[Accepted manuscript]

# The CELI Corpus: design and linguistic annotation of a new online learner corpus

Stefania Spina, Irene Fioravanti, Luciana Forti, Fabio Zanda

## Abstract

This article introduces the CELI Corpus[1], a new learner corpus of written Italian consisting of ca. 600,000 tokens, evenly distributed among CEFR proficiency levels B1, B2, C1 and C2. The collected texts derive from the language certification exams administered by the University for Foreigners of Perugia all around the world. The corpus contains rich metadata pertaining to text-related and learner-related variables. It expands the domain of learner corpora by being, among other things, both freely available online to the research community, and by focusing on a target language other than English. The article also presents and evaluates the pos-tagging procedure, thus contributing to best practices in learner corpus annotation.

## 1. Introduction

In learner corpus research, critical reflection on design criteria is crucial in structuring the rich reservoir of empirical data that is typical of corpora in line with the needs of SLA research (Tono, 2003; Gilquin, 2015). As language learning is, by definition, a developmental process taking place over time, empirical data collected and organised longitudinally or pseudo-longitudinally are of considerable interest (Myles, 2005; Gilquin, 2015). Furthermore, in the case of pseudo-longitudinal designs, text attribution to proficiency level is critical in order to ensure comparability among different studies (Carlsen, 2012). Additionally, the presence of

[1] https://www.unistrapg.it/cqpwebnew/, https://lt.eurac.edu/cqpweb/

balanced subcorpora within a corpus can allow systematic comparisons among the different parts that make up the corpus (e.g., among different proficiency levels) (Tracy-Ventura and Paquot, 2021). Finally, target languages other than English are needed in order to gain a broader view of second language acquisition processes and dynamics (Vyatkina, 2016; Lozano, 2021). However, an inspection of the learner corpora listed in the *Learner corpora around the world*[2] list reveals that most corpora developed so far lack one or more of these features. Most of them, in fact, are characterised by a cross-sectional design, while very few have a longitudinal or pseudo-longitudinal design, covering a significant timeframe or including balanced sets of proficiency levels (Meunier, 2015). Furthermore, the vast majority of learner corpora built so far refer to English as the target language, despite a few notable exceptions (e.g., Lozano, 2021; Vyatkina, 2016). Another issue related to corpus design concerns the ways in which a learner text is attributed to a certain proficiency level. This is an issue that has seldom been at the centre of learner corpus research discussion, despite proficiency level being arguably a "fuzzy variable" in the design of learner corpora (Carlsen, 2012).

In this paper, we seek to address some of the gaps that still characterise learner corpus research, by introducing the CELI corpus, a new corpus of L2 Italian writing. Our goal is to highlight the contribution that this corpus could make, with special reference to the domain of Italian L2 studies, which is still under-resourced as far as corpora are concerned. More particularly, the aim of this paper is twofold: (1) to present the CELI corpus, by illustrating its general architecture, the text- and learner-related variables it includes, the methods adopted in compiling it, and its contents; (2) to discuss the quality of the annotation procedures conducted on the corpus, by reporting on a study that measured and evaluated the performance of the pos-tagger, in light of the features that most typically characterise learner language. The next

[2] Centre for English Corpus Linguistics, Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (last accessed: 27/02/2023).

section reviews existing online learner corpora of Italian, with respect to size, design, proficiency levels, criteria for text attribution to CEFR proficiency levels, balancing criteria. A description of the CELI corpus, along with an evaluation of the reliability of the tagging procedures that were applied to it, will follow.

## 2. Online learner corpora of Italian

In this section, we review the learner corpora of written Italian currently available and searchable online. Although oral corpora of learner Italian are also available, we focus our review on written corpora only so as to reflect the specific domain in which the CELI corpus is situated. A total of eight corpora emerges from our search, which are listed in Table 1 in alphabetical order and in relation to size, design[3], proficiency levels, criteria adopted for text attribution to CEFR proficiency levels, and criteria adopted to create balanced subcorpora within the corpus. With specific reference to the last two aspects, we see that in order to attribute a text to a certain proficiency level, placements tests were used in half of the cases, that is in the corpora CAIL2 (Bratankova, 2015), COLI (Spina, forthcoming), CORITE (Bailini and Frigerio, 2018), and LOCCLI (Spina and Siyanova-Chanturia, 2018). Particularly in the context of Italian L2 language testing and assessment, placement tests, however, often lack the breadth and solidity of CEFR-based certification exams, and this hinders the reliability of text attribution to proficiency level. In the case of KOLIPSI, the texts were attributed to proficiency levels by professional CEFR raters, while in the case of MERLIN_IT they derived from language certification exams. In the cases of VALICO and LEONIDE_IT, no explicit proficiency levels are recorded. As for the balancing criteria, we can see that these are either non-existent (CORITE, KOLIPSI, VALICO), or alternatively refer to time spent studying

[3] Corpus design was categorised in terms of cross-sectional (e.g., one homogeneous proficiency level), pseudo-longitudinal (e.g., multiple proficiency levels, each of them represented by texts produced by different leaRner samples), and longitudinal (e.g., multiple proficiency levels, each of them represented by texts produced by the same learner sample).

Italian (CAIL2), number of learners per level (COLI), number of texts/learners per data collection point (LEONIDE_IT, LOCCLI), and number of texts per CEFR level (MERLIN_IT).

**Table 1**. Learner corpora of written Italian available and searchable online.

| | **Corpus name** | **Size** | **Design** | **Proficiency levels** | **Criteria for text attribution to CEFR proficiency levels** | **Balancing criteria** |
|---|---|---|---|---|---|---|
| 1 | CAIL2 | ca. 237,000 | Cross-sectional | B1, B2, C1, C2 | Placement test | Time spent studying Italian (in months) |
| 2 | COLI | ca. 44,637 (written component only) | Pseudo-longitudinal | B1, B2, C1 | Placement test | Number of participants per level |
| 3 | CORITE | ca. 103,000 | Part longitudinal, part pseudo-longitudinal | A1, A2, B1 | Placement test | N/A |
| 4 | KOLIPSI | ca. 800,000 | Cross-sectional | Intermediate, advanced | Raters assigned the texts to a specific CEFR level on the basis of a CEFR grid | N/A |
| 5 | LEONIDE_IT | ca. 93,000 | Longitudinal | Lower proficiency level | N/A | Number of texts/learners per data collection point |

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | LOCCLI | ca. 97,000 | Longitudinal | A1, A2, B1 | Placement test | Number of texts/learners per data collection point |
| 7 | MERLIN_IT | ca. 92,400 | Pseudo-longitudinal | A1, A2, B1, B2 | Raters assigned the texts to a specific CEFR level on the basis of a CEFR grid, in the context of language certification exams | Number of texts per CEFR level |
| 8 | VALICO | ca. 380,000 | Cross-sectional | Year of Italian language study | N/A | N/A |

## 3. The CELI corpus: description

### 3.1. Design

The CELI corpus is a pseudo-longitudinal corpus of Italian L2; its main goal is to be representative of written Italian produced by learners belonging to the intermediate and advanced levels of proficiency according to CEFR.

