Introducing the CORYL Corpus: What it Is and How We Can Use it to Shed Light on Learner Language

Angela Hasselgreen and Kari Telstad Sundet

This article presents CORYL (CORpus of Young Learner language), and demonstrates how the corpus can help reveal or shed further light on many phenomena which are manifested in the written English language of Norwegian school pupils. The paper begins with an introduction of CORYL, then focuses on learner language and the role of corpora in the study of this. The following sections are devoted to what we term *Computer-aided Error Analysis* and *Interlanguage Analysis* (not involving errors). Within these sections, extracts and other findings from CORYL-searches are presented to illustrate what we believe CORYL is able to indicate about the language of these learners.

1. Introducing CORYL

1.1. Background and development

CORYL (see *http://clarion.uib.no/korpuskel*), is a learner corpus, which consists of English texts written by pupils in Norwegian schools. CORYL is hosted by Corpuscle, which "is a corpus management and analysis system for annotated corpora" (*http://clarino.uib.no/korpuskel/page).* The tool is being developed at Uni Research Computing, in collaboration with colleagues at the University of Bergen. The work has been supported by grants from The Research Council of Norway, the Meltzer foundation and the CLARINO infrastructure. The authors have been involved in the tagging and proofing of the texts in the corpus.

The aim of compiling CORYL was to allow researchers to carry out a range of investigations into the *interlanguage* (Selinker 1972) of these

Remaining Relevant: Modern Language Studies Today

Bergen Language and Linguistics Studies · vol. 7 · 2017 © Angela Hasselgreen and Kari Telstad Sundet.
DOI: http://dx.doi.org/10.15845/bells.v7i0

learners, which can be regarded as the learner's current mental version of the target language. While corpora are not well suited to studying the idiosyncrasies of an individual's interlanguage, they can expose patterns in the language use of the group, or subsections of the group. This may involve using the corpus data to investigate phenomena that are already thought to characterize the language of these pupils (corpus-based/informed research), or to allow the corpus to reveal patterns in the language, which can then be further investigated (corpus-driven research). It is important to note that while this will often involve studying errors, non-errors, according to Callies (2015, 41), "can reveal as much about learner language as errors, e.g. when it comes to avoidance phenomena, […]."

The term 'errors' has somewhat lost favour in recent decades, as the Interlanguage perspective has gained ground in SLA (Second Language Acquisition). 'Error' is, however, used here in accordance with other literature on corpus linguistics, e.g. Granger et al (2015). This is because it is necessary in corpus tagging to have a precise notion of what is being selected for analysis. A term such as 'non-targetlike form' may be open to interpretation. Following the recommendation of Ellis and Barkhuizen (2005, 59), the errors identified in CORYL are restricted to *absolute errors*, as opposed to *dispreferred forms*, which would involve subjective judgments.

The texts that make up the current corpus were collected in the course of the National Tests of English in 2004 and 2005, and can thus be considered largely synchronic, although the recent addition of new texts adds a diachronic dimension to the corpus. The tests were designed to measure pupils' writing abilities in English, the pupils being given a number of tasks, such as: *Write a post card*, *Describe what you see in the picture*, and *Write a letter to the editor*. The language in the corpus can be considered 'natural' in that pupils were given a free hand as to the actual language they used. The genres and themes are limited to what was prescribed by the tasks; texts are annotated for the task type/genre.

The texts were taken randomly from pupils in the 7th, 10th and 11th grades, to ensure some degree of representativeness in the texts. The texts were collected in the course of National Tests of English Writing, and were from randomly selected schools distributed widely across the country. The corpus at the time of writing contains 272 texts from pupils in

7th and 10th grades, and a number of texts from the 11th grade will soon be uploaded to the corpus. Recently collected 7th-grade texts, which add a diachronic dimension, will also be added to the corpus. The texts are annotated for gender, as well as the approximate age of the pupils: 11–12 years (7th grade) and 15–16 years (10th/11th grade). No information is available regarding the L1 background of the writers, although it is assumed that most pupils have Norwegian as an L1 or an L2. The original texts were handwritten and had to be typed manually. Each of the transcribed texts has been proofread by a rater to ensure that it has been accurately copied. To safeguard the pupils' privacy, the texts have been anonymized and any potentially identifying information has been removed from the material.

