositional modifiers (`appos`, 75.8% $F_1$).

We next examined the most frequent parser 'confusions'. The erroneous assignment of dependency relation labels is quantified using the error matrix in Figure 7, in which each column represents the assigned label by the parser, while each row represents the actual label. Only the more frequent errors (i.e., occurring at least four times) have been included, and together account for 38% of all parsing errors. The number at the end of each row is the absolute frequency at which the label on that row was erroneously assigned, with each cell containing the percentage of this number that got erroneously assigned by the corresponding column label.

One prominent parsing error was misanalysis of direct objects (107 cases). In 37% of these cases, the direct object was mis-analysed as a nominal subject. This case is illustrated with Figure 6, in which the noun *laundry* is wrongly assigned to the nominal head *afternoon* as subject. Few errors were expected, such as noun compounds (NN) with (ADJECTIVAL MODIFIER), or a few indirect objects (IOBJ) with direct objects (DOBJ) or nominal subjects (NSUBJ), but there are no clear outliers. Only the DEPENDENT relation (DEP) has been used often as it is the most underspecified category in the dependency scheme.

[13] The $F_1$-score combines precision ($P$) and recall ($R$), and is defined as $2 * P * R/(P + R)$.

**Table 7:** Grammatical relations occurring at least five times. Precision (*Prec*), recall (*Rec*) and $F_1$ scores are listed for each grammatical relation, together with its absolute frequency (*#GRs*).

| *Relation* | *Description* | *Prec* | *Rec* | $F_1$ | *#GRs* |
|---|---|---|---|---|---|
| root | root | 95.2 | 97.2 | 96.2 | 999 |
| dep | dependent | 92.5 | 65.7 | 76.8 | 210 |
| aux | auxiliary | 98.0 | 98.9 | 98.4 | 646 |
| auxpass | passive auxiliary | 96.1 | 96.1 | 96.1 | 78 |
| cop | copula | 96.6 | 98.7 | 97.6 | 313 |
| arg | argument | | | | |
| agent | agent | 100.0 | 87.5 | 93.3 | 8 |
| comp | complement | | | | |
| acomp | adjectival complement | 78.6 | 68.8 | 73.3 | 16 |
| ccomp | clausal complement with internal subject | 93.0 | 86.0 | 89.4 | 189 |
| xcomp | clausal complement with external subject | 93.8 | 94.2 | 94.0 | 243 |
| complm | complementizer | 98.1 | 100.0 | 99.0 | 52 |
| obj | object | | | | |
| dobj | direct object | 93.8 | 95.7 | 94.7 | 694 |
| iobj | indirect object | 81.8 | 100.0 | 90.0 | 18 |
| pobj | object of preposition | 100.0 | 85.2 | 92.0 | 27 |
| mark | marker (word introducing an advcl) | 91.9 | 97.8 | 94.8 | 93 |
| subj | subject | | | | |
| nsubj | nominal subject | 97.7 | 96.4 | 97.0 | 1343 |
| nsubjpass | passive nominal subject | 95.6 | 94.2 | 94.9 | 69 |
| csubj | clausal subject | 66.7 | 50.0 | 57.1 | 8 |
| cc | coordination | 96.4 | 100.0 | 98.2 | 28 |
| conj | conjunct | 93.2 | 97.6 | 95.3 | 463 |
| expl | expletive (expletive “there”) | 100.0 | 100.0 | 100.0 | 40 |
| mod | modifier | | | | |
| amod | adjectival modifier | 96.9 | 95.9 | 96.4 | 657 |
| appos | appositional modifier | 89.3 | 65.8 | 75.8 | 38 |
| advcl | adverbial clause modifier | 86.6 | 97.0 | 91.5 | 100 |
| det | determiner | 98.3 | 99.4 | 98.9 | 1024 |
| predet | predeterminer | 90.0 | 100.0 | 94.7 | 10 |
| preconj | preconjunct | 83.3 | 100.0 | 90.9 | 5 |
| infmod | infinitival modifier | 85.7 | 80.0 | 82.8 | 15 |
| partmod | participial modifier | 93.9 | 86.1 | 89.9 | 36 |
| advmod | adverbial modifier | 95.2 | 95.4 | 95.3 | 507 |
| neg | negation modifier | 98.9 | 100.0 | 99.5 | 91 |
| rcmod | relative clause modifier | 92.9 | 89.7 | 91.2 | 58 |
| quantmod | quantifier | 87.0 | 100.0 | 93.0 | 20 |
| nn | noun compound modifier | 92.7 | 89.9 | 91.3 | 300 |
| npadvmod | noun phrase adverbial modifier | 95.7 | 91.7 | 93.6 | 24 |
| tmod | temporal modifier | 84.6 | 87.3 | 85.9 | 63 |
| num | numeric modifier | 96.9 | 95.4 | 96.2 | 132 |
| number | element of compound number | 100.0 | 100.0 | 100.0 | 10 |
| prep | prepositional modifier | 95.7 | 96.3 | 96.0 | 975 |
| poss | possession modifier | 99.1 | 98.8 | 99.0 | 336 |
| prt | phrasal verb particle | 95.2 | 100.0 | 97.5 | 61 |
| parataxis | parataxis | 62.5 | 100.0 | 76.9 | 5 |