As Gilquin (2015) argues, a learner corpus should be designed by adopting specific criteria, "given the highly heterogeneous nature of interlanguage, which can be affected by many variables related to the environment, the task and the learner him-/herself" (2015: 16). Furthermore, Tracy-Ventura, Paquot and Myles (2021) suggest several recommendations to be considered in designing a learner corpus: to include L2s other than English; to build more multilingual corpora to promote cross-linguistic comparisons; to document all the stages of learning development including not only intermediate and advanced learners but also beginner learners; to include learners with different ages and with different L1s and from different contexts of learning; to reconsider what a 'control' corpus is and how it can be used in

comparing data; to collect metadata systematically and document them accurately including more learner and task variables; to document transcription and annotation stages; to include spoken data; and to collect longitudinal data. Moreover, they recommend making the learner corpus freely available (Tracy-Ventura, Paquot & Myles, 2021).

Among these recommendations, we adopted the following five: i. to include L2s other than English; ii. to include learners at different levels of proficiency, from different age groups and with varied L1s; iii. to collect metadata systematically and document them accurately; iv. to include varied task assignments[4]; and v. to make the learner corpus freely available. Further, another criterion was followed in designing the CELI corpus: to balance subcorpora in terms of tokens and make them comparable.

The above adopted criteria make the CELI corpus a reliable tool in the investigation of L2 Italian. First, it is representative of an L2 different from English (i.e., Italian), which is still an under-represented L2 in the LCR context. Second, it includes learners from different ages and from different levels of proficiency providing varied objective measures of proficiency. Third, metadata were systematically collected and are fully documented (as will be shown in subparagraphs 3.2. and 3.3., the CELI corpus presents different variables for both texts and learners). Fourth, its subcorpora are equally designed according to the same criteria and balanced in terms of tokens in order to make them comparable. Finally, the CELI corpus is a freely available and searchable corpus. Searchability is another crucial factor to consider in designing a learner corpus allowing different kinds of queries. To this end, the CELI Corpus is searchable from a CQPweb interface, on the basis of a range of metadata including CEFR level, learners' sex, learners' age, learners' nationality, exam centre location, task assignment ID, text genre and text type.

[4] Tracy-Ventura, Paquot and Myles (2021) recommend to include varied tasks in designing a learner corpus. In this specific case, we could not include different tasks as learner productions are derived from a specific production task of the written exam. However, we included different task assignments to which correspond different textual typologies and genres, as illustrated in Table 4.

### 3.2. Text variables

Written texts produced by Italian L2 learners were collected from the written examinations for the language certificates of Italian as a Foreign Language (CELI - *Certificati di Lingua Italiana*) developed by the *Center for Language Evaluation and Certification* (CVCL – *Centro Valutazione Certificazioni Linguistiche*) at the University for Foreigners of Perugia (Italy). For the purpose of the present project, the written texts were collected from CELI 2, CELI 3, CELI 4 and CELI 5, which certify Italian language knowledge with respect to proficiency levels B1, B2, C1, and C2 respectively. The CELI exams consist of an oral part and a written part.

The written part is articulated in different components: i. reading comprehension, ii. written production; iii. language competence; and iv. listening comprehension (Grego Bolli, 2004). The written production includes a series of production tasks. The texts contained in the CELI corpus were collected from one specific production task, for each CELI exam (Spina et al., 2022). Details of the production tasks for each CELI exam are shown in Table 2.

Table 2 - Production tasks' typology for each CELI exam.

| Exam | Typology | Word range |
|---|---|---|
| CELI 2 (B1) | A short letter or e-mail to write following a given task assignment. | 90-100 |
| CELI 3 (B2) | A short composition on personal experiences, situations, themes and topics of general interest to be chosen from two different task assignments. | 120-180 |
| CELI 4 (C1) | A composition to be chosen from two different task assignments on problems and phenomena in today's society, or a story about personal events and experiences, or a formal letter. | 220-250 |
| CELI 5 (C2) | A free composition to be chosen from three different task assignments that may relate to a report or essay, a | 330-360 |

| | |
|---|---|
| | fictional story, or a description of personal experiences including aspects of Italian civilisation. |

Several metadata are recorded for each text:

1) The identification number of the text;
2) The identification number of the exam centre where the candidates took the exam;
3) The task assignment to which the text is associated;
4) The CEFR level for which the candidate took and passed the certification exam (B1; B2; C1; C2);
5) The total score assigned to the whole exam;
6) The score band of the score on the whole exam (A; B; C);
7) The total score assigned to the written part of the exam;
8) The total score assigned to the production task;
9) Scores related to four assessment criteria (vocabulary control; grammar accuracy; sociolinguistic appropriateness; and coherence and cohesion).

The total score assigned to the whole exam derives from the sum of the score assigned to the written part and the score assigned to the oral part, and it is associated with a score band (A = "excellent"; B = "good"; C = "passing grade"). The score assigned to the production task is derived from the sum of the scores related to the aforementioned four assessment criteria. Table 3 shows, for each proficiency level, the score ranges with their associated score bands for the whole exam, and the score ranges related to the written part of the exam. The score ranges of the production task for each proficiency level, and the maximum scores related to the four assessment criteria, are shown in Table 4[5].

[5] The total score assigned to the whole exam, to the written part of the exam, to the production task, and to the four assessment criteria were assigned by official raters of CVCL, who were in charge of the certification exams.