A notable feature of the corpus is the fact that most of the texts have also been assigned to levels and half levels on the CEFR by multiple raters. Thus, while no equivalent native speaker corpus exists for purposes of contrasting language phenomena, a comparison can be made between texts not only from older and younger pupils, but also from pupils placed at different levels of proficiency.

Finally, it should be mentioned that the limited nature of the CORYL corpus (currently totaling 129,421 words) places limitations on the elements that can be searched for, and yields low numbers of occurrences in many instances. No claim is made that generalizations are possible based on the corpus findings; these should be regarded as no more than indications. Thus, it is particularly important that CORYL corpus searches are followed up by or combined with experimental research, for example in the language classroom.

1.2. Error annotation in CORYL

Because an important intended use of the corpus is to reveal patterns of errors, the texts are manually annotated, or tagged, for all errors, following the coded classification shown in the appendix. The tagging has been carried out by a native speaker of Norwegian with high-level English competence, and checked by a native speaker of English, with a good knowledge of Norwegian. As it is impossible to reconstruct any part of a corpus text in a 'correct' form, a certain degree of interpretation is inevitable, in order to hypothesise the intended target meaning (Ludeling & Hirschmann, 2015, 141–43). In the case of CORYL this has sometimes proved difficult, especially in low-level texts, with a consensus having to

be sought between tagger and proofreaders; occasionally a tag of 'nonsense' has had to be used. Those who subsequently work with the corpus data must be free to disagree with the tagging, which reinforces the need in corpus research to combine automatic searching with 'hands on' judgements. In order to be able to search for words without depending on correct spelling, CORYL contains corrected version of all spelling errors.

The decision was made from the start not to use an automatic Part-of-Speech (POS) tagger, due to the large number of errors in many of the texts, especially in the 7th grade. Even when used on standard English texts, these have been shown to misidentify many tokens; Manning (2011) cites a misidentification rate of 270 words for every 10 000. With so many CORYL texts being written at CEFR levels A1–A2, POS tagging would necessitate the manual correction of many thousands of errors.

Below is an example of a sentence that it would be impossible for a machine to tag correctly, but that can be interpreted by a native speaker familiar with the context in which it was written:

Example 1: L1 phrase
theyr **in theyr best age**. (Norwegian: De er i sin beste alder)
(p07–10)

It is important to point out at this point that while 'L1' is used frequently here, and in the actual tagging, it should not be interpreted as strictly meaning the first language of the writer, as this was not identified in the material used to compile CORYL. It can, in fact, rather be interpreted as 'Norwegian', which is the school language for these pupils, and the first language of most. Recent additional texts have been collected to expand the corpus, which contained details of pupils' first languages. While these have not yet been fully analysed, first indications would suggest that Norwegian has a strong influence on the writing of pupils with first languages other than Norwegian.

2. *Studying learner language with the help of* CORYL

The study of learner language using an electronic corpus has two specific advantages over more traditional SLA studies, according to Granger et al. (2015,1). Firstly, language samples are accessible from a large number of speakers or writers, who are arguably more representative than in tradi-

tional studies that sometimes use only a handful of subjects. Secondly, the search for phenomena can be done rapidly, without the cost in time and resources that would be necessary for manually studying samples. An additional advantage is that data can be sorted and contrasted; for example, comparing the language of pupils placed on different CEFR levels or in different age groups. While the CORYL corpus is relatively small, and limited in terms of themes (and therefore vocabulary), it is sufficiently large to allow for the study of many phenomena that occur frequently in the language of pupils in this context. Either by setting out to investigate issues that we know to be salient, or by allowing the corpus to reveal phenomena we were not previously aware of, it is possible to identify characteristics of learner language which may prove useful for researchers, teachers and course-book writers, as well as others involved in the English language of young learners in the Norwegian school context, and hopefully beyond.

