Dictionaries and encoding examples to support language production

Ana Frankenberg-Garcia
Centre for Translation Studies, School of English and Languages
University of Surrey

Learners’ dictionaries can help users not just with meanings, but also with collocations, syntax and common errors. Although there are several ways in which information for language production can be presented, previous research has shown that learners tend to prefer to obtain it from examples. However, early studies on the effects of dictionary examples did not disclose much evidence of their value. This could be because dictionaries do not distinguish between examples to reinforce comprehension and examples to support production, and also because learners may need more than one example to figure out how words are used. Following recent findings by Frankenberg-Garcia (2012, 2014), where encoding examples appeared to help language production, the present study aimed to further investigate the effects of such examples by refining previous elicitation procedures, by using a different and larger population sample, and by testing whether learners react better to an optimum number of examples.

Keywords: learners’ dictionaries, dictionary examples, language production, corpora

1. Introduction
Dictionaries have traditionally been used to mainly look up what words mean or how they are spelled. However, many of today’s dictionaries provide additional information that goes beyond spellings and definitions (or translations, in the case of bilingual dictionaries). English language pedagogical lexicography is particularly advanced in this respect, with state-of-the-art dictionaries containing linguistic data gleaned from corpora that can help learners choose appropriate collocations, use the correct syntax and even avoid common mistakes.

1.1 Encoding information in dictionaries
There are various ways in which dictionaries summarize information that can help users with language production. This includes information on word grammar, which, as exemplified in table 1, can be conveyed in a variety of ways. As can be seen in the table, a combination of more than one method is often used. Despite advances such as these on the part of dictionary makers, users in general are not very good at utilizing dictionaries to their full potential, especially with regard to looking up the sort of grammatical information illustrated in table 1. An experiment involving 211 undergraduates studying for non-language related degrees at a Portuguese university carried out by Frankenberg-Garcia (2011) showed that only a small minority would turn to a dictionary to look up what prepositions to use after intransitive verbs. In an earlier study, Béjoint (1981) carried out a questionnaire to find out how a group of 122 French learners of English used dictionaries and found that less than half the respondents said they referred to the syntactic coding information in dictionaries. Even undergraduates studying languages, whom one would expect to be more aware of syntactic constraints, do not seem to fare any better. Herbst and Stein (1987) reported that many German university students of English and even some teachers did not know that dictionaries contained information on the syntactic patterns of words.

In an empirical study conducted some years later, Bogaards and van der Kloot (2002) observed how a group of 117 Dutch learners of English used information about verb constructions in English dictionaries by asking the students to underline the information that had been useful to them in a translation into English task. The study showed that grammar codes were rarely used, that explicitly-worded grammar patterns were important to university students of English but not to high-school students in general, and that definitions that embed information about syntax and collocation like those used in the COBUILD dictionary (see Hanks 1987) were used for looking up meanings, but not so much as an aid to writing. On the other hand, the most widely used source of grammar information for language production purposes were the example sentences supplied in an entry.

Table 1. Different ways of presenting information on word grammar in dictionaries

<table>
<thead>
<tr>
<th>Example for approve in the sense of ‘to think that sth/sb is good or suitable’</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic codes</td>
<td>V + of</td>
</tr>
<tr>
<td>Syntactic information spelled out in words</td>
<td>approve of</td>
</tr>
<tr>
<td></td>
<td>approve of somebody/something</td>
</tr>
<tr>
<td></td>
<td>approve of somebody doing something</td>
</tr>
<tr>
<td></td>
<td>approve of somebody’s doing something</td>
</tr>
<tr>
<td></td>
<td>approve of</td>
</tr>
<tr>
<td></td>
<td>approve of someone doing something</td>
</tr>
<tr>
<td>Syntactic patterns embedded in definitions</td>
<td>If you approve of an action, event, or suggestion, you like it or are pleased about it; If you approve of someone or something, you like and admire them.</td>
</tr>
<tr>
<td>Explanations alerting to common errors (plus examples)</td>
<td>When approve means ‘to have a positive feeling towards something or someone’, it is followed by of, not by a direct object. ✓ I do not approve of violence. ✗ I do not approve of violence.</td>
</tr>
<tr>
<td>Examples</td>
<td>He seemed to approve of my choice.</td>
</tr>
<tr>
<td></td>
<td>I don’t really approve of children wearing make-up.</td>
</tr>
<tr>
<td></td>
<td>She doesn’t approve of my friends.</td>
</tr>
<tr>
<td></td>
<td>He doesn’t approve of smoking.</td>
</tr>
<tr>
<td></td>
<td>I thoroughly approve of what the government is doing.</td>
</tr>
<tr>
<td></td>
<td>Do you approve of my idea?</td>
</tr>
<tr>
<td></td>
<td>She doesn’t approve of me leaving school this year.</td>
</tr>
<tr>
<td></td>
<td>She doesn’t approve of my leaving school this year.</td>
</tr>
<tr>
<td></td>
<td>Not everyone approves of the festival.</td>
</tr>
<tr>
<td></td>
<td>You’ve never approved of Henry, have you?</td>
</tr>
<tr>
<td></td>
<td>Catherine’s parents now approve of her marriage.</td>
</tr>
<tr>
<td></td>
<td>I don’t approve of cosmetic surgery.</td>
</tr>
</tbody>
</table>

In a similar study on the use of dictionaries for language production purposes, Dziemianko (2006) carried out an impressive large-scale study with 606 high-school
and university students in Poland to investigate the user-friendliness of different types of systems used in dictionaries for describing the syntactic behaviour of verbs. She found that grammar codes were used only by the university students (who happened to be attending courses in Linguistics and English Grammar), but they preferred explicitly-worded information on grammar over just abbreviated codes. In the same way as in the Dutch experiment, Dziemianko found that in the Polish context too COBUILD-type definitions were not at all popular for retrieving information about verb syntax, whereas examples were again perceived as the best option.