**Table 3**. Score ranges and score bands for each CELI exam and proficiency level.

| **CELI exam (proficiency level)** | **Score Range of the whole exam** | **Score band** | **Score range of the written part** |
|---|---|---|---|
| CELI 2 | 138 - 160 | A | 72 - 120 |
| (B1) | 115 - 137 | B | |
| | 94 - 114 | C | |
| CELI 3 | 173 - 200 | A | 84 - 140 |
| (B2) | 144 – 172 | B | |
| | 117 - 143 | C | |
| CELI 4 | 173 - 200 | A | 84 - 140 |
| (C1) | 144 – 172 | B | |
| | 117 - 143 | C | |
| CELI 5 | 173 - 200 | A | 89 - 150 |
| (C2) | 144 – 172 | B | |
| | 117 - 143 | C | |

**Table 4**. Score ranges for the production task and the maximum score pertaining to the four assessment criteria for each proficiency level.

| **Proficiency level** | **Range of the score of the production task** | **Vocabulary control** | **Grammar accuracy** | **Sociolinguistic appropriateness** | **Coherence and cohesion** |
|---|---|---|---|---|---|
| B1 | 12 - 20 | 5 | 5 | 5 | 5 |
| B2 | 12 - 20 | 5 | 5 | 5 | 5 |
| C1 | 18 - 30 | 8 | 8 | 6 | 8 |
| C2 | 21 - 35 | 9 | 8 | 9 | 9 |

Each text is associated with its task assignment. Each writing prompt is reported in the corpus with an identification number which allows to derive the information about the exam session (when the candidate has performed the exam). Further, the task assignment is associated with the other metadata indicated in Table 5.

**Table 5**. Task assignment variables.

| Variables | Value |
|---|---|
| ID_TASK_ASSIGNEMENT | Identification number of the task assignment |
| SESSION | The date of the exam session |
| CEFR | The proficiency level of the language certification exam |
| TOT_SCORE_MAX | The maximum score that can be obtained in the whole exam |
| W_SCORE_MAX | The maximum score that can be obtained in the written part of the exam |
| TASK_SCORE_MAX | The maximum score that can be obtained in the production task |
| LEX_SCORE_MAX | The maximum score that can be assigned to vocabulary control in the production task |
| GRAM_SCORE_MAX | The maximum score that can be assigned to grammatical accuracy in the production task |
| SOCIO_SCORE_MAX | The maximum score that can be assigned to sociolinguistic appropriateness in the production task |
| CC_SCORE_MAX | The maximum score that can be assigned to coherence and cohesion in the production task |
| GENRE | Text genre elicited by the task assignment (letter; e-mail; blog; article; essay) |
| TYPE | Text type elicited by the task assignment (descriptive; narrative; argumentative; mixed) |

### 3.3. Learner variables

For each learner, the metadata about sex (F/M), age and student registration number are reported. Further, candidates performed the language test in different exam centres located not only in Italy, but also elsewhere in Europe and in other countries worldwide. Another variable that should be considered in the design of a learner corpus is the learners' L1 (Tracy-Ventura, Paquot and Myles, 2021). This information cannot be derived from the CELI certification, as candidates are asked to report only their nationality, which does not always reflect the learners' mother tongue (Spina et al., 2022), as in the case of the EFCAMDAT corpus (Murakami & Ellis, 2022). In any case, learners' nationalities were kept as balanced as possible by collecting the same nationalities for each subcorpora and the same number of candidates of a specific nationality for each subcorpora.

In the CELI corpus information about learners' proficiency is provided through different objective indexes: a) the CEFR level of the CELI certificate; b) the score obtained in the whole exam; c) the score band; d) the score obtained in the written part of the exam; and e) the score assigned to the production task. Texts were included in the corpus if learners obtained at least the passing grade in the production task. Furthermore, we included in the corpus only learners that passed the whole exam within a single exam session. Learners that did not obtain the passing grade at the oral part as well as at the written part were excluded from the data collection.

### 3.4. Data collection and transcription criteria

The handwritten exam texts were manually typed and digitised (Spina et al., 2022). Data collection started in February 2020 and ended in February 2021. Texts were reproduced as faithfully as possible. However, learners' errors could complicate the pos-tagging procedure

(see next section). Thus, a manual error correction was carried out according to the target hypothesis (TH), which is the assumed 'correct' form. As Vyatkina (2016) points out, several types of THs are possible, so it should be specified which criteria are adopted. We used the minimal TH or TH1 layer (Reznicek et al., 2013), which usually corrects only spelling and morpho-syntactic mistakes. Specifically, we normalised only learners' spelling errors, such as the unnecessary doubling of letters (1 and 2) or the absence of graphic accents (2), as exemplified below:

1) *[...] il nostro *svillupo è stato sorprendente. *Abiamo scoperto un modo di [...]*

**svillupo* ('progression', 'development') > *svi**l**u**pp**o*; **Abiamo* ('We have') > *A**bb**iamo*

'Our progression was amazing. We discovered a way to [...]'

2) *Ti chiedo *scussa che non sono fatta viva *pero sono stata molto occupata*.

**scussa* ('sorry', 'pardon') > *scu**s**a*; **pero* ('but') > *per**ò***

'I am sorry that I did not get in touch with you but I have been very busy.'

Further, we normalised word forms with spelling errors when the Part-of-Speech (POS) was ambiguous, and the correct POS could be disambiguated taking into account the context. For example, learners frequently produced the verb 'to have' without the grapheme for the unvoiced fricative (see example 3). Given that these forms can be easily confused with conjunctions during automatic tagging procedures, they were corrected (e.g., *è*, 'is' vs. *e*, 'and'; *ho*, 'to have' vs. *o*, 'or').

3) ***o** visitato* ('I visited/have visited') > *ho visitato*

Finally, we normalised phonographematic errors[6], as shown in the following examples:

4) ***di**cisamente* ('definitely') > ***de**cisamente*

[6] We defined phonographematic errors as graphemic errors caused by an incorrect phonological perception of the target word.

5) *ce***re***monia* (‘ceremony’) > *ce***ri***monia*

By contrast, errors ascribable to a possible L1 influence, lexical mistakes, and mismatches in the morpho-syntactic agreements, were left unmodified. All these cases are illustrated in the examples below:

6) *Non si può* ****fare*** *la colpa ai social media*. (Lexical mistake)

‘We cannot blame social media’.

7) *Butto nella plastica* ***la*** *confezion****i*** *di yogurt*. (Agreement mismatch)

‘I throw the packet of yogurt into the plastic’.

8) Una **lenda antiqua** [...] in cui si può vedere la **alma** di Portogallo. (Possible L1 influence)

‘An ancient legend [...] about the soul of Portugal’.

In (6) learner produced a lexical mistake in the Italian collocation *dare la colpa* (‘to blame’) by substituting the typical verb *dare* (‘to give’) with *fare* (‘to do’). Further, the example 7 shows an agreement mismatch between the article *la* (singular) and the noun *confezioni* (plural). Finally, in the example 8 the forms *lenda antiqua* and *alma* have not been normalized as they are probably produced through a transfer from learner’s L1.

### 3.5 Composition of the corpus and its subcorpora

The CELI corpus contains 3,041 texts amounting to 608,614 tokens and 24,698 types. Its subcorpora, one for each proficiency level (B1; B2; C1; C2), present the same design and are balanced with respect to number of tokens (see Table 6).