Two distinct areas of study related to learner language, using CORYL, are identified here: *Computer-aided Error Analysis* (CEA) and *Interlanguage Analysis* (IA). These terms are based on the presentation of the topic in Callies (2015, 38–41). For the second of these areas, Callies employs the term *Contrastive Interlanguage Analysis* (CIA), since such studies normally contrast the interlanguage in the corpus with the language of native speakers. However, in the case of CORYL, no parallel corpus of native-speaker language is available, which corresponds with regard to the age of the writers and the tasks responded to. Thus, the term *Interlanguage Analysis* is used to describe the study of language in the corpus that does not (necessarily) involve errors, but nonetheless reveals patterns in the language that warrant further investigation.

Extracts from the CORYL corpus will be used to exemplify features of language use. These are accompanied by a reference to the writer (pupil); for example, p04–10 or p151–07, which denotes pupil ID and grade.

3. Computer-aided Error Analysis (CEA)

In this section, the focus is on writing errors detected in the texts in the corpus, and the way these can be analysed, or used as the basis for further study. It has long been acknowledged, for example by Odlin (1989), that many errors appear to be caused by the interference of the writer's L1, or another familiar language, on the interlanguage. Other errors appear not

to be influenced by the specific L1, but may be due to developmental factors. These two types of error will be categorized respectively as 'interference errors' and 'non-interference errors'.

3.1. Interference errors

Interference errors are frequently ascribed to *transfer*. Ellis and Barhuizen (2005, 65) state: "Transfer relates to the introduction of an L1 form into the interlanguage system." This can be illustrated by the phrase in example 2:

> *Example 2: L1 phrase*
> […] the old man […] **looks** angry **out**. (Norwegian: *Den gamle mannen ser sint ut*)
> (p18–10)

Ellis and Barhuizen (2005) distinguish transfer from *borrowing*, by which an L1 item is simply used as it stands, e.g. using the word Norwegian word 'paraply' for 'umbrella'. 'Transfer' is preferred here to the term 'crosslinguistic influence', frequently used in SLA literature, e.g. Kellerman and Sharwood Smith (1986), as it is felt the latter term could be interpreted as including borrowing.

CORYL contains numerous examples of both transfer and borrowing. The latter is easiest to identify, and is simply tagged in CORYL as L1. With regard to transfer, caution has to be employed when ascribing an error to this category. According to Osborne (2015, 351): "It is not enough to identify errors in learner production and attribute them to transfer solely on the grounds that similar forms exist in the learners' L1, if other sources of error have not reasonably been ruled out". In the CORYL tagging process, errors have primarily been tagged according to the surface form of the error; for example, wrong preposition. Where it seems likely, by comparison with Norwegian, that transfer may have occurred, the error is also tagged as L1, but the researcher is warned that this is a fairly subjective judgment. It is convenient for a researcher using CORYL to search for instances where transfer is likely to have occurred, but this should not be regarded as more than an indication. The tagging in not intended to explain *why* an error has occurred; that is a task which researchers using the corpus might address. Examples 3 and 4 are tagged as preposition and modal errors respectively, but are also tagged as L1-influenced.

Example 3: Preposition/L1
I'm scared **for** sleep in the wood (Norwegian preposition in this context: *for*)
(p248–7)

Example 4: Modal/L1
John was like a hero when he s**hould** save a pizza from horrible things. (Norwegian modal in this context: *skulle*)
(p45–7)

Transfer is frequently found in lexical errors, notably in those apparently caused by cognates or 'false friends', which Ringbom (1983) ranks as the greatest single cause of written errors among Swedish-speaking learners of English. *Nature*, as used in Example 5, is one of numerous false friends found in the corpus:

Example 5: Wrong word/L1 (*cognate*)
They throw sigaretts in the **nature (**Norwegian equivalent in this context: *naturen*)
(p146–10)

The top ten wrongly used words in CORYL are shown in Table 1. Many of these might be described as 'core words', referred to in Section 4, that learners tend to over-rely on. Another favourite cognate, *mean,* which is often used to denote *think*, is ranked number 24.