In a different corner of the world, Chan (2011) gave a grammaticality judgment task accompanied by dictionary consultation to a group of 31 Cantonese students of English. While some dictionary entries given to the students contained explicit grammar information, this information was purposefully removed from other entries. Self-reporting protocols and an interview then showed that it was again the examples present in the entries rather than the explicit grammar information which the students valued the most. Dziemianko (2012) then conducted another impressively detailed large-scale study with 893 Polish university and junior high school students where, among other things, she explicitly compared how they reacted to examples and other ways of presenting information about word grammar in dictionaries. Again, she found that examples trumped all other forms of conveying syntactic information.

1.2 Studies on dictionary examples
Despite the fact that example sentences illustrating the use of words in context in dictionaries tend to very appreciated by language learners, early studies on the effect of dictionary examples per se did not disclose much evidence about their actual worth (Summers 1988, Laufer 1993, Nesi 1996, Al-Ajmi 2008). A possible explanation for this put forward in Frankenberg-Garcia (2012, 2014) could be that dictionaries do not seem to make clear a distinction between examples whose purpose is to help learners understand words and phrases and examples meant to help learners employ them while speaking or writing. As pointed out by Humblé (2001), examples to facilitate comprehension (decoding examples) should contain contextual clues about meaning, while examples to facilitate production (encoding examples) should illustrate grammar patterns and collocation. Dictionary example (1), given under the headword feign in the Oxford Advanced Learner’s Dictionary (OALD) online, contains contextual clues like survived the massacre and death, which can help a learner understand what feign means. As there is nothing particularly challenging about the way this transitive verb is used in context, one can say that (1) is primarily a decoding example. In contrast, dictionary example (2), from the headword cater in the Macmillan English Dictionary (MED) online, illustrates that the verb is intransitive and is followed by the preposition for, but does not seem particularly useful in terms of helping users understand what cater means. One can therefore say that (2) is primarily an encoding example. There are of course also examples that can help with both language comprehension and production at the same time. In example (3), from the MED online headword believe, previous supporters used in conjunction with stopped can give users an idea of what believe might mean, and the example also shows that in this particular sense the verb is intransitive and is followed by the preposition in.

(1) He survived the massacre by feigning death.
(2) I’m catering for a wedding reception next week.
(3) Some of his previous supporters have stopped believing in him.
As noted in Frankenberg-Garcia (2012), however, there do not seem to be many dictionary examples that give clear contextual clues about meaning and help with grammar at the same time. This should not be regarded as a shortcoming of dictionaries, but rather as a consequence of the simple fact that people do not give contextual clues about what words mean every time they employ them. It suffices to browse through a random set of concordance lines in a corpus to confirm this. As shown in (4), the randomly selected first five concordances for deliberately in the English corpus behind SkELL² (Baisa and Suchomel 2014) suggest that this adverb is typically used in the context of a passive construction, but the concordances do not give very clear contextual clues about what deliberately actually means.

Therefore health issues are deliberately brushed aside. The advertising rules surrounding children are deliberately strict. The word "consultation" is deliberately used here. This dire situation has been deliberately created. Its distribution was deliberately delayed three times.

Coming back to the above-mentioned earlier studies on the lack of much effect of examples (Summers 1988, Laufer 1993, Nesi 1996, Al-Ajmi 2008), the fact that they drew on real dictionary examples – as was indeed their intention – means there was no guarantee that the influence of examples on language comprehension was tested with decoding examples, and that the influence of examples on language production was tested with encoding examples. As discussed in Frankenberg-Garcia (2012, 2014), if one wishes to test the value of examples for aiding comprehension, the examples given to learners must at least have the potential to aid comprehension. If examples only illustrate target syntax or collocation without providing contextual clues that help to convey meaning, then it is unlikely that they will be of much use in comprehension. Likewise, examples for language production must have the potential to aid production. If they fail to exhibit the specific lexico-grammatical patterns required to satisfy a learner’s encoding needs, then they cannot be helpful in production.

Another problem of earlier experiments on dictionary examples is that the elicitation procedures of the studies looking at examples and language production (Summers 1988, Laufer 1993, Nesi 1996) asked the participants to write sentences with words that they were unlikely to have seen before. Thus the subjects probably had to look up the meaning of those words before they could employ them in writing and, as indeed Summers (1988) and Laufer (1993) themselves recognized, it is not natural for writers to use words that they have to look up for comprehension. Frankenberg-Garcia (2014) argued that in those experiments problems with language comprehension may have constituted an important intervening variable, distorting the results obtained with regard to the (lack of) effect of examples upon language production.

More recently, Frankenberg-Garcia (2012) and Frankenberg-Garcia (2014) – the latter being a conceptual replication of the former³ – revisited the question of the value of examples in language comprehension and language production by conducting a couple of experiments where, instead of resorting to real dictionary examples, the examples used were carefully selected from corpora so as to ensure that the participants (respectively 48 undergraduates and 50 high school students learning English in Portugal) were given decoding examples for the comprehension tests and encoding examples for the production tests. Also unlike previous studies, the production tests did not feature words that were unknown to the participants. Instead, the words tested were sufficiently easy for the participants to understand (to control for the comprehension

variable), but were often used incorrectly by Portuguese learners of English (to encourage the participants to look up how those words were used). Moreover, because learners might need to see more than just one example to be able to deduce what words mean or learn how to employ them grammatically, Frankenberg-Garcia (2012, 2014) also compared the differences between presenting learners with a single and multiple examples.