**Table 6**. Composition of CELI corpus and its subcorpora.

| **Subcorpora** | **texts** | ***tokens*** | ***token average*** | ***types*** | **sentences** | ***token* x sentence** |
|---|---|---|---|---|---|---|
| B1 | 1212 | 156612 | 129.21 | 7397 | 13514 | 11.58 |
| B2 | 840 | 152251 | 181.25 | 9519 | 8438 | 18.04 |
| C1 | 585 | 149859 | 256.16 | 12546 | 7508 | 19.95 |
| C2 | 404 | 149892 | 371.01 | 14153 | 7196 | 20.82 |
| **TOTAL** | **3041** | **608614** | **-** | **-** | **36656** | **-** |

As the four subcorpora are equally designed and balanced in terms of tokens, they can be easily compared in terms of number of learners that have taken the exam in Italy or elsewhere outside Italy, and scores obtained in the different tasks (Table 7).

**Table 7**. The four subcorpora compared by scores.

| **Subcorpora** | **B1** | **B2** | **C1** | **C2** |
|---|---|---|---|---|
| % of the Exam centres abroad | 73% | 79% | 77% | 64% |
| Average of the scores of the whole exam | 124/160 | 157/200 | 154/200 | 153/200 |
| % of the score band A | 26% | 22% | 17% | 16% |
| % of the score band B | 17% | 57% | 56% | 53% |

| | | | | |
|---|---|---|---|---|
| % of the score band C | 57% | 21% | 27% | 31% |
| Average of the written part | 91/120 | 107/140 | 104/140 | 109/150 |
| Average of the production task | 16/20 | 16/20 | 24/30 | 28/35 |

## 4. Pos-tagging of the CELI corpus: procedure and evaluation

Most of the annotation work on learner corpora has traditionally been focused on error tagging (Lüdeling and Hirschmann, 2015; Van Rooy, 2015). In recent decades, the focus has shifted from error tagging to a more "purely linguistic annotation, irrespectively of errors" (Valverde Ibañez, 2011: 214), therefore relying even more extensively on automated annotation tools, such as, among others, part-of-speech tagging.

However, pos-tagging of learner corpora has received limited attention in the literature (Picoral et al., 2021), with a prominent focus on ICLE (de Haan, 2000; Meunier and de Mönnink, 2001; Van Rooy and Schäfer, 2002, 2003), and on other corpora of L2 English (the MACLE: *Malaysian Corpus of Learner English or Spanish*; Aziz and Don, 2019), on corpora of L2 Spanish (the CORANE corpus: *Corpus para el análisis de errores de aprendices de E/LE*; Valverde Ibañez, 2011), of L2 German (the KANDEL corpus: *Kansas Developmental Learner corpus*; Vyaktina, 2016), and of L2 French (the FLLOC corpus: *French Learner Language Oral Corpora*; Marsden et al., 2002). In most of these cases, learner data were processed using taggers, tagsets and training procedures that are commonly used to process corpora of native data (Campillos Llanos, 2016).

Accurate pos-tagging allows more sophisticated corpus queries, in order to investigate more thoroughly learners' interlanguage, and can be followed by other language processing tasks, such as parsing.

### 4.1. Annotation procedure

The pos-tagging of the CELI corpus involved three distinct stages: i. the automatic tagging procedure; ii. a semi-automatic post-editing step, aimed at correcting recurrent tagger errors; iii. a final manual resolution of all the lemmas that were unknown to the tagger.

The 3,041 learner texts included in the CELI corpus were first automatically tokenised, lemmatised, and annotated for POS using *TreeTagger* (Schmid, 1994). In line with what is considered common practice, we opted for a domain transfer solution, consisting in the use of a version of the tagger that was pre-trained on native Italian texts, which had already been used to tag native Italian corpora (Spina, 2014). According to previous studies (de Haan, 2000; Van Rooy and Schäfer 2002, 2003; Vyatkina, 2016), taggers trained on error-free native texts can be used on non-native texts with fairly good results in terms of accuracy. For the benefit of accuracy, the texts included in the CELI corpus underwent a limited normalisation process prior to pos-tagging, which particularly concerned spelling errors such as double consonants instead of single consonants (and vice versa), and few very frequent word pairs that are orthographically similar in Italian and are often confused by learners (cfr. Section 3.4). Abundant evidence (de Haan, 2000; Valverde Ibañez, 2011) indicates that the learner errors mostly affecting the accuracy of the tagger are spelling errors, especially when they involve non-standard forms that correspond to existing words in the target language, as in the examples provided in 3.4. A similar evidence on the relevance of spelling errors was provided for dependency parsing of learner data (Huang, 2018).

The second stage of the pos-tagging process was a semi-automatic editing procedure, which was carried out on specific POS tags with the aim of correcting recurrent tagger errors, revealed by previous analyses on Italian native corpora (Spina, 2014). These post-editing operations involved frequent and grammatically ambiguous forms, such as *come*, *dove*, *che* (‘like’, ‘where’, ‘that’) or verbal forms with incorporated clitic pronouns that are not included in the

lexicon[7], and therefore are not recognised by the tagger (e.g. *spronarsi,* 'to push oneself*'; raccontartene,* 'to tell you something of it'). Through the use of a set of regular expressions searches, this post-editing process allowed us to correct almost 2,800 tagging errors.

In the final stage, we proceeded with a manual resolution of all the lemmas tagged as "unknown" by the tagger. Many of these were non-standard forms produced by learners, which had not been normalised during the data transcription, such as **devano* for *devono* (lemma *dovere,* 'must') in example 9. In this case, we simply replaced the "unknown" label applied by the tagger with the lemma *dovere* (9).

9) *Penso che i giovani* ***devano*** *navigare nelle reti sociali con molta precauzione.*

'I think that young people should browse social networks very carefully.'

devano VER:fin <unknown> --> devano VER:fin dovere

### 4.2. Measuring and evaluating tagger performance

This evaluation process relied on the use of a tagger pre-trained on native Italian data to annotate texts produced by learners. It addressed three specific objectives: i. measuring the performance of *TreeTagger* on L2 Italian texts; ii. analysing the most frequent tagger errors; iii. investigating to what extent and how tagger errors are related to learner errors.

