The use of the general nouns *people* and *thing* by L2 learners of English

– A corpus-based study

Göran Gerdin
Abstract

With the advent of corpora documenting learner English, a new and interesting field of research has become available. Learner corpora provide a new type of data which can inform thinking both in second language acquisition research and in foreign language teaching research. Analyses of learner corpora normally report on features which are typically ‘overused’ and ‘underused’, when contrasted to comparable native speaker corpora, in addition to those which are ‘misused’ by the learners. Ringbom (1998) conducted a study in which he identified one common aspect of non-native speaker corpora: the high frequency of general nouns, such as people and thing.

The aim of this paper was to test Ringbom’s findings and attempt to identify how English as a second language learners’ usage of these particular nouns in written production differ from that of native speakers by conducting a corpus comparison of comparable learner and native speaker corpora. The results of this study clearly support Ringbom’s findings; additionally, it was found that the learners’ written production does not appear vaguer and ‘non-native like’ merely because they overuse the general nouns people and thing, but it also seems as if the learners use these nouns in a more restricted range of meanings whereas the natives’ usage is more diversified. Moreover, this study has identified some of the issues that teachers of English as a second language should be aware of when helping their students to avoid using the general nouns people and thing in a non-native like manner.
Contents

1. Introduction .............................................................................................................................4
   1.1 Background ......................................................................................................................................4
   1.2 Aim and Scope ................................................................................................................................6
   1.3 Outline .......................................................................................................................................7

2. Literature Review ....................................................................................................................7
   2.1 Corpora ........................................................................................................................................7
      2.1.1 Frequency .................................................................................................................................. 8
      2.1.2 Phraseology .............................................................................................................................9
      2.1.3 Collocations ..........................................................................................................................10
   2.2 Corpus Types ............................................................................................................................11
   2.3 Learner Corpora .......................................................................................................................11
   2.4 Learner Corpora and Language Teaching .............................................................................14

3. Data Description ....................................................................................................................16
   3.1 The ICLE Corpus .....................................................................................................................16
      3.1.1 Learner Variables ..................................................................................................................17
      3.1.2 Task Variables .......................................................................................................................17
      3.1.3 Size and Representativeness ...............................................................................................17
      3.1.4 PICLE & SWICLE ...............................................................................................................18
   3.2 The Native Speaker Corpora ................................................................................................18

4. Methodology ..........................................................................................................................19
   4.1 General Nouns .........................................................................................................................19
   4.2 Meanings of people and thing ...............................................................................................20
   4.3 Data Collection and Coding .................................................................................................21

5. Results .....................................................................................................................................23

6. Discussion ................................................................................................................................28
   6.1 Frequencies ..............................................................................................................................28
   6.2 Collocates and Meanings .......................................................................................................32

7. Conclusions ............................................................................................................................38

8. Implications for Teaching ......................................................................................................39

9. References ...............................................................................................................................43
1. Introduction

This study examines the insights to be gained into the nature of second language acquisition from the analysis of ‘learner corpora’, which are digital representations of the performance or output, typically written, of language learners. With the advent of corpora documenting learner English, a new and interesting field of research has become available. There are many aspects of English grammar and discourse that can fruitfully be explored in learner corpora to shed light on both practical and theoretical questions with applications in the teaching of English as a second language. One such area is the usage of high generality vocabulary items, also known as ‘general nouns’.

1.1 Background

Learner corpora provide a new type of data which can inform thinking both in second language acquisition (SLA) research, which tries to understand the mechanisms of SLA, and in foreign language teaching (FLT) research, the aim of which is to improve the learning and teaching of foreign languages (Granger 2002:5). A widespread approach adopted by several learner corpora researchers is to work from a common observation or impression about learner language, develop a hypothesis to explain the observation, and test the hypothesis through a comparison of learner and native speaker corpora (Cobb 2003). Analyses of learner corpora normally report on features which are typically overused and underused, when contrasted to comparable native speaker corpora, in addition to those which are misused by the learners (Leech 1998). Ringbom (1998), for instance, identified one common aspect of non-native speaker corpora: the high frequency of vocabulary items of high generality, such as *people* and *thing*. These words are often used by learners of English as a second language (L2) in situations where native speakers usually employ another, less ambiguous word, as (1) produced by one of the Polish L2 learners illustrates:

(1) Child-clones for sale would be a blessing for infertile *people* craving for an offspring for long. (PICLE)

Although native speakers might use *people* in this case in speech, it is most likely that they in written production would rather use an alternative word, such as *couples* or *individuals*. This was found to be true when a native speaker asked to write an appropriate word after *infertile* in the above sentence, did in fact use *couples*. In (2) below another Polish L2 learner uses *thing* where a
native speaker most probably, at least in writing, would rather have used another word, such as *substance* or *drug*.

(2) Addicts simply cannot do without the *thing* they are addicted to. (PICLE)

According to Hunston (2002), this type of overuse makes the learners’ production appear somewhat vague when compared to native speaker production. Even at