The two studies rendered very similar results with regard to the effects of examples upon comprehension: one decoding example helped learners understand previously unknown words a little, and multiple decoding examples worked as effectively as dictionary definitions. Although more replication studies are still necessary, in the present study I will not concern myself any further with the value of examples for language comprehension, for information on meaning is also effectively available from the definitions provided in monolingual dictionaries and from the translations supplied in bilingual dictionaries.

1.3 Encoding examples

For language production, on the other hand, examples remain crucial, for despite the fact that syntactic codes, or explicitly worded grammar, or syntactic information embedded in COBUILD-type definitions can help, examples seem not only to be more appreciated by learners, but they also seem to be more effective. In Frankenberg-Garcia (2012), both a single and multiple encoding examples helped the participants perform significantly better than the control group, and the participants with recourse to multiple encoding examples significantly outperformed the ones who had access to only one example. In Frankenberg-Garcia (2014), however, only the participants using multiple encoding examples obtained significantly higher scores than the control group. As discussed in Frankenberg-Garcia (2014), this discrepancy may be have occurred because of the different ways in which the data was elicited in the two studies. While the 2012 study the students were given an error correction test and were thus made aware that the words in focus had been misused, in the 2014 study the data was elicited via a translation task, where the participants were not told that the words being tested were usually problematic. Put differently, the participants in the 2014 study may have assumed the words they were required to employ in the translation task were very straightforward and may not have consulted the examples provided to them in the test as carefully as the participants in the 2012 study, who had been explicitly asked to try and correct errors. The difference between the two elicitation procedures is not trivial. There is no reason why learners should consult any references if they think they already know how to use the language, even if they are wrong about this. As discussed in Frankenberg-Garcia (2005, 2011), a major reason why learners do not use dictionaries as much as they should do is that they are often not aware of their own language limitations and reference needs. Thus no matter how well dictionaries present information on word grammar, learners will not use that information if they do not feel the need to look it up.

To summarize, what we know so far is that although examples are not the only source of grammatical information to help learners with language production, generally speaking learners tend to prefer to obtain this information from examples and seem to learn more from examples than from syntactic codes or from explicitly worded grammar information or from COBUILD-type definitions. Moreover, examples exhibiting the target syntax seem to be quite effective, particularly if learners are made aware of their reference needs and if more than one example is supplied.
Although space is a major concern of printed dictionaries, in an age where e-dictionaries are becoming more and more mainstream, the number of examples given will not have a great impact on the cost of the dictionary. In fact, in some of the electronic versions of major English learners’ dictionaries certain entries contain hyperlinks to further examples. But do learners actually click on those further examples at all? And if they did, would these further examples be useful? Too many examples could be distracting or even overwhelming. There could be an information overload that is actually detrimental to learning. Furthermore, it is not just the quantity of examples that is relevant. The type of examples also matters. When we click on the ‘examples bank’ of the LDOCE 4 CD-ROM when looking up approve, for instance, the examples given are not sorted according to meanings – to agree to something and to think that somebody/something is acceptable or suitable – which can be important for decoding purposes, and neither are they sorted in terms of transitivity, which can be important for encoding purposes. In the first five examples given in the LDOCE 4 ‘bank of further examples’ shown in (5), for instance, a decoding learner can be confused by the mix of senses and an encoding learner can be confused by the mix of transitive and intransitive uses of the verb. This means that without guidance, the information in the ‘dictionary examples bank’ can be as hard to interpret as raw corpus data.

Although space is a major concern of printed dictionaries, in an age where e-dictionaries are becoming more and more mainstream, the number of examples given will not have a great impact on the cost of the dictionary. In fact, in some of the electronic versions of major English learners’ dictionaries certain entries contain hyperlinks to further examples. But do learners actually click on those further examples at all? And if they did, would these further examples be useful? Too many examples could be distracting or even overwhelming. There could be an information overload that is actually detrimental to learning. Furthermore, it is not just the quantity of examples that is relevant. The type of examples also matters. When we click on the ‘examples bank’ of the LDOCE 4 CD-ROM when looking up approve, for instance, the examples given are not sorted according to meanings – to agree to something and to think that somebody/something is acceptable or suitable – which can be important for decoding purposes, and neither are they sorted in terms of transitivity, which can be important for encoding purposes. In the first five examples given in the LDOCE 4 ‘bank of further examples’ shown in (5), for instance, a decoding learner can be confused by the mix of senses and an encoding learner can be confused by the mix of transitive and intransitive uses of the verb. This means that without guidance, the information in the ‘dictionary examples bank’ can be as hard to interpret as raw corpus data.

(5) Bob doesn’t approve of alcohol.
 Congress voted not to approve the President’s plan for cutting the arms budget.
 Do you approve of doctors using human embryos for research?
 Don will only buy the motorcycle if his parents approve.
 Forty percent of registered voters approved of Senator Campbell.

In the discussion section of Frankenberg-Garcia (2014), I argued that when presenting learners with multiple examples, it was important to make sure that the examples were of the same kind. While decoding examples should be sorted by meaning, that is, according to the different senses of polysemous words, encoding examples need to be sorted according to different lexico-grammatical patterns of use. Moreover, I also recommended that there should be enough examples of each type so as to provide learners with repeated exposure to such patterns.

1.4 Aims of the study
In view of the above context and discussion, the aims of present study were to further explore whether exposure to encoding examples would help learners in language production. In particular, in this study the elicitation procedure was be refined so as to ensure that the students were not doing a simple error correction test, and that at the same time the students were motivated to consult the examples supplied to them, even when the language production test used to elicit the data looked very easy at first. In addition to this, I attempted to explore whether there was an optimum number of examples to be given to learners. In Frankenberg-Garcia (2012, 2014) the students had been given just one or exactly three examples. But would learners benefit from two examples? And would four examples help more? In this study I therefore also compared the effects of one, two, three and four examples.