Table 1. The top ten words used incorrectly in CORYL (expressed as percentage of total number of wrong words)

1. got (2,30%)
2. take (1,91%)
3. the (1,71%)
4. came (1,45%)
5. took (1,38%)
6. get (1,32%)
7. nature (1,32%)
8. go (1,18%)
9. went (1,12%)
10. back (1,05%)

In addition to affecting single vocabulary items, transfer appears to result in the direct translation of whole phrases from the L1 (tagged as L1P), as seen in example 6, which show the influence of the Norwegian phrase *ta telefonen* = *'take the telephone'*.

Example 6 also contains another instance of the word *to take* (*took*) used incorrectly. While this may also be a direct translation of the Norwegian word 'ta', it could be being used in the sense of 'picked up' or 'grabbed' rather than as part of a telephone-related phrase.

Example 6: L1 phrase
Then I took the telefone and I ring my dad. But he **not take it.**
(p280–7)

This finding confirms those of Salido (2016), obtained in an error analysis of 'Support Verb Constructions' (SVCS) in written Spanish learner corpora. This study maintained that one of the most frequently collocation types used correctly and incorrectly in the corpus was that of SVCS. These consist of a verb+noun construction, in which the verb carries little lexical meaning and this is provided by the noun. Salido found evidence that learners, while choosing the noun correctly, had problems identifying the correct support verb, often due to the restricted range of these available to them, compared to native speakers.

An example of an error in CORYL in which *got* is incorrectly used as a support verb is found in example 7, where *had* would have been correct.

Example 7: L1 phrase
my mother **got** a heartattack
(p176–10)

There is another source of interference error, which cannot be directly attributed to transfer, but nonetheless reflects the relationship between the L1 and target language. This is the phenomenon Hasselgren[1] (1994) refers to as *divergence*, which occurs when a single item in one language has a number of equivalents in another language. An example of this is the present tense verb form, which in Norwegian has only the simple form, while English has both simple and progressive forms. While the use of

1 Later spelt Hasselgreen.

the simple tense when the progressive is required can perhaps be directly ascribed to transfer, teachers of English in Norway, and possibly beyond, know that the progressive form is frequently used when the context requires the simple form, as demonstrated in example 8:

Example 8: Verb tense (*overuse of progressive form*)
The bus **is leaving** the railway station at 14.00
(p04–10)

Another example of divergence is the Norwegian *det er*, which can be translated as *it is* or *there is* in English. Example 9 shows an error in this feature:

Example 9: It/there is
[…] **it is** a little bit to walk to get there.
(p272–10)

A divergence-related finding that was unexpected was revealed when searching for wrong function words in CORYL. At the very top of the list, with twice the frequency of the word that was in second-place, we find the word 'the'. On closer examination of the texts, it becomes apparent that, while there are many of the more predictable article errors, *the* also commonly seems to be perceived as the most natural equivalent of the Norwegian word *det* (which may also be translated as *there*, *it or that*) and used accordingly. Examples 10 and 11 show the use of 'the' to convey a variety of meanings based on *det*:

Example 10: Wrong function word (*the = there*)
the were a dog on the flor.
(year 7)[2]

Example 11: Wrong function word (*the = that/it*)
Wath is **the** ?
(year 7)

2 The pupil id is unavailable for this examples

Divergence is also a common source of mischosen lexis. In a study by Hasselgren (1994) of the English vocabulary of young Norwegian adults in translation texts, it was found that over half of the words wrongly chosen could be ascribed to divergence.