To address these objectives, we randomly selected 24 texts included in the CELI corpus, so that they would meet the following balancing criteria: we extracted one text for each of the six most represented countries (Greece, Spain, Romania, Switzerland, Albania and Germany), for each of the four proficiency levels. The total length of the 24 selected texts was approximately 8,000 tokens, that were manually annotated by two pairs of linguists (the four authors of this paper), so that each pair of annotators would tag 12 texts. According to a well-established

[7] Verbs are an open class in the vocabulary, and therefore texts may contain verbs not included in the lexicon used to train TreeTagger. Similarly, some infrequent verb+clitic forms may not be included in the lexicon. All these forms are not recognised by the tagger, and tagged with the label "unknown".

practice (e.g., Vyatkina, 2016), the two annotators, working separately on the same texts, discussed the cases where there was disagreement in the chosen tags until they reached a shared consensus. Once consensus was reached for the total POS tags, the manually pos-tagged texts were identified as the gold standard, that is the human-produced labels used for comparison against the labels produced by a software (Picoral et al., 2021). To measure the performance of *TreeTagger* and evaluate its accuracy on learner data, this gold standard was used in two distinct evaluations: in the first one, the gold standard was compared to the raw product of the pos-tagging of the same sample of 24 texts, carried out with *TreeTagger*; a second evaluation compared the gold standard to the product of the following, semi-automatic post-editing stage (the second stage of our pos-tagging procedure, as described in section 4.1), performed on the same sample of 24 texts. In both raw and post-edited tagger output evaluations, we identified correct POS tags as the tags where the tagger annotation matched the gold standard, and the incorrect ones as those where this match was not found.

Three measures were used to quantify different aspects of the tagger performance (Picoral et al., 2021): the most basic measure of accuracy, calculated by dividing the number of correct tags by the total number of tags; precision, calculated by dividing the number of tokens correctly assigned to a POS "x" by the total number of tokens tagged as "x"; and recall, calculated by dividing the number of tokens correctly assigned to a pos "x" by the total number of "x" in the data.

Table 8 shows the values of overall accuracy for both the raw and the post-edited annotation. These two accuracy values are compared to the accuracy values obtained from

the evaluation of the TreeTagger performance on Italian native data (Spina, 2014), which adopted the same procedure. As the native Italian corpus was much larger, this evaluation was carried out on a larger sample of approximately 22,000 tokens. The two datasets were, however, symmetrical to those used for the CELI corpus: the first one included the original raw

data, unmodified with respect to the direct product of automatic pos-tagging, and the second one contained the data corrected through the same semi-automatic post-editing procedure used for the CELI corpus.

**Table 8**. The overall accuracy of both the raw and the post-edited annotation.

| | **Accuracy** | |
|---|---|---|
| | **raw sample** | **post-edited sample** |
| CELI (L2 Italian) | 97.2% | 97.7% |
| Perugia corpus (native Italian) | 97.3% | 98.1% |

The data on accuracy reveals two different results. Firstly, the tagger performs in a similar way with native and learner data. The peculiarities of learners' interlanguage, whether errors or other non-standard forms, do not seem to affect the correct automatic identification of grammatical categories. This is true mostly for the raw sample, where accuracy values are almost identical, while for the post-edited sample there is a slight difference, which may suggest a somewhat higher effectiveness of post-editing on native data. This result is in line with previous studies on pos-tagging accuracy on learner data (de Haan, 2000; Meunier and de Mönnink, 2001; Valverde Ibañez, 2011; Van Rooy and Schäfer, 2002), which consistently demonstrated that the learner errors that have the greatest negative impact on tagger performance are spelling errors. As already outlined in sections 3.4 and 4.1, the orthographic errors associated with highly frequent forms were systematically corrected in the CELI corpus at the data transcription stage, thus eliminating the most common source of erroneous annotation, and making learner data more similar to those produced by native speakers.

Secondly, a series of post-editing operations, aimed at correcting a core of recurrent errors identified by previous studies (Spina, 2014), is effective for both native and learner data.

Our analysis of tagger performance then focused on the POS tags that resulted as the most challenging for the tagger. Table 9 shows the eight POS tags with the lowest precision values in the two evaluations.

**Table 9**. The eight POS tags with lowest precision values in both raw and post-edited samples, with their respective recall values.

| **POS** | **Precision** | | **Recall** | |
|---|---|---|---|---|
| | **raw sample** | **post-edited sample** | **raw sample** | **post-edited sample** |
| QST interrogative adverb | 57.1% | 100% | 70% | 100% |
| INT interjection | 76.9% | 84.6% | 81.2% | 86,7% |
| AUX:ppast past-participle of an auxiliary verb | 80% | 80% | 83.3% | 83.3% |
| SUB subordinator | 86.4% | 85.9% | 88% | 87.6% |
| RELA relative pronoun | 90% | 97.6% | 90.91% | 97.7% |
| DET:indef indefinite determiner | 90.2% | 90.2% | 91% | 91% |
| ADJ adjective | 90.4% | 90.6% | 91.2% | 91.4% |
| PRO:indef indefinite pronoun | 93.5% | 93.5% | 93.9% | 93.9% |

The POS tag assigned by the tagger which returned the highest number of errors is that of the interrogative adverb (QST), which has a precision value of 57% in the raw sample. This tag is also problematic in data produced by native Italians (Spina, 2014), as it mostly involves grammatically ambiguous forms (*quanto,* 'how much'; *quando*, 'when'; *dove*, 'where'; *quale*, 'which'), which can function as interrogative adverbs, or pronouns, subordinators or relative pronouns. (10) is an example of wrong attribution of the pos QST to a subordinator (*quante*). In this case, moreover, the semi-automatic post-editing operations were able to correct all the errors made by the tagger in the raw sample, reaching accuracy and recall values of 100% in the sample.

10) *Non esistono statistiche per sapere **quante** persone cambiano radicalmente di lavoro.* quante --> *QST --> SUB

'There are no statistics showing how many people change jobs completely.'

A similar case is the opposite, the wrong attribution of the tag SUB (subordinator) to an interrogative adverb (QST), as in (11):

11) *Ciao Marco, **come** stai?* → ***~~SUB~~** → **QST**

'Hi Marco, how are you?'

These tagger errors are therefore more due to the inherent ambiguity of the forms, rather than to learners' interlanguage errors.

Interjections also had a relatively low precision value in the raw sample (76.9%), which reached 84.6% after the post-editing operations. Again, tagger errors do not appear to be due to learner errors, but to ambiguities in the forms that the tagger fails to resolve, as in the case of (12), where *grazie* ('thanks to') is labelled as an interjection, while it is a noun.

12) *Però **grazie** ai miei amici sono riuscita a superare tutto questo dramma.* → ***INT** --> NOUN

‘But thanks to my friends I managed to get through all this tragedy.’

