Collecting, analysing and using longitudinal learner data for language teaching: the case of LONGDALE-IT

Erik Castello

Abstract. This study aims to investigate the effectiveness of Data-Driven Learning (DDL) teaching materials based on learner corpus data. The data analysed is made up of texts written by a group of Italian university students and collected as part of the Italian component of the Longitudinal Database of Learner English (LONGDALE) project: LONGDALE-IT. All the students participating in the project wrote two texts as a response to the same prompt, one in the first and the other in the second year at university. The data collected in the first year was used to create some DDL materials, which were administered to the students in the second year with the aim of helping them improve on various aspects of their writing, including their use of \textit{it}-extraposition. In this paper, the first-year learner data will be compared to the second-year learner data, and subsequently to a sub-section of the Louvain Corpus of Native English Essays (LOCNESS). Quantitative and qualitative findings concerning the use of \textit{it}-extraposition in the learner texts will be discussed, with a view to determining the impact of DDL teaching materials on the learning process.

Keywords: learner corpora, DDL teaching materials, \textit{it}-extraposition constructions, LONGDALE project.

1. Introduction

In the last few decades there has been a burgeoning interest in the use of learner corpora, i.e. electronic collections of authentic, continuous and contextualised foreign or second language texts (Granger, 2009, p. 14), for language learning

1. University of Padua, Italy; erik.castello@unipd.it

Collecting, analysing and using longitudinal learner data for language teaching...

and teaching. While a large number of learner corpus-based studies have already identified a variety of linguistic features of learner language (e.g. lexical, grammatical, phraseological), more studies are needed on the use of this data to inform pedagogical practice (e.g. Cotos, 2014). In the literature, the term Data-Driven Learning (DDL) is used to refer to the use of either native or learner corpora in language learning/teaching (e.g. Chamber, 2010; Gilquin & Granger, 2010), and Learning-Driven Data (LDD) is the expression proposed by Seidlhofer (2002) to indicate data collected from specific teaching contexts which are investigated with the aim of “understand[ing] local conditions of relevance” and ultimately devising relevant classroom applications (pp. 213-214). Studies on DDL and LDD have discussed various issues, including how to appropriately combine native and learner data in DDL to fit the local needs, and what the best choice is between autonomous, hands-on, partly serendipitous corpus exploration and the use of paper-based, “hands-off” materials prepared by the teacher (e.g. Boulton, 2012).

This study draws on research conducted on the Italian component of the Longitudinal Database of Learner English (LONGDALE), an international project which was launched by the University of Louvain², Belgium, in 2008. The aim of the project is to collect a large longitudinal database consisting of data from learners with various mother tongue backgrounds who are followed over a period of at least three years. To date, the project has involved the collection of various data types, including argumentative essays, narratives, and informal interviews. The database also contains comprehensive information about the learners.

This paper aims to explore the effects of DDL teaching materials based on LDD administered to a group of Italian university students over two years. The main focus is on the use of it-extrapolation in the students’ written production collected as part of the Italian component of the LONGDALE project (LONGDALE-IT).

2. Data and method

2.1. The learner population and the data

The data consists in a longitudinal corpus of texts produced by 138 Italian undergraduate students of Linguistic and Cultural Mediation at the University of Padua (Italy), who attended the first-year English language course in 2013-2014 and the second-year course in 2014-2015. At the end of both courses the students were asked to write a text in response to the same prompt, i.e. about their views

on standard and non-standard English, native-likeness and their expectations for their level of English by the end of the course. The texts, written under no time constraints and of about 450-500 words in length, were gathered through LimeSurvey. During the second-year course the learners were exposed to both hands-on and hands-off DDL teaching materials which were based on first-year data and compared to a sub-section of the Louvain Corpus of Native English Essays (LOCNESS), a collection of essays written by British and American university students. Not only were the DDL materials used to illustrate and contextualize contrastive lexico-grammatical phenomena, such as the structure of English noun phrases, *it*-extraposition constructions and conjunctive adjuncts, but also to give the students general feedback and to make them think critically about what they had already achieved and what remained to be improved. Generally speaking, as regards productive skills, the target CEFR levels of the first- and second-year courses were B1 and B2 respectively. *Table 1* shows quantitative data about the corpora used for the study: the first-year corpus (LONGD_pd1), the second-year corpus (LONGD_pd2), and the LOCNESS sub-corpus.