This can be exemplified by the Norwegian word *gå,* which in English can be translated as either *go* or *walk*. Example 12 is one of occurrences in the corpus, where the word *went* is used with the meaning *walked*.

Example 12: Wrong word/L1 (*divergence/cognate*)
My mom **went** through my door and shouted…
(p222–10)

3.2. Non-interference errors

The examples shown from CORYL so far have involved errors believed to be influenced by interference from the Norwegian language, mainly due to transfer of Norwegian forms on pupil's English. However, many errors cannot be ascribed to interference, but rather to the stage of the development of the interlanguage.

As a learner's interlanguage grows, its development will, to some extent, reflect the learner's L1, but the findings of researchers such as Pienemann and Johnston (1987) revealed that common stages in the acquisition of certain structures seemed to exist universally, regardless of the L1 of the learner. Lightbown and Spada (2006) consider a number of these, including negation, question formation, possessive determiners, relative clause structures and references to the past. Errors that are developmental in nature can be studied in CORYL, by comparing the language of pupils placed on a range of CEFR levels.

Regarding the use of the past tense, it is interesting to note a development from A1 to A2. A1 pupils commonly use only *was* to indicate past tense, but on reaching A2, they start attempting to use other forms of the past tense, although this generally involves simply applying the *ed*-ending to most or all verbs.

The findings outlined below derive from a fairly superficial preliminary search for evidence in CORYL of the way verb tenses develop across CEFR levels. A more thorough search would be very worthwhile. The examples here illustrate what appears to be a typical progression in verb tense development across CEFR levels.

At **level A1**, pupils tend to consistently use *was* to indicate past tense, while virtually all other verbs are in the present or infinitive form. This is illustrated in Example 13:

Example 13: A1
I **go** in and **sad** HALLO! But nobady **sas** noting. The window **wos** knust. I **go** into my brothers room, and all hes things **wos** gone. I **call** my mom and dad.
(p30–7)

At **level A2**, pupils are generally still not using the past tense appropriately, but they are making various attempts, commonly involving an incorrect use of progressive forms. They may also use *ed*-endings wrongly. Examples 14 and 15 show typical A2 attempts:

Example 14: A2
Thay **are starting to build** a hut
(p215–7)

Example 15: A2
I **go** to the silver to see **wats has happening** with that. But it was ther. But many thing was **stoled**
(p26–7)

These findings confirm the findings of a study carried out by Salaberry (2000) on the acquisition of English past-tense verbs by Spanish leaners of English. It was found that irregular past-tense forms of frequently used verbs appeared to be acquired more readily than regular verb forms. This was attributed to the greater effect of saliency and frequency compared to that of the lexical aspects in the early stages of acquiring inflectional morphpology.

According to Salaberry: "In essence, the prediction is that the more frequent and irregular the verb, the more likely it will appear first in the development of past marking of adult instructed L2 learners, […]" (2000, 138); and he goes on to conclude: "In sum, the analysis of data from this study provided evidence that irregular morphology (e.g. extended past

tense marking of stative *be* as irregular *was*) correlated more strongly than lexical aspect with morphological past tense marking" (2000, 148).

At **level B1,** pupils demonstrate similar types of errors to those of A2 pupils, but here there are fewer verb errors relative to other error types. More complex verb errors of the type shown in example 16, in which the context demanded *grew up* or *have grown up*, start occurring relatively more frequently:

Example 16: B1
Because adults **are grown** up in an evironment that they had to work hard to get food.
(p53–10)

At **level B1/B2**, errors tend to occur in more advanced verb constructions, as shown in examples 17 and 18.

Example 17: B1/B2
I am looking forward **to see** you there.
(p55–10)

Example 18: B1/B2
I **remembered I forgot** my food
(p80–10)

At **level B2** there are very few verb errors, and these are typically quite subtle, as shown in example 19.