The other POS tags that were most often wrongly attributed (past participles of auxiliary verbs, subordinators, relative and indefinite pronouns, adjective and indefinite determiners; cfr. Table 8) range from 80% to 93.5% of precision. With regard to the effectiveness of the post-editing operations performed after pos-tagging, the comparison of respective precision and recall values highlights three possible scenarios. In most cases, post-editing increased - sometimes in a highly significant manner, as in the case of interrogative adverbs and interjections (examples above, inserire numeri) - the accuracy of pos-tagging, by removing many of the tagger errors. In a few cases, post-editing had no effect on accuracy, as precision and recall values remained unchanged. This happened for example with indefinite pronouns (13) :

13) *Conosco tantissime parole e ti posso trovare mille significati [...].* → ***PRO:indef --> DET: indef**

‘I know so many words and I can find a thousand meanings for you [...].

In order to analyse more closely the most common types of tagger errors in the annotation of learner data, Table 10 shows in a reduced form the complete matrix of the number of errors per POS. The POSs involved in the most frequent tagger errors are seven (adjective, adverb, noun, finite mood verb, finite mood auxiliary verb, preposition, past participle). The errors that occur more frequently are the tagging of an adjective as a past participle (frequency in the raw sample = 22) (14), with its opposite (a past participle as an adjective: frequency = 4) (15), and a noun as an adjective (frequency = 16) (16), with its opposite (an adjective as a noun: frequency = 6) (17).

(14) *Ma i suoi genitori non erano **convinti** della sua scelta.* → ***VER:ppast** (instead of ADJ)

‘Yet her/his parents were not certain about her/his choice.’

(15) *Mezz’ora dopo essermi **sdraiata** [...]* → ***ADJ** (instead of VER:ppast)

‘Half an hour after I had been lying down [...].’

(16) *Parenti, amici, **vicini**, tutti abbiamo almeno una conoscenza che [...].* → ***ADJ** (instead of NOUN)

‘Relatives, friends, neighbours, we all have at least an acquaintance that [...].’

(17) *E se ci si sente soli, non è perché siamo **soli**, ma [...].* → ***NOUN** (instead of ADJ)

‘And if you feel alone, it is not because you are alone, but [...].’

Errors between noun/adjective and adjective/past participle are also very common in the pos-tagging of texts produced by native Italian speakers (Spina, 2014): the contexts in which the two pairs of grammatical categories occur are, in fact, very similar, and this makes the tagger's task more complex. Another error that occurs frequently is the tagging of prepositions as adverbs (frequency = 13) (18).

(18) *[...] nessuno ci pensa due volte **prima** di scrivere [...].* → ***ADV** (instead of PRE)

‘[...] nobody thinks twice before writing [...]’

Again, the similarity of the contexts in which the two POSs occur also apply in this case. However, given that the two POSs are either closed categories (prepositions) or categories including a limited number of forms (adverbs), the post-editing phase was effective and led to a reduction of errors by 61%.

This data confirms what has already been shown in the previous paragraphs: there are no substantial differences in the tagger accuracy with data from native Italians and learners, and a post-editing phase aimed at specific recurrent pos errors is able to improve the tagger performance.

**Table 10**. The number of the most frequent tagger errors in the raw sample (in parentheses, the number of errors in the post-edited sample, where there are differences). Row 1 indicates the POS assigned by the tagger, column 1 the correct POS.

| | **adjective** | **adverb** | **noun** | **finite mood verb** | **finite mood auxiliary verb** | **preposition** | **past participle** |
|---|---|---|---|---|---|---|---|
| adjective | - | 6 | 6 | 3 (2) | 0 | 0 | 22 |
| adverb | 4 (2) | - | 0 | 5 (1) | 0 | 3 | 0 |
| noun | 16 | 5 | - | 5 | 0 | 0 | 2 |
| finite mood verb | 0 | 0 | 5 | - | 10 | 0 | 2 |
| finite mood auxiliary verb | 0 | 0 | 0 | 1 | - | 0 | 0 |
| preposition | 0 | 13 (5) | 0 | 0 | 0 | - | 0 |
| past participle | 4 (5) | 0 | 1 (0) | 1 | 0 | 0 | - |

### 4.3. The Impact of Learner Language on POS-tagging performance

By performing a more thorough analysis of the POS errors, we were able to verify that only a limited number of tagging errors actually coincide with learner errors. In particular, incorrect tagging usually occurs when learners' erroneous forms turn out to be homographs with other common Italian words.

One of the most common learner errors types which affected the automatic POS tagging can be identified as typographic. For example, in (19), *giungo,* which is the first person singular of the present tense of the verb *giungere* ('to arrive'), was employed by the learner instead of the noun *giugno* ('June'), probably due to confusion caused by the closeness in their spelling. In this particular case, *TreeTagger* assigned to this instance of non-standard language a *VER:fin tag (*giungo*) instead of a NOUN tag (*giugno*), leading to a tagging non-compliant with the target hypothesis, if compared with the manual gold standard annotation of the sample:

19) *Durante la cerimonia, organizzata il 2* ***giungo*** *nella Facoltà di [...].* → ***VER:fin**

'During the ceremony, which was organised on the 2nd of June at the Faculty of [...].'

Another type of learner errors leading to POS tagging non-compliant with the target hypothesis is represented by morphological errors, as shown in (20). Here, *interesso,* which is the first person of the present tense of the verb *interessare* ('to interest'), was used in place of the noun *interesse* ('interest')[8], thus resulting in a *VER:fin tag (*interesso*) in place of a NOUN tag (*interesse*).

Similarly, in (21) the feminine plural adjective *deserte* ('desolate', 'deserted') was employed instead of the masculine plural noun *deserti* ('deserts'), causing an *ADJ (*deserte*) in place of NOUN (*deserti*) tagging not compliant with the target hypothesis:

[8] In this case, the use of *interesso* in place of *interesse* may be interpreted as a product of an overregularisation process, since masculine singular nouns in Italian tend to present a more common *-o* ending.

20) *Penso che questo articolo è di importanza vitale per l'**interesso** dei vostri lettori.* →***VER:fin**

'I think that this article is vital for your readers' interests.'

21) *[...] una splendida natura: mari (oceani), laghi, boschi, montagne, **deserte** [...].* → ***ADJ**

'[...] wonderful nature: seas (oceans), lakes, woods, mountains, deserts [...].'

Furthermore, we could also observe some lexical errors, such as in (22), where *conosciuto* ('known'), which is the past participle of the verb *conoscere* ('to know'), was used inappropriately instead of the noun *conoscente* ('acquaintance'), producing a *VER:ppast (*conosciuto*) in lieu of NOUN (*conoscente*) non-compliant tagging.

Similarly, in (23), the feminine indefinite pronoun *qualcuna* ('somebody') was used instead of the determiner *qualche* ('some'), returning a *PRO:indef (*qualcuna*) instead of a DET:indef (*qualche*) non-compliant tagging.