2. Method

2.1 Subjects
The subjects participating in the present study were 161 undergraduates at a Portuguese university. Unlike in many studies on dictionary use, they were not language or linguistics students: the participants were a mix of undergraduates studying for degrees in Tourism (49 students), Hotel Management (87 students), Business Administration (4 students), Information Systems for Business (7 students), Secretarial Studies (4 students), Marketing (6 students), and Human Resources (4 students). All students had two hours of credit-bearing English classes per week. Coming from the Portuguese secondary school educational system where English is compulsory until year eleven and optional in year twelve, the level of proficiency of students enrolled in the university’s level 4 English groups from which the data was elicited was equivalent to B1 to B2 CEFR levels. The subjects were on average 22.7 years old, and 64 were male, 96 female. One student did not fill in the field for gender.

2.2 Elicitation procedure
The test given to the participants in order to elicit the data for the present study was administered during English classes, where the students were grouped together by level of English proficiency and period of study (mornings, afternoons or evenings) rather than by their respective academic disciplines. The test was conducted by four different lecturers teaching eight level 4 English groups in the same afternoon shift. Most students in this shift study Hotel Management or Tourism, which explains why there were fewer students from other disciplines taking the test.

The lecturers responsible for administering the test were given written instructions on how to proceed and the instructions were also discussed with them orally to ensure they fully understood the strict experimental conditions necessary. As the lecturers normally spoke English only in class, the test was administered in English, but clarifications in Portuguese were allowed.

The experiment was carried out in test conditions, but without a time limit. It took between five and ten minutes for the students to complete the pre-test, and then another five to ten minutes for them to complete the post-test carried out in the same session. Both of these will be described next.

2.2.1 Pre-test translation task. In the first, pre-test part of the experiment, the subjects were given a partial translation into English task likely to bring about typical mistakes by native speakers of Portuguese. The participants were not allowed to use any references for this part of the test, and were required to use a green biro that had been supplied to them, for reasons that would be explained to them in the post-test phase of the test. The ten test items given to the students and the classic mistakes associated with them are listed in figure 1. As can be seen, the translations were all very straightforward in terms of the target English vocabulary needed, but at the same time required the students to employ words that were difficult to use in the given context. For example, in test item one, B1/B2 CEFR-level Portuguese learners of English usually have no problem in retrieving the lexis required for the translation in question, but because of the influence of Portuguese, they often produce the error *it’s the first time someone does something* (rather than it’s the first time someone has done something). Although the students might also make other types of errors when translating those sentences, such as spelling mistakes or translating à noite in test item seven into at night instead of tonight, as shall be seen later in section 2.2.5, the focus of this study was limited to evaluating the specific errors highlighted in figure 1.
1. É a primeira vez que venho a este restaurante.  
   It’s the first time I have come/been to this restaurant. ✓
   It’s the first time I come to this restaurant. ✗

2. A piscina tem vinte metros de comprimento.  
   The swimming pool is twenty metres long. ✓
   The swimming pool has twenty metres. ✗

3. Aprova o trabalho do presidente?  
   Do you approve of the president’s work? ✓
   Do you approve the president’s work? ✗

   He has worked/has been working with us since 2010. ✓
   He works with us since 2010. ✗

5. Eles nunca confiaram em nós.  
   They (have) never trusted us. ✓
   They never trusted in us. ✗

6. Eu sempre segui os seus conselhos.  
   I (have) always followed your advice. ✓
   I always followed your advices. ✗

7. Ligue-me hoje à noite para o 212 745 975.  
   Call me tonight at/on 212 745 975. ✓
   Call me tonight to 212 745 975. ✗

   Congratulations on your new job. ✓
   Congratulations for your new job. ✗

9. Telefone para o banco amanhã de manhã.  
   Telephone the bank tomorrow morning. ✓
   Telephone to the bank tomorrow morning. ✗

10. Votei no candidato errado.  
    I voted for the wrong candidate. ✓
    I voted in the wrong candidate. ✗

Figure 1. Partial translation task, target translations, and typical errors foreseen

2.2.2 Experimental groups. After having completed the pre-test translation task described above, the participants were asked to hand back the green biros they had used in the translations and were randomly assigned to five different groups in each class taking the test. Group one (totaling 32 students) was supplied with a reference sheet containing one encoding example per test item, group two (33 students) received a reference sheet with two encoding examples per test item, group three (32 students) got three encoding examples and group four (33 students) was given encoding four examples per test item. Group five (31 students) was the control group and therefore did not get any reference sheets.

2.2.3 **Encoding examples.** The encoding examples supplied to the participants were selected from general English corpora (the British National Corpus, the Corpus of Contemporary American English and the enTenTen corpus). A few concordances were shortened to make them fit into a single line. Care was taken to ensure that all the encoding examples supplied focused on the target syntax addressed in the test. For instance, the concordances in (6) were selected to show that the present perfect is normally used after the expression *it's the first time* in test item one.

(6) It is the first time she's ever been here.
   It is the first time we have attended the exhibition.
   This is a brand new idea, it is the first time I have presented it in public.
   If it is the first time you have had the symptoms, consult your doctor.

The selection procedure focused on ensuring the examples chosen were easily readable by the students, with 95% of the vocabulary used falling within the above mentioned Oxford 3000 list. Words lying outside the Oxford 3000 were either proper names or had cognates in Portuguese. Note, however, that in a few cases the set of examples supplied were not neatly symmetrical. For example, as highlighted in (7), for test item six, different structures were used to illustrate that *advice* is uncountable: *a lot of advice,* *words of advice,* no quantifier and *pieces of advice.* Since in all four cases the Portuguese equivalent would have been a plural, they all provided opportunities for the students to realize that one cannot say *advices* in English. Although this emulates a learner’s natural encounters with the word in non-experimental conditions, in future experiments it would perhaps be best to ensure the examples given to the different groups were as comparable as possible.