Example 19: B2
If there's anything that **bothers** you I'll do my best
(p79–10)

Regarding lexis, Hasselgren (1994) revealed that, in her data, about one in three of the wrong-word errors appeared to be caused by a misselected synonym. Although *much* is arguably a function word rather than a lexical word, it can be regarded as belonging to a group of synonyms: *much, many, a lot (of),* where the selection depends more on grammatical/syntactic considerations than semantic meaning. Its similarity with

the Norwegian *mye* may account for the fact that it tends be a favourite choice among learners, as illustrated by example 20:

Example 20: Wrong function word
Music means **much** to almost everybody
(p267–10)

Tagging items as errors on the grounds of inappropriate style or connotation is done with caution in CORYL, as choices may have been made consciously for rhetorical purposes, such as the use of irony. However, there are instances where a clearly unfortunate choice is made, as illustrated in example 21, which is an extract from a letter to the editor:

Example 21: Wrong word (*style/connotations*)
They put garbarge on the ground and **don't give a shit.**
(p278–10)

This section has been devoted entirely to illustrating how CORYL can be used to investigate some areas of pupil error, whether due to interference or the developmental stage of the interlanguage. The examples found have been tagged for specific errors and this tagging has been used to retrieve manifestations of these, leading to a discussion of how the error may have arisen, drawing at times on the literature of corpus linguistics or SLA.

In the next section, the focus moves from errors in the interlanguage of pupils, to those aspects in which actual errors do not necessarily take place, and yet the language is stamped by their development as learners. This means that we are no longer able to depend on the corpus tags to guide us to the features we are interested in (as in corpus-driven research). We need to begin any search with a hypothesis or at least a 'hunch' as to what to look for in the data.

4. Interlanguage analysis

While the language of learners is certainly not always 'incorrect' there are many instances in which it is simply 'different' from that of a native speaker in an equivalent situation. One of the principal ways in which

this is manifested is through the avoidance of certain forms, often accompanied by the overuse of others (Osborne, 2015, 336).

In the case of CORYL, while it may be suspected that Norwegian learners underuse a certain word or structure, the absence of a parallel native speaker corpus for comparison makes this difficult to establish; experimental study with pupils may be the best way to find evidence of over- and underuse. It is, however, possible to undertake a comparison between subgroups of CORYL, in order to find indications that this may be happening.

A comparison of age groups can shed light on what elements learners, even at similar CORYL levels, only acquire as they mature. One such area appears to be the acquisition of conjunctions/adverbial conjuncts. Hasselgreen and Moe (2006) investigated the use of conjunctions in the writing of Norwegian pupils aged 12/13 and 15/16, on levels A2 and B1, and concluded that regardless of level, the younger pupils used a distinctly narrower range of conjunctions than the older pupils. This is perhaps unsurprising, given that studies such as Nippold (2006) cited in Hasselgreen & Caudwell (2016), found that the development of conjunctions and adverbial conjuncts in English as a first language does not take off fully until adolescence.

Table 2 contrasts a selection of conjunctions/adverbial conjuncts in CORYL, for the different age groups. With the exception of *because*, it can be seen that those used by the older pupils, though admittedly in small in number, are virtually never used by the 7th graders.

Table 2: a selection of conjunctions/adverbial conjuncts in CORYL, contrasting age groups

	Age: 15/16	Age: 12/13
However	16	0
Since	16	1
Even if	11	0
Although	9	0
Even though	5	0
Therefore	1	0
Because	146	72

Because is used with great frequency by both groups. The ease with which these pupils, even at very low levels on the CEFR use *because* can account for the rather high frequency of complex sentences among these pupils.

It is worth noting that a range of aspects of language development, such as syntactic attainment (e.g. adverbial conjuncts) and lexis (including abstract concepts and derivational morphology), as well as certain aspects of discourse and pragmatics, occur late in childhood, in some cases even extending into young adulthood (Nippold, 2006). When investigating the ability of an L2 learner to use a phenomenon, it can be important to determine whether the learners can use its equivalent in his/her L1.