### *3.4. Discussion*

The most striking result of this evaluation is the high accuracy scores of the automated annotation tools which demonstrates that the tools are robust to learner language. These scores are high even for lower proficiency levels (Table 6). There are a number of explanations for this result.

The first relates to the overall simplicity of learner language in comparison to native productions. For instance, learner sentences tend to be shorter and therefore less demanding for the parser. The average word length of sentences in this evaluation set is 11 words, whereas that of sentences in the British National Corpus (BNC; Burnard, 1995) is 16 words, and that of the WSJ is 21 words. It is possible then that the shorter average sentence length of learner productions compensates for the effect of learner errors on the parser.

A second explanation relates to the nature of errors and the finding that automated tools show robustness to at least half the learner errors. Many errors are semantic and do not affect parser performance which targets syntactic structure. Consider for instance the examples in (5) from our sample. The parser has correctly identified the relevant grammatical dependencies, despite the learner errors (this is also true for also examples (3-b) and (3-c)).

(5) a. ... I play all other **kind** of music....
b. .... the best convenience of **the** modern technology ....
c. .... it is a **potential** large group...
d. ... my older brother is 40 and my **littler** brother is ...
e. ... my wife is wearing **a** white and pink pants....
f. ....I try to visualise pleasant and restful **place** ...

Similarly, a good part of what appear to be subcategorization errors or choice of preposition errors do not affect the parser's ability to obtain the correct dependencies, see examples (2-e) and (2-f) and (6).

(6) a. I encourage you to continue your support **in** respect **of** my presidency
b. I am looking forward **for** your feedback...
c. The body language in Russia is very similar **with** American one.

The erroneous patterns we see in (5) and (6) require a systematic investigation to be modelled accurately. At the same time, the robustness of the parser is welcome since it allows us to obtain the main syntactic patterns that *are* used by learners. Consider for instance, the sentence mentioned earlier with an adverb intervening between the verb and its object *it brings rarely such connotations*. The parse of this sentence is shown in Figure 8. The parser 'ignores' the word order violation and correctly analyzes *rarely* as an adverbial modifier and *such connotations* as a nominal direct object. Similarly, the parser identifies a relative clause in (3-a) and analyzes *to reaching* in (3-d) as a complement of the main verb. The robustness of the parser in these cases is important, because it captures syntactic structures that *are* used in learner language. In sum, the parser often abstracts away from local morphosyntactic information and word order violations but succeeds with the underlying syntactic dependencies. While this results in an incomplete picture of learner language, it nevertheless reveals an important aspect of the grammar underlying learner language. The current dependency annotations can be used to investigate the range of word order violations and the way different types of dependents (e.g., adverbs vs. indirect objects) may intervene between the verb and the direct object.

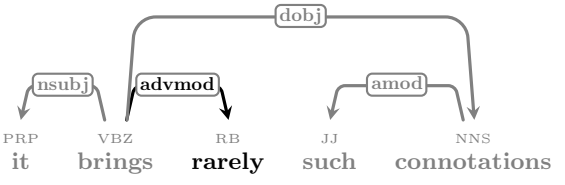


**Figure 8:** Sentence with a word order error.

Let us briefly consider the mismatches raised by Díaz-Negrillo et al. (2010) and Ragheb and Dickinson (2011). In general, the parser has resolved conflicting evidence in favor of morphological information (and our annotators have accepted morphology-based POS tags). Thus, *catched* in (4-a) and *certificated* in (4-c) are tagged as VBN (verb, past participle). Similarly, *want* in (1-a) is tagged as VBP (verb, non-3rd-person form). Such annotations do not reflect the relevant mismatches but enable a systematic investigation of the environments and lexical elements involved in them.

Finally, despite the very high accuracy scores, some patterns have been particularly challenging for the parser. For instance, absence of critical morphological information may lead to a category error in tagging which can then result in the wrong parse. For instance, *change* in (3-e) has been analyzed as a noun.

In conclusion, the relatively high accuracy scores of the parser can be explained through a combination of the overall simplicity of learner productions and the robustness of the parser that tends to focus on the primary dependency relations 'ignoring' local morphosyntactic information, some word order violations as well as semantic violations. Current NLP tools allow us to obtain reliable morphosyntactic annotations for learner language that are vital for investigating a wide range of lexical and morphosyntactic phenomena of L2 grammars.