(7) She gave him a lot of advice and help.
   Thank you so much for your words of advice.
   The first step is do not be afraid to seek advice.
   I was given three pieces of advice: don't write such complicated sentences, slow down and smile.

The reference sheet with the full set of concordances supplied to the experimental group using four examples is available in the appendix. The experimental group referring to a single example was given just the first concordance of each set of four in that list, the experimental group using two examples was given the first two concordances, and the experimental group using three examples was given the first three concordances.

2.2.4 **Post-test revision task.** After the appropriate reference sheets were distributed to the different experimental groups, the students were informed that the sentences they had just translated may have looked very easy but were in actual fact rather problematic, as Portuguese native speakers often made mistakes when writing those sentences in English. The participants were then asked to start the second, post-test part of the test, where they were told to go over the translations they had just written in green ink and, if they thought they had made any mistakes, they should supply a new translation underneath, using their own blue or black biros. The change of colour was meant to ensure the students could still see what they had written in the pre-test phase but did not feel tempted to overwrite those translations, and also to clearly set off any modifications made from the pre-test to the post-test phase. The students in the control group were required to go over their translations without resorting to any references, but like the other groups, they were made aware that the translations in the pre-test phase may have
been less straightforward than they might have initially thought. The remaining students were told the references sheets they had been given contained sentences that might help them improve their translations and were asked to read those sentences carefully while reviewing their translations. It was further explained to the students that the sentences under number one in the reference sheets could help with test item one, that the sentences under number two could help with test item two, and so on. This was done because, in the same way as in the study by Bogaards and van der Kloot (2002), the focus of the present study was on the effectiveness of the corpus examples supplied rather than on their findability.

The present pre-test and post-test format was an attempt to refine the elicitation procedures previously adopted in Frankenberg-Garcia (2012) and Frankenberg-Garcia (2014). As described in the introduction, in the 2012 study the subjects had been given an error correction test, which means they had been asked to correct errors which they themselves might not have committed in the first place, compromising to a certain extent the ecological validity of the experimental design. In the 2014 study, although the participants were given a partial translation task similar to the one used in the present study, they were not explicitly told that the seemingly easy sentences they were required to produce were actually prone to errors. As discussed earlier, this means that some of the more (over) confident participants in the 2014 study may not have consulted the references supplied to them during the test very carefully or at all, which may have in turn affected the results obtained. In the present study, care was taken to ensure that the students were led to reassess their own initial translations (rather than errors that they themselves might not have made), and that they were encouraged to use the references supplied (rather than naively dismissing them as superfluous because the test looked easy).

2.2.5 Marking the translations. The tests were all marked by a single marker in no particular order. To ensure consistency, the tests were marked item by item instead of participant by participant. The personal details of each participant and, most importantly, information about the experimental group to which they belonged, were in the cover sheet of the test, which was not visible to the marker during the procedure.

One point was given to each correct target grammar structure used. As shown in figure 1, variations in the translations were accepted, including variations that were not anticipated in figure 1 such as the president’s work vs. the work of the president in test item three or tomorrow morning vs. tomorrow during the morning in item 9. Errors such as spelling mistakes that were unrelated to the target area of difficulty in the tests or inappropriate lexical choices like your new work instead of your new job in text item 8 were disregarded. The vast majority of the errors identified were nevertheless exactly the ones that had been predicted in figure 1. Partial corrections were assigned half marks. This happened only in test item two, where the students added long after *This swimming pools has twenty metres, but failed to replace the verb has with is. Although *This swimming pool has twenty metres long remains incorrect, the students were at least able to notice the presence of the adverb at the end of the construction, hence the half mark.

As the tests were not returned to the students, no attempt was made to provide any corrective feedback, although the lecturers responsible for administering the tests were told they could supply the students with the correct answers in the following session if they wished.

Each test was assigned two grades. The pre-test grade was issued to the initial translations in green ink. The post-test grade was given to the translations in phase two,
after the participants had been encouraged to review their original translations with the help of the reference sheets supplied to the experimental groups or, in the case of the control group, without any references. This procedure enabled one to compare the pre-test and post-test scores of each individual student, and then compare the overall differences in scores between and within the groups. The results obtained will be presented in the next section.

3. Results

The initial results of the present study are summarized in tables 2a and 2b. The pre-test scores in table 2a show that the partial translation task used to elicit the data was quite difficult for the majority of the students, with the overall mean score obtained being below 2, and the median just 1 point out of 10. This was a good sign that the test questions were sufficiently difficult for there to be potential for improvement in the post-test phase. There was nevertheless a small number of students who performed far better than the others in the pre-test, as can be seen from the figures presented in the range column. Looking at the pre-test scores for each group, it can be seen that the differences between groups were not remarkable, a fact which reflects the random way in which the students were assigned to each group. Note, however, that one student in the control group with a score of 9, two students in the two-examples group with scores of 8 and 9, and one student in the three-examples group with a score of 10 were considered outliers, whose performance would be impossible to improve much or at all in the post-tests given the exceptionally high scores already obtained in the pre-tests. A conscious decision was therefore made to reject these outliers and exclude them from further analysis, on the grounds that their performance in the pre-test was too good for them to qualify as suitable candidates for the present experiment. With initially 161 participants, the loss of just four would not represent a major loss to the experiment. After rejecting these four outliers, the new, recalculated values for the pre-test scores are presented in table 2b.

The post-test scores are in turn presented in table 3. It shows that the mean and median scores of all four experimental groups increased, whereas there was only a slight increase in the mean and no change in the median score of the control group. Interestingly, there was also an increase in the standard deviations and ranges for all groups except the control group, suggesting that the encoding examples provided to the experimental groups may have made individual learner differences become more salient.