What Hasselgreen (1994, 237) terms *Lexical teddy bears*—words "we regularly clutch at" and "feel safe with"—may be associated with errors, but can also be seen to be present in language where no actual error occurs; this can be clearly seen in the case of core words, as demonstrated by Hasselgreen (1994). Here all-purpose forms, such as *have, get*, or a *lot*, were shown to be used in contexts where native speakers chose a word more specific to the collocation, such as *acquire, undergo* or *profusely*. This reflects Salido's (2016) findings in his study of support verb constructs, referred to above; Salido noted that SVCs were common in both correct and incorrect collocations in learner language. A study of core words in CORYL might involve searching for the word itself, for example *get,* and finding which words it tends to collocate with. Then these words could be examined to see which alternative words they occur with, looking for a variation in these across CEFR levels. This investigation has not been undertaken at this stage. What has been noted is the extremely high frequency of occurrences of *get,* and the virtual absence of its synonyms as given in the Oxford Dictionary. Table 3 illustrates this:

Table 3: occurrences of get and synonyms in CORYL

Word	Number of occurrences in CORYL
Get/got	446
Acquire(d)	0
Receive(d)	1
Fetch(ed)	0
Achieve(d)	0

A result such as this is probably best followed up by experimenting with pupils, to see how adept they are at using appropriate synonyms of core words.

5. Conclusion

This article has hopefully demonstrated that the CORYL corpus has the potential to add significantly to our knowledge of the development of the English of young learners in Norway. We may begin by searching for phenomena we intuitively know to exist, for example, looking for prepositions used wrongly; or by allowing the corpus to surprise us, for example in the list of 'wrong words' or 'wrong function words' found in the corpus. Some of these may seemingly be due to interference, others to developmental factors such as learner progress across CEFR levels or related to maturity. It can also shed light on non-erroneous language forms. However, as has been stated several times in the article, corpus findings on their own are rarely sufficient for us to draw absolute conclusions, and should be followed up, for example in classroom-based studies.

The article is not based on full-scale studies undertaken using CORYL; such studies are in their infancy at the time of writing. It has certainly not provided absolute answers regarding the way written English as an L2 in Norwegian schools develops. Many aspects of language, such as vocabulary development, have not been touched upon, being largely beyond the capacity of a small and limited corpus such as CORYL. It has however shown some directions CORYL can take us in an effort to learn more about the language of these learners. It is informed by what we, as compilers of the corpus, have been struck by over the time we have worked with it. Virtually every search has been an 'aha- experience' in some way. Just playing with CORYL has been inspirational, and one hopes that the reader will be tempted to spend some time doing some searching of their own in this unique and fascinating body of language.

References

Callies, M. 2015. "Learner corpus methodology." In *Learner Corpus Research*, editied by S. Granger, G. Gilquin and F. Meunier, 35–56. Cambridge: Cambridge University Press.

Ellis, R. and G. Barkhuizen. 2005. *Analysing Learner Language*. Oxford: Oxford University Press.

Granger, G. Gilquin and F. Meunier (Eds). 2015. *Learner Corpus Research*. Cambridge: Cambridge University Press.

Hasselgreen, A. and E. Moe. 2006. "Young Learners' Writing and the CEFR: Where Practice Tests Theory." Paper presented at the 3rd EALTA conference, Krakow, May 18–21.

Hasselgreen, A. and G. Caudwell. 2016. *Assessing the Language of Young Learners: British Council Monographs,* vol. 1. Sheffield: Equinox.

Hasselgreen, A. 1994. "Lexical Teddy Bears and Learner Language." *International journal of Applied Linguistics* 8 (2):237–60.

Kellerman, E. and M. Sharwood Smith (Eds). 1986. *Crosslinguistic Influence in Second Language Acquisition*. Oxford: Pergamon Press.

Lightbown, P.M. and N. Spada. 2006. *How Languages are Learned*. Oxford: Oxford University Press.