## 4. Conclusions

EFCAMDAT is a new L2 English database built at the University of Cambridge, in collaboration with EF Education First. It consists of scripts submitted to *Englishtown*, the online school of EF Education First. It stands out for its size and rich individual longitudinal data from learners with a wide variety of backgrounds. It is an open access resource available to the research community via a web-based interface at *http://corpus.mml.cam.ac.uk/efcamdat/*, subject to a standard user agreement.

EFCAMDAT is tagged and parsed for POS and dependency relations using Penn Treebank tags and the Stanford parser. The core of this paper presented the results of an evaluation study on parser performance on a sample of 1,000 sentences from EFCAMDAT. The results indicate that parsing English L2 with current NLP tools generally yields performance close to that of parsing well-formed native English text. This is despite the fact that 33.8% of our sample of 1,000 learner sentences contains at least one learner error. Our analysis shows that current NLP tools are robust to a significant part of learner errors. This is because a good part of errors involve local morphosyntactic, subcategorization, word order, and semantic errors that do not affect the main dependency relations the parser targets. As a result, the parser can successfully capture syntactic patterns that are used by learners and provide valuable annotations for the investigation of a wide range of phenomena and SLA hypotheses. At the same time, some form errors (spelling, morphosyntactic) can lead to category (POS) errors when the distributional information cannot reliably define a category (e.g., co-ordination) or the form is ambiguous between a noun and verb category.

The robust performance of current NLP tools on L2 English demonstrates that they can be successfully used for automatic annotation of large scale databases like EFCAMDAT. Moreover, they provide a critical starting point for development of tools that can accurately model the erroneous and untypical patterns of learner language.

In this work we have used one particular state-of-the-art parsing algorithm, the Stanford parser, but various others could be used as well, such as the C&J reranking parser (Charniak & Johnson, 2005) and the Berkely parser (Petrov, Barrett, Thibaux, & Klein, 2006). It would be worth exploring other systems, in isolation and by using ensemble models that combine independently-trained models at parsing time to exploit complementary strengths (Sagae & Tsujii, 2007).

## Appendix: The Penn Treebank POS tagset (excl. punctuation tags)

| | | | | | |
|---|---|---|---|---|---|
| 1. | CC | Coordinating conjunction | 19. | PP | Possessive pronoun |
| 2. | CD | Cardinal number | 20. | RB | Adverb |
| 3. | DT | Determiner | 21. | RBR | Adverb, comparative |
| 4. | EX | Existential there | 22. | RBS | Adverb, superlative |
| 5. | FW | Foreign word | 23. | RP | Particle |
| 6. | IN | Preposition/subord. | 24. | SYM | Symbol |
| 7. | JJ | Adjective | 25. | TO | to |
| 8. | JJR | Adjective, comparative | 26. | UH | Interjection |
| 9. | JJS | Adjective, superlative | 27. | VB | Verb, base form |
| 10. | LS | List item marker | 28. | VBD | Verb, past tense |
| 11. | MD | Modal | 29. | VBG | Verb, gerund/present participle |
| 12. | NN | Noun, singular or mass | 30. | VBN | Verb, past participle |
| 13. | NNS | Noun, plural | 31. | VBP | Verb, non-3rd ps. sing. present |
| 14. | NNP | Proper noun, singular | 32. | VBZ | Verb, 3rd ps. sing. present |
| 15. | NNPS | Proper noun, plural | 33. | WDT | wh-determiner |
| 16. | PDT | Predeterminer | 34. | WP | wh-pronoun |
| 17. | POS | Possessive ending | 35. | WP | Possessive wh-pronoun |
| 18. | PRP | Personal pronoun | 36. | WRB | wh-adverb |

## References


Amaral, Luiz, & Meurers, Detmar. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, *23*, 4–24.

Barchan, Jonathan, Woodmansee, B., & Yazdani, Masoud. (1986). A PROLOG-based tool for French grammar analysis. *Instructional Science*, *15*, 21–48.

Bley-Vroman, Robert. (1989). What is the logical problem of foreign language learning? In Susan M. Gass, & Jacquelyn Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 41–68). New York: Cambridge University Press.

Burnard, Lou. (1995, May). *Users Reference Guide for the British National Corpus*. http://info.ox.ac.uk/bnc/.

Cer, Daniel, Marneffe, Marie-Catherine De, Jurafsky, Daniel, & Manning, Christopher D. (2010). Parsing to Stanford dependencies: Trade-offs between speed and accuracy. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 1628–1632). Valletta, Malta.