**Table 2a.** Preliminary pre-test scores

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean pre-test score (out of 10)</th>
<th>Standard Deviation</th>
<th>Range</th>
<th>Outliers</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>31</td>
<td>2.00</td>
<td>2.37</td>
<td>0-9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 encoding example</td>
<td>32</td>
<td>1.44</td>
<td>1.56</td>
<td>0-5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2 encoding examples</td>
<td>33</td>
<td>2.18</td>
<td>2.26</td>
<td>0-9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3 encoding examples</td>
<td>32</td>
<td>2.16</td>
<td>2.30</td>
<td>0-10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4 encoding examples</td>
<td>33</td>
<td>1.70</td>
<td>2.05</td>
<td>0-7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>All groups</td>
<td>161</td>
<td>1.90</td>
<td>2.21</td>
<td>0-10</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2b. Pre-test scores (after removing outliers with scores 8-10)

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean pre-test score (out of 10)</th>
<th>Standard Deviation</th>
<th>Range</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>30</td>
<td>1.77</td>
<td>2.01</td>
<td>0-7</td>
<td>1</td>
</tr>
<tr>
<td>1 encoding example</td>
<td>32</td>
<td>1.44</td>
<td>1.56</td>
<td>0-5</td>
<td>1</td>
</tr>
<tr>
<td>2 encoding examples</td>
<td>31</td>
<td>1.77</td>
<td>1.61</td>
<td>0-6</td>
<td>1</td>
</tr>
<tr>
<td>3 encoding examples</td>
<td>31</td>
<td>1.90</td>
<td>1.83</td>
<td>0-7</td>
<td>1</td>
</tr>
<tr>
<td>4 encoding examples</td>
<td>33</td>
<td>1.73</td>
<td>2.05</td>
<td>0-7</td>
<td>1</td>
</tr>
<tr>
<td>All groups</td>
<td>157</td>
<td>1.72</td>
<td>1.81</td>
<td>0-7</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Post-test scores

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean post-test score (out of 10)</th>
<th>Standard Deviation</th>
<th>Range</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>30</td>
<td>2.00</td>
<td>1.98</td>
<td>0-7</td>
<td>1</td>
</tr>
<tr>
<td>1 encoding example</td>
<td>32</td>
<td>4.53</td>
<td>2.59</td>
<td>0-9</td>
<td>4</td>
</tr>
<tr>
<td>2 encoding examples</td>
<td>31</td>
<td>5.32</td>
<td>2.58</td>
<td>0-10</td>
<td>5.5</td>
</tr>
<tr>
<td>3 encoding examples</td>
<td>31</td>
<td>6.03</td>
<td>2.61</td>
<td>0-10</td>
<td>7</td>
</tr>
<tr>
<td>4 encoding examples</td>
<td>33</td>
<td>5.18</td>
<td>2.20</td>
<td>1.5-9</td>
<td>5</td>
</tr>
<tr>
<td>All groups</td>
<td>157</td>
<td>4.64</td>
<td>2.74</td>
<td>0-10</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Next, table 4 summarizes the differences between pre-test and post-test scores. It shows that on average the participants improved almost three points in the post-tests, but the control group improved very little or not at all (with a mean of 0.23 and a median of 0), while the three-examples group showed the most improvement, with a mean difference of 4.13 and a median of 4. Note, however, that there was a lot of variability within groups, with overall changes in scores ranging from -2 to 8.5.

Table 4. Difference between pre-test and post-test scores

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean difference of pre-test and post-test scores</th>
<th>Standard Deviation</th>
<th>Range</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>30</td>
<td>0.23</td>
<td>1.16</td>
<td>-2-4</td>
<td>0</td>
</tr>
<tr>
<td>1 encoding example</td>
<td>32</td>
<td>3.19</td>
<td>2.21</td>
<td>0-8</td>
<td>3</td>
</tr>
<tr>
<td>2 encoding examples</td>
<td>31</td>
<td>3.55</td>
<td>2.26</td>
<td>0-8</td>
<td>3</td>
</tr>
<tr>
<td>3 encoding examples</td>
<td>31</td>
<td>4.13</td>
<td>2.21</td>
<td>0-8.5</td>
<td>4</td>
</tr>
<tr>
<td>4 encoding examples</td>
<td>33</td>
<td>3.45</td>
<td>1.69</td>
<td>0.5-7</td>
<td>3.5</td>
</tr>
<tr>
<td>All groups</td>
<td>157</td>
<td>2.94</td>
<td>2.36</td>
<td>-2-8.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

To obtain a clearer picture of the differences between pre-test and post-test scores for each separate group and test whether any improvement observed was significant, the two data sets were compared using the Wilcoxon signed-ranked test. This was preferred over the more common matched t-test used for paired data because the difference between pre-test and post-test scores for each group were not normally distributed. The null hypotheses were that, for each group, there would be no improvement from pre-test to post-test scores. The results of this analysis are summarized in table 5, which shows

that the null hypotheses can be rejected for all experimental groups with a 95% degree of confidence, but not for the control group. Table 5 also shows that the majority of the 30 participants assigned to the control group did not score better in the post-test, and five of them even got lower marks after being asked to revise their initial translations. In contrast, in all experimental groups, the majority of the participants scored better in the post-test, only a few were unable to improve their translations, and none actually made the translations worse. Note, however, that there were a few cases where students were able to correct the translation of one test item, but then introduced an error in another test item, resulting in a net effect of zero.