<table>
<thead>
<tr>
<th></th>
<th>LONGD_pd1</th>
<th>LONGD_pd2</th>
<th>LOCNESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>tokens</td>
<td>43392</td>
<td>64494</td>
<td>35399</td>
</tr>
<tr>
<td>types</td>
<td>2324</td>
<td>3307</td>
<td>4547</td>
</tr>
<tr>
<td>n. Sentences</td>
<td>1585</td>
<td>2574</td>
<td>1698</td>
</tr>
<tr>
<td>words per sentence</td>
<td>27,38</td>
<td>25</td>
<td>20,8</td>
</tr>
<tr>
<td>sentences per text (average)</td>
<td>11,4</td>
<td>18,6</td>
<td>21,7</td>
</tr>
<tr>
<td>average word number per text</td>
<td>314,4</td>
<td>467,3</td>
<td>453,8</td>
</tr>
<tr>
<td>n. of texts</td>
<td>138</td>
<td>138</td>
<td>78</td>
</tr>
</tbody>
</table>

2.2. Methodology for studying *it*-extraposition

The methodology used to conduct the study is Contrastive Interlanguage Analysis (CIA), whereby the learner language represented in LONGD_pd1 and LONGD_pd2 “is analyzed in its own right […] longitudinally” (Granger, 2009, p. 18) and then compared to native data. The aspect that was investigated is the use of *it*-extraposition constructions, that is, clauses containing a formal or anticipatory subject as well as a notional subject, which takes the form of an extraposed embedded clause (Kaltenböck, 2003, p. 236). The types of clauses that can be

---

extraposed are that-clauses (e.g. it could be argued that ...), wh-clauses (e.g. it is debatable whether ...), to-clauses (e.g. it would be interesting to ...), and for/to-clauses (e.g. it is really hard for a learner to achieve them). By contrast, extraposed ing-clauses and noun phrases have a borderline status (Kaltenböck, 2003, p. 247), and are usually not acceptable in formal writing. It-extraposition constructions, which function to “depersonalize text and create an impression of [...] objectivity”, are most common in academic prose, and are syntactically complex, which is why they can present an area of difficulty for L2 writers (Hinkel, 2013, p. 10). Speakers of Italian, in particular, are likely to encounter additional difficulties, in that Italian has no counterpart to it-clauses.

The analysis involved the study of all the instances of it-extraposition constructions in the corpora, which were retrieved by means of The Sketch Engine⁵, an online corpus query system. The frequencies of occurrence of various phenomena related to it-extraposition were compared using the Log-Likelihood statistic (LL)⁶, which indicates whether and to what extent differences are statistically significant.

3. Results and discussion

Over the two years, an increase occurred in the overall use of it-extraposition constructions (respectively 0.48% and 0.64% of the number of tokens), which is likely to be due to the “awareness raising” effect of the DDL activities administered to the students during the second-year course. It must be pointed out, however, that in both years the students produced more it-extraposition constructions than the native speakers (0.25%), which confirms the results of other studies of learner language. Figure 1 represents the distribution of extraposed embedded clauses in the three corpora. The figures are given per total number of it-clauses.

As can be seen from Figure 1, in the second year, the learners produced slightly fewer to-clauses and opted for more that-clauses (LL: +32.59, p<0.05), which brings them closer to the native students. It can also be noted that the number of ing-clauses decreased over the two years (LL: - 10.54, p<0.01), which suggests that most of the learners became aware of the fact that extraposed ing-clauses are not appropriate for academic writing. It should then be noticed that in the second year, the learners extraposed a larger number of noun phrases than in the first year (LL: +2.82, p>0.05), thus producing more instances of erroneous constructions, such as “*it is necessary a real contact with native-speakers to apply our knowledge”.

---

⁵. http://www.sketchengine.co.uk
⁶. http://ucrel.lancs.ac.uk/llwizard.html
Though the DDL materials specifically addressed this aspect, which is notoriously challenging for Italian learners, it clearly remained difficult for the learners, whose production revealed persistent L1 interference. The use of *it*-extraposition in combination with other features of academic English (e.g. passive voice, complex noun phrases, intricate sentences) also proved challenging.

Figure 1. Distribution of extraposed embedded clauses in LONGD_pd1, LONGD_pd2, and in the LOCNESS sub-corpus

By contrast, other types of mistakes concerning the use of *it*-extraposition decreased significantly over the two years. The decrease in the omission of anticipatory *it* (e.g. “*I think is extremely difficult to be able to talk like a native speaker*”), in particular, was highly significant (LL: -16.27, \( p < 0.0001 \)).

4. **Conclusions**

The main purpose of this study was to test the effectiveness of DDL materials based on LDD data, and ascertain whether a group of Italian university students who were exposed to them improved their use of *it*-extraposition over two years. The results of the analysis have shown that the texts they wrote in the second year are generally more accurate, and suggest that the students’ use of *it*-extraposition is evolving towards the native speaker norm. The main exception to this trend is the increase in the number of extraposed noun phrases, which highlights that
this is a persistently challenging aspect for Italian learners. The study thus hints that exposing the learners to concordance lines containing their own mistakes and making them reflect on them does not necessarily bring about their complete eradication. This could depend on various factors, including the quality of the DDL materials and the way the students approached them. In order to investigate these aspects further, an analysis could be conducted into the features of the DDL activities, as well as into the students’ perception of them. Besides *it*-extraposition, other features of the texts could be investigated, such as the use of the passive voice, the complexity of noun phrases and the intricacy of sentences.

References