Charniak, Eugene, & Johnson, Mark. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 173–180). Stroudsburg, PA, USA: Association for Computational Linguistics.

De Marneffe, Marie-Catherine, & Manning, Christopher D. (2008). The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation* (pp. 1–8).

Díaz-Negrillo, Ana, Meurers, Detmar, Valera, Salvador, & Wunsch, Holger. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, *36*, 139–154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

Dickinson, Markus, & Lee, Chong Min. (2009). Modifying corpus annotation to support the analysis of learner language. *CALICO Journal*, *26*, 545–561.

Dickinson, Markus, & Ragheb, Marwa. (2009). Dependency annotation for learner corpora. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud, & Frank Van Eynde (Eds.), *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories* (pp. 59–70). Milan, Italy.

Dickinson, Markus, & Ragheb, Marwa. (2011). Dependency annotation of coordination for learner language. In Kim Gerdes, Eva Hajicova, & Leo Wanner (Eds.), *Proceedings of the International Conference on Dependency Linguistics (Depling)* (pp. 135–144). Barcelona, Spain.

Education First. (2012). Englishtown. http://www.englishtown.com/.

Feldweg, Helmut. (1991). *The European Science Foundation Second Language Database*. Nijmegen, Netherlands: Max Planck Institute for Psycholinguistics.

Granger, Sylviane. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO*, *20*, 465–480.

Granger, Sylviane, Kraif, Olivier, Ponton, Claude, Antoniadis, Georges, & Zampa, Virginie. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, *19*, 252–268.

Jensen, Karen, Heidorn, George E., Miller, Lance A., & Ravin, Yael. (1983). Parse fitting and prose fixing: Getting a hold on ill-formedness. *Computational Linguistics*, *9*, 147–160.

Klein, Dan, & Manning, Christopher D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (pp. 423–430). Stroudsburg, PA, USA: Association for Computational Linguistics.

Krivanek, Julia, & Meurers, Detmar. (2011). Comparing rule-based and datadriven dependency parsing of learner language. In *Proceedings of the International Conference on Dependency Linguistics (Depling 2011)* (pp. 128–132).

Lüdeling, Anke, Walter, Maik, Kroymann, Emil, & Adolphs, Peter. (2005). Multi-level error annotation in learner corpora. In *Proceedings from the Corpus Linguistics Conference Series* (Vol. 1). Birmingham, UK.

Marcus, Mitchell P., Marcinkiewicz, Mary Ann, & Santorini, Beatrice. (1993, June). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

Menzel, Wolfgang, & Schröder, Ingo. (1999). Error diagnosis for language learning systems. *ReCALL*, *11*, 20–30.

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

Myles, Florence, & Mitchell, Rosamond. (2007). French learner language oral corpora. http://www.flloc.soton.ac.uk/. University of Southampton.

Nicholls, Diane. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics Conference* (pp. 572–581). Lancaster University: University Centre for Computer Corpus Research on Language.

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kubler, Sandra, … & Marsi, Erwin. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, *13*, 95.

Ott, Niels, & Ziai, Ramon. (2010). Evaluating dependency parsing performance on German learner language. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (NEALT 2010)* (pp. 175–186).

Petrov, Slav, Barrett, Leon, Thibaux, Romain, & Klein, Dan. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 433–440). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ragheb, Marwa, & Dickinson, Markus. (2011). Avoiding the comparative fallacy in the annotation of learner corpora. In Gisela Granena, Joel Koeth, Sunyoung Lee-Ellis, Anna Lukyanchenko, Goretti Prieto Botana, & Elizabeth Rhoades (Eds.), *Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA research, dimensions, and directions* (pp. 114–124). Somerville, MA, USA: Cascadilla Proceedings Project.

Sagae, Kenji, & Tsujii, Jun'ichi. (2007). Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL* (Vol. 7, pp. 1044–1050).

Tesnière, Lucien, & Fourquet, Jean. (1959). *Eléments de syntaxe structurale*. Paris: Klincksieck.

Wagner, Joachim, & Foster, Jennifer. (2009). The effect of correcting grammatical errors on parse probabilities. In *Proceedings of the 11th International Conference on Parsing Technologies* (pp. 176–179). IWPT '09. Paris, France: Association for Computational Linguistics.

# Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines

edited by
Ryan T. Miller, Katherine I. Martin,
Chelsea M. Eddington, Ashlie Henery,
Nausica Marcos Miguel, Alison M. Tseng,
Alba Tuninetti, and Daniel Walter

**Cascadilla Proceedings Project Somerville, MA 2014**



ISBN 978-1-57473-464-5 library binding



## Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

## Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Geertzen, Jeroen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In *Selected Proceedings of the 2012 Second Language Research Forum*, ed. Ryan T. Miller et al., 240-254. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #3100.