**Table 5. Differences between pre-test and post-test scores (Wilcoxon signed-rank tests)**

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>post-test score higher</th>
<th>post-test score lower</th>
<th>post-test score equal</th>
<th>Z value*</th>
<th>W value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>30</td>
<td>7</td>
<td>5</td>
<td>18</td>
<td>1.0198</td>
<td>26</td>
</tr>
<tr>
<td>1 encoding example</td>
<td>32</td>
<td>28</td>
<td>0</td>
<td>4</td>
<td>4.6226*</td>
<td>0*</td>
</tr>
<tr>
<td>2 encoding examples</td>
<td>31</td>
<td>30</td>
<td>0</td>
<td>1</td>
<td>4.7821*</td>
<td>0*</td>
</tr>
<tr>
<td>3 encoding examples</td>
<td>31</td>
<td>29</td>
<td>0</td>
<td>2</td>
<td>4.7030*</td>
<td>0*</td>
</tr>
<tr>
<td>4 encoding examples</td>
<td>33</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>5.0119*</td>
<td>0*</td>
</tr>
</tbody>
</table>

* = significant at p ≤ 0.05 (one-tailed)

Having analysed the differences in pre-test and post-test scores for each group separately, the next step in the analysis was to compare the different groups with each other. To carry out this comparison, a non-parametric Kruskal-Wallis test was preferred over a one-way ANOVA because the data was not normally distributed. The null hypothesis was that there would be no difference between the groups. With an alpha level of 0.05 and a degree of freedom of 4 (given the five groups being compared), the critical $X^2$ value (9.48773) was found to be smaller than the calculated $H$ value (55.986), which means the null hypothesis can be rejected. One can therefore conclude that there are significant differences between the five groups. Looking back at the actual differences in pre-test and post-test scores in table 4, it seems logical to attribute the major cause of this difference to the control group, which did not receive any experimental treatment.

The above result does not, however, shed any light on whether the number of examples presented to the participants in each group actually made a difference. Therefore, to focus on the differences between the number of examples, the Kruskal-Wallis test was used again, but this time purposefully leaving out the control group. The null hypothesis was that there would be no difference between the treatments (i.e. supplying the participants with one, two, three or four examples). The critical $X^2$ value at the 0.05 level with a degree of freedom of 3 (7.81473) was found to be greater than the calculated $H$ value (3.713), which means the null hypothesis cannot be rejected. It is therefore not possible to conclude that the differences between the different experimental groups are significant. Thus despite the slightly higher overall post-test improvement noted among the three-examples group in table 4, this difference was not statistically significant.

**4. Discussion**
The overall results presented in the previous section are in accordance with the general findings of Frankenberg-Garcia (2012, 2014), and, as in Bogaards and van der Kloot

(2002), Chan (2011) and Dziemianko (2006 and 2012), they lend further support to the idea that examples indeed seem to help language production. More specifically, in the present study it was possible to show that simply telling people that their translations are prone to errors does not in itself result in improvement, whereas supplying them with examples in order to assist them revise their initial translations appears to be beneficial. In fact, the lower post-test scores of some of the participants in the control group indicate that just drawing learners’ attention to problematic areas can even be counterproductive. Of course, when supplying learners with examples, it is not just any example that helps: unlike earlier studies focusing on dictionary examples (c.f. Summers 1988, Laufer 1993 and Nesi 1996), and unlike existing examples in learners’ dictionaries, which do not distinguish between encoding and decoding examples, the examples used in the present study and in Frankenberg-Garcia (2012, 2014) were hand-picked so as to ensure they exhibited the target syntax or collocation required.

Having ascertained that encoding examples that focus on the target syntax or collocation are beneficial, the challenge that remains is to make sure dictionary users are able to access suitable encoding examples whenever the need arises. The present study (and Frankenberg-Garcia 2012, 2014) was carried out in a controlled experimental setting, where the participants were basically spoon-fed with appropriate examples. In the real world, however, learners would need to procure those examples by themselves. A possible solution would be for learners to retrieve their own examples from corpora. However, these would not be sorted according to meaning or usage, and grappling with corpora and raw corpus data can be quite daunting at first. Another possibility would be for learners’ dictionaries to supply separate encoding examples for language production purposes. But even if this were to be done, dictionary user research shows that people continue to use dictionaries primarily for language comprehension, and many of the dictionary resources to aid language production remain underexploited and secondary to information to aid comprehension. More research is therefore needed in order to develop efficient ways in which suitable encoding examples can be made accessible to learners.

The other main purpose of the present study was to try and determine whether the number of examples mattered. From the results obtained here, it cannot be concluded that the number of encoding examples supplied makes a difference. Thus unlike Frankenberg-Garcia (2012), where three encoding examples helped more than just one example, and unlike Frankenberg-Garcia (2014), where one example did not lead to significant improvement but three examples did, in the present study it was not possible to observe that more examples were more effective than just one example and neither it was possible to define an ideal number of examples.

There are a number of factors that could account for this outcome. The first point to bear in mind is sample size. With 157 subjects (after the 4 outliers were excluded), the present study counted with over three times as many participants as the 2012 study (48 subjects) and as the 2014 study (50 subjects). Another related issue is the sampling itself. Whereas the participants in the 2012 study were recruited from three groups of Tourism students at the same university and the ones in the 2014 study were from two groups of year eleven students at the same secondary school, as reported in section 2.1.1, in the present study the participants were mixed groups of undergraduates pursuing degrees in different subject areas. The larger sample size and more varied backgrounds of the participants should contribute to rendering the present findings more generalizable than those of the previous studies.

In addition to this, the refined elicitation procedure adopted in the present study also enabled one to assess progress from pre-test to post-test rather than evaluate just.
the post-treatment results, as was the case in the 2012 and 2014 studies. Although those experiments made use of control groups to account for previous knowledge of the language tested, the pre-test and post-test procedure used in the present study helped to further tease out the possible effect of this variable - so much so that a conscious decision was made to exclude from the current sample the four outliers with scores of 8 and over in the pre-test phase.