22) *[...] che ci fanno vedere chi è un vero amico e chi è solo un **conosciuto**.* → ***VER:ppast**

'[...] that make us see who is a real friend and who is just an acquaintance.'

23) *[...] oppure sussurrare mentre sta passando **qualcuna** ragazza.* → ***PRO:indef**

'[...] or whispering while some girls are passing by.'

Nevertheless, there are learner errors which do not affect the pos-tagging process. These learner errors "have effective information that helps determine the POSs" (Mizumoto and Nagata, 2017: 55). For instance, in our sample we found the sentence in (24), where the misspelt word **divettando* (corr. *diventando,* 'becoming'), while producing an unknown lemma, was correctly tagged as VER:ger, as it presents the typical characteristics of the gerund form of first-conjugation verbs, i.e. the ending in *-ando*.

In (25), *spero*, which is the first person singular of the present tense of *sperare* ('to hope') was used inappropriately instead of the verb *aspettare* or *attendere* ('to wait for'), probably due to L1 influence (cf. Spanish *esperar*, 'to wait for'/'to hope'), producing anyhow the correct POS tag, VER:fin.

24) *Da piccola, mi popolavano i sogni gli eroi dei libri,* ***divettando*** *anche i miei eroi personali.* → **VER:geru**

'When I was a child, my dreams were filled up with heroes from books, who ended up becoming my personal heroes too.'

25) ***Spero*** *con ansia la tua risposta.* → **VER:fin**

'I'm looking forward to your reply.'

Although our sample is very small, we found that learner errors which did affect the automatic pos-tagging represent 6% of the total POS tag errors in the raw sample and 5% in the post-edited sample[9]. Furthermore, we were also able to spot differences in terms of learners' rate of errors affecting the tagger accuracy at different proficiency levels. As expected - even though we need to take these findings cautiously with such a limited sample size -, a computation of errors on a 1,000 token basis shows that the learner errors actually affecting the pos-tagging are 2.3 for the B1 texts, 1.8 for the B2, 1.4 for the C1, and 0.7 for the C2 texts in our sample. As we hypothesised above, these are learner errors that led to actual tagging errors, which were not detected even in the further post-tagging phase, mainly because they often involve forms belonging to open grammatical categories, such as verbs, nouns and adjectives, for which a semi-automatic correction of tagger errors cannot be envisaged.

[9] There are a few dubious cases, which were not counted here.

## 5. Conclusions

Design and annotation criteria are key issues in learner corpus design. As for the former, while still not receiving the attention it deserves, rigorous proficiency level attribution in learner corpora allows reliable comparability between different learner corpora so as to inform a sound discussion of empirical findings within the broader domain of second language acquisition research. A corpus such as the CELI corpus goes in this direction with a design including balanced subcorpora of written texts produced in a language certification context, and with reference to a language other than English, i.e., Italian.

The CELI corpus also contributes to learner corpus design from the perspective of the annotation criteria adopted. The annotation procedure involved an automatic pos-tagging, followed by a semi-automatic post-editing step to correct frequent tagger errors on grammatically ambiguous forms, and a final manual resolution of the lemmas which the tagger did not recognise. The effort produced to make the pos-tagging as effective as possible seems to have been worthwhile: an evaluation of the tagger's performance revealed that its accuracy on learner data is comparable to that on data produced by native Italian speakers. Data on accuracy also suggested that the post-editing procedure resulted in a further improvement in annotation accuracy, by removing a small number of recurrent tagger errors.

All in all, the CELI corpus introduces new ways to analyse the acquisition of Italian L2 from an empirical perspective, with the advantages deriving from a pseudo-longitudinal perspective, while relying on solid annotation procedures. It is hoped that many studies can stem from it, thus helping us expand our knowledge of Italian L2 acquisition dynamics and, more generally, of the multiple affordances that learner corpora entail in the different domains of second language teaching and learning.

**References**

Alderson JC (2007) The CEFR and the Need for More Research. *The Modern Language Journal* 91: 659–663.

Alexopoulou T, Michel M, Murakami A, and Meurers D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. Language Learning, 67, 180–208.

Aziz RA and Don ZM (2019) Tagging L2 Writing: Learner Errors and the Performance of an Automated Part-of-Speech Tagger. *Gema Online Journal of Language Studies* 19(3): 140–155. https://doi.org/10.17576/gema-2019-1903-09

Bailini S and Frigerio A (2018) CORESPI e CORITE, due nuovi strumenti per l'analisi dell'interlingua di lingue affini. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos* 5(2): 313-319.

Bratankova L (2015) *Le collocazioni Verbo + Nome in apprendenti di italiano L2*. Unpublished PhD thesis, University for Foreigners of Perugia, Italy.

Campillos Llanos L (2016) PoS-tagging a Spanish oral learner corpus: Criteria, procedure, and a sample analysis. In Alonso-Ramos M (ed) *Spanish Learner Corpus Research: Current trends and future perspectives*. Studies in Corpus Linguistics 78: 89–116. Amsterdam: John Benjamins Publishing Company.

Carlsen C (2012) Proficiency Level—a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics* 33(2): 161–183.

Council of Europe (ed) (2001) *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Corino E, Colombo S and Marello C (2017) *Italiano di stranieri: I corpora VALICO e VINCA*. Perugia: Guerra.

De Haan P (2000) Tagging non-native English with the TOSCA–ICLE tagger. In Mair Ch. and Hundt M (eds) *Corpus linguistics and linguistic theory*. Amsterdam: Rodopi, pp. 69–79.

Dulay H, Burt M and Krashen S (1982) *Language Two*. Oxford: Oxford University Press.

Gass S and Selinker L (2008) *Second Language Acquisition. An Introductory Course*. New York: Routledge.

Geertzen J, Alexopoulou T and Korhonen A (2014) Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). *Selected Proceedings of the 31st Second Language Research Forum (SLRF)*. MA: Cascadilla Press.

Gilquin G (2015) From design to collection of learner corpora. In Granger S, Gilquin G and Meunier F (eds) *The Cambridge handbook of learner corpus research.* Cambridge: Cambridge University Press, pp. 9–34.

Glaznieks A, Frey JC, Stopfner M, Zanasi L and Nicolas L (2022) Leonide: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research* 8(1): 97–120.

Glaznieks A, Frey JC, Nicolas L, Abel A and Vettori C (in preparation) The Kolipsi Corpus Family. A collection of Italian and German L2 learner texts from secondary school pupils.

Granger S (1996) From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Aijmer K, Altenberg B and Johansson M (eds) *Languages in Contrast*. Lund University Press: Lund, pp. 37–51.

Granger S (ed) (1998) *Learner English on Computer*. London: Longman.