Ludeling, A. and H. Hirschmann. 2015."Error Annotation Systems". In *Learner Corpus Research*, edited by S. Granger, G. Gilquin and F. Meunier, 135–58. Cambridge: Cambridge University Press.

Manning, C. 2011. "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In *CICLing 2011, Part 1. LNCS 6608*, 171–89. Berlin Heidelberg: Springer Verlag.

Nippold, M.A. 2006. *Later Language Development*. Austin, TX: Pro-ed.

Odlin, T. 1989. *Language Transfer and Cross-linguistic influence in Language Learning*. Cambridge: Cambridge University Press.

Osborne, J. 2015. "Transfer and Learner Corpus Research." In *Learner Corpus Research*, edited by S. Granger, G. Gilquin and F. Meunier, 333–56. Cambridge: Cambridge University Press.

Pienemann, M. and M. Johnston. 1987. "Factors Influencing the Development of Fluency." In *Applying Second Language Research*, edited by D. Nunan, 45–141. Adelaide: National Curriculum Research Centre.

Ringbom, H. 1983."Borrowing and Lexical Transfer." *Applied Linguistics,* 4 (3):207–12.

Salaberry, M.R. 2000. "The Acquisition of English Past Tense in an Instructional Setting." *System* 28:135–52.

Salido, M.G. 2016." Error Analysis of Support Verb Constructions in Written Spanish Learner Corpora." *The Modern Language Journal*, 100 (1):362–76.

Selinker, L. 1972. "Interlanguage". *International Review of Applied Linguistics in Language Teaching,* 10 (3):209.

Appendix: CORYL error tags
(taken from the CORYL website: clarino.uib.no/korpuskel/)
Below you will find a list of all the error code tags used to annotate CORYL, with a brief description of what they stand for and how we applied them.

Code	Stands for
SP	Spelling (correct form supplied)
APOS	Apostrophe (correct form supplied)
CM	Compound, e.g. holiday cottage (correct form supplied)
SEP	Separation of word into 2 parts (correct form supplied)
L1	Norwegian word
SMS	SMS or slang, e.g. L8 or wanna
V	Wrong form of verb, excl. concord. This includes a wide variety of wrongly formed/used verbs, and also cases where a word wrongly used as weak regular verb, with –ed ending, e.g. *comed* (also tagged as NW in some cases).
VC	Concord error, e.g. he were
MV	Modal verb used wrongly
WFO	Wrong form of word (excl. verb), e.g. plural or –ly ending
NW	Nonexistent or unintelligible word (excluding clear verb- or word-form errors)
MFS	Missing full stop
NONSE	Nonsensical/ impossible to understand sentence
SE	Incomplete or multiple sentence, e.g. linking with commas
ART	Any clear article error (overuse, underuse or wrong choice of article)
PREP	Wrong preposition
SY	Anything other than the above categories that affects the syntax of the clause, e.g. word order or prepositions that are missing or added. This also includes missing copula 'to be', e.g. *he coming*.

WW	A wrong lexical word (i.e. not of a closed class)
WFU	A wrong function (i.e. non-lexical) word, such as a pronoun, possessive (e.g. my) demonstrative (e.g. these), conjunction (e.g. and) other linking word (then, when, because, so, as, that, like, [...]). This includes 'the' used in wrong translation of 'det' (not as article). Also *much/many/a lot*
CAP	Missing or overused capital, e.g. i
IT	It/there errors
WPH	A phrase that is wrong, most often with multiple, complex errors, but where it is virtually impossible to decide exactly what is wrong or why.
WI	Wrong idiom. A pupil has used a phrase that appears to be an idiomatic one, but which does not exist in Eng or Nor. This also includes 'good' idioms which are used wrongly in the context, and appear to have been misunderstood.
L1P	L1 phrase. When the whole phrase has a Norwegian formulation translated, e.g. *I hope you have it nice, very fun,* [...]