Another non-trivial factor related to the elicitation procedure is that, as discussed in the introduction, in the 2012 study the students did a test that focused on errors that they may not have committed or even noticed, while in 2014 study the participants were not made aware that the sentences they had been required to translate were problematic. In contrast, in the present study the students were explicitly asked to reassess possible problems in their own translations. It is therefore possible that they were more engaged in using the examples supplied to them to figure out ways of improving their own writing than in previous studies. This could explain why even the participants with access to just one example performed as well as the participants supplied with more than one example.

Individual differences among the subjects must also be taken into account when interpreting the present results. If the number of examples presented to learners does not seem to make a difference overall, it may well be the case that some individuals prefer and react better to fewer examples, while others favour and require more examples, and there may be even a few people for whom encoding examples do not help at all. Indeed, the higher standard deviations and ranges observed for all groups except the control group in the post-test phase suggests that the examples supplied to the experimental groups may have exacerbated individual differences. Thus it is also necessary to carry out more research about individual preferences and needs with regard to an ideal number of examples.

5. Conclusion
This study arose from the need to further explore the topic of helping learners with language production. Language users seem to appreciate examples from dictionaries, but there is not much proof that the examples in dictionaries as they are today suffice. I argued that when presenting learners with examples aimed at helping in language production, it was important to use encoding (as opposed to decoding) examples, and to sort these examples according to different lexico-grammatical patterns of use. Based on previous findings reported in Frankenberg-Garcia (2012, 2014), it also seemed necessary to supply learners with more than just one example. The present study therefore proceeded to further investigate whether encoding examples would help learners in language production using a refined elicitation procedure and larger experimental and control groups, and to test whether learners reacted better to an optimum number of examples. The results obtained reinforce the idea that encoding examples are helpful to writers, but it was not possible to establish what constitutes an ideal number of examples, suggesting that there might be significant learner variability in this respect. Further research remains necessary, particularly with regard to exploring individual learners’ preferences in relation to encoding examples, and with regard to developing efficient ways of rendering examples accessible to them.

Notes
1 The information in table 1 was gathered from the following five major online learners’ dictionaries on 03/11/2014: Collins English for Learners (COBUILD), Oxford Advanced Learners’ Dictionary (OALD),


2 Retrieved from http://skell.sketchengine.co.uk/ on 17/11/2014

3 In addition to using a different population sample, changes in the experimental design were also made.


5 100% of the target words required for this task lie within the Oxford 3000 keywords, which are held to be the core words of the English language in terms of frequency and usefulness. See http://oald8.oxfordlearnersdictionaries.com/oxford3000/.

6 Respectively available at http://corpus.byu.edu/bnc/ (BNC), http://corpus.byu.edu/coca/ (COCA) and http://www.sketchengine.co.uk/ (enTenTen) on 15/04/2015. See also Davies (2008) and Kilgarriff et al. (2004) for further details about these corpora.

7 The term sentences was deliberately used instead of concordances, in order not to confuse the participants with corpus-linguistics terminology.

8 According to the John Tukey test (Tukey 1977), outliers fall below the first quartile or above the third quartile.

9 H = Kruskal-Wallis Test statistic

10 Note, however, that great progress has been made with regard to developing interfaces capable of bringing corpora closer to laypersons with little or no knowledge of corpus linguistics. A remarkable recent example is the Sketch Engine’s SkELL corpus-cum-online interface, at http://skell.sketchengine.co.uk/run.cgi/skell. See also Baisa and Suchomel (2014).

References

A. Dictionaries


B. Other literature


**Appendix**

**Reference sheet 4.** *The sentences below may help you improve your translations. Read sentences 1 to help you with translation 1, read sentences 2 to help you with translation 2, and so on. Please do not write anything on this reference sheet.*

1. It is the first time she's ever been here.  
   It is the first time we have attended the exhibition.  
   This is a brand new idea, it is the first time I have presented it in public.  
   If it is the first time you have had the symptoms, consult your doctor.

2. The tank was about ten metres wide and fifteen metres long.  
   The wall was roughly 80 metres long and more than 3 metres high.  
   The Panama Canal lock chambers are 305 metres long and 33.5 metres wide.  
   The tunnel was about four hundred metres long.

3. Her great-grandmother would not only approve of Julia's plan but applaud it.  
   She smiled at him, as if she approved of what he said.  
   He approves of her every action and would do anything for her.  
   A majority of Americans approve of interracial marriage

4. He has been in charge since April 2000.  
   They have lived in Edinburgh since 1984.  
   Since then, I have been writing about and teaching art and theater.  
   All of the amateur winners have since turned pro.

5. Always trust your instincts.  
   Don’t trust anyone.  
   Make sure you can trust your friends.  
   I use their services because I trust their judgment.

6. She gave him a lot of advice and help  
   Thank you so much for your words of advice.  
   The first step is do not be afraid to seek advice.  
   I was given three pieces of advice: don’t write such complicated sentences, slow down and smile.

7. Call him directly at 214-555-049 to give him your response.  
   Give us a call at 800-989-8255.  
   For information or directions, call the church at 281-297-5900.  
   Call me at 719-579-6407 between 9:00 AM and 4:30 PM.

8. Congratulations on your graduation!  
   First of all, congratulations on the success of your new e-book.  
   I thanked him for his congratulations on my birthday.  
   Congratulations on your 50th wedding anniversary.

9. I must telephone the police.
   She telephoned the office on a Sunday night using the after-hours number.
   Bea telephoned the hospital and asked the nurse if Corren had gone home.
   The patient telephoned the pharmacy

10. I tried to vote for you today, but they told us we couldn't vote any more.
    I didn't vote for the president's healthcare plan.
    Are they voting for the other guy?
    Why should people vote for democrats who act like republicans?