Granger S (2015) Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1): 7–24.

Granger S, Dupont M, Meunier F, Naets H and Paquot M (2020) *The International Corpus of Learner English*. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger S, Gilquin G and Meunier F (eds) (2015) *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Granger S, Hung J and Petch-Tyson (eds) (2002) *Computer learner corpora, second language acquisition, and foreign language teaching*. Amsterdam: John Benjamins Publishing Company.

Grego Bolli G (2004) Measuring and evaluating the competence in Italian as a foreign language. In: *Studies in Language Testing, 18: European Language Texting in a Global Context* (eds M Milanovic and C Weir). Proceedings of the ALTE Barcelona Conference, July 2001, pp. 271–283. Cambridge (UK): Cambridge University Press.

Guiraud P (1954). *Les caractères statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Harsch C (2014) General Language Proficiency Revisited: Current and Future Issues. *Language Assessment Quarterly* 11(2): 152–169.

Harsch C and Malone ME (2021) Language Proficiency Frameworks and Scales. In Winke P Brunfaut T (eds) *The Routledge Handbook of Second Language Acquisition and Language Testing*, London-New York: Routledge.

Huang Y, Murakami A, Alexopoulou T and Korhonen A (2018) Dependency parsing of learner English. I*nternational Journal of Corpus Linguistics* 23(1): 28–54.

Hulstijn JH (2007) The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal* 91: 663–667.

Hulstijn JH, Alderson JC and Schoonen R (2010) Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In Bartning I, Martin M and Vedder I (eds) *Eurosla Monograph series 1. Communicative*

*proficiency and linguistic development: intersections between SLA and language testing research*. European Second Language Association, pp. 11–20.

Larsen-Freeman D and Cameron L (2009) *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Lozano C (2021) CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*.

Lüdeling A and Hirschmann H (2015) Error annotation systems. In Granger S, Gilquin G and Meunier F (eds) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, pp. 135–158.

Marsden E, Myles F, Rule S & Mitchell R (2002). Oral French Interlanguage Corpora: Tools for Data Management and Analysis. Centre for Language in Education Occasional Papers no. 58. University of Southampton.

Meurers D (2015) Learner corpora and natural language processing. In Granger S, Gilquin G and Meunier F (eds) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, pp. 537–566.

Myles F (2005) Interlanguage corpora and second language acquisition research. *Second Language Research* 21(4): 373–391.

Huang Y, Murakami A, Alexopoulou T and Korhonen A (2018) Dependency parsing of learner English. *International Journal of Corpus Linguistics* 23(1): 28–54.

Meunier F (2015) Developmental patterns in learner corpora. In Granger S, Gilquin G, Meunier F (eds) *The Cambridge Handbook of Learner Corpus Research,* Cambridge: Cambridge University Press, pp. 379-400

Meunier F and De Mönnink I (2001) Assessing the success rate of EFL learner corpus tagging. In De Cock S, Gilquin G and Granger S (eds) *Future Challenges for Corpus Linguistics*. Proceedings of the 22nd ICAME Conference, pp. 59–60.

Mizumoto T and Nagata R (2017) Analyzing the Impact of Spelling Errors on POS-Tagging and Chunking in Learner English. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, Taipei, Taiwan, pp. 54–58. Asian Federation of Natural Language Processing.

Murakami A and Alexopoulou T (2016) L1 Influence on the Acquisition Order of English Grammatical Morphemes. *Studies in Second Language Acquisition* 38(3): 365–401.

Murakami A and Ellis N C (2022) Effects of Availability, Contingency, and Formulaicity on the Accuracy of English Grammatical Morphemes in Second Language Writing. *Language Learning*, 1-42.

Paquot M and Le Bruyn B (eds.) (2021). Learner corpus research meets second language acquisition. Cambridge: Cambridge University Press.

Picoral A, Staples S and Reppen R (2021) Automated annotation of learner English. *International Journal of Learner Corpus Research* 7(1): 17–52.

Reznicek M, Lüdeling A and Hirschmann H (2013) Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In Díaz-Negrillo N, Ballier N and Thompson P (eds) *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, pp. 101–124. doi: 10.1075/scl.59.07rez

Schmid H (1994) *Probabilistic part-of-speech tagging using decision trees*. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK.

Spina S (2014) Il Perugia Corpus: Una risorsa di riferimento per l'italiano: Composizione, annotazione e valutazione. In: Basili R, Lenci A, and Magnini B (eds) Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014, volume 1. Pisa: Pisa University Press, pp. 354–59.

Spina S (forthcoming) Task effects on phraseological complexity in learners' written and oral production. In: Ackerley K, and Castello E (eds) Continuing Learner Corpus Research: Challenges and Opportunities, Corpora and language in use series, Presses Universitaires de Louvain.

Spina S and Siyanova-Chanturia A (2018) The longitudinal corpus of Chinese learners of Italian (LOCCLI). In: Poster presented at the 13th Teaching and Language Corpora conference, University of Cambridge, UK.

Spina S, Fioravanti I, Forti L, Santucci V, Scerra A, and Zanda F. (2022) Il corpus CELI: Una nuova risorsa per studiare l'acquisizione dell'italiano L2. Italiano LinguaDue 1: 116–38.

Tono Y (2003) *Learner corpora: Design, development and applications*. Paper Presented at the Corpus Linguistics 2003 Conference (CL 2003) Lancaster.

Tracy-Ventura N and Paquot M (2021). *The Routledge Handbook of Second Language Acquisition and Corpora*. Abingdon: Routledge.

Valverde Ibañez MP (2011) An Evaluation of Part of Speech Tagging on Written Second Language Spanish. In Gelbukh AF (ed) *Computational Linguistics and Intelligent Text Processing*. Proceedings, Part I, pp. 214–226, Springer Berlin Heidelberg.

Van Rooy, B. (2015). Annotating learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics, pp. 79-106). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.005

Van Rooy B and Schäfer L (2002) The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies* 20: 325–335.

Van Rooy B and Schäfer L (2003) An Evaluation of Three POS Taggers for the Tagging of the Tswana Learner English Corpus. In Archer D, Rayson P, Wilson A and McEnery T (eds)

*Proceedings of the Corpus Linguistics 2003 conference Lancaster University (UK) vol. 16*, pp. 835-844, University Centre for Computer Corpus Research on Language Technical Papers.

Vyatkina N (2016) The Kansas Developmental Learner corpus (KANDEL): A developmental corpus of learner German. *International Journal of Learner Corpus Research* 2(1): 101–119.

Wisniewski K (2017) Empirical Learner Language and the Levels of the Common European Framework of Reference. *Language Learning* 67(S1): 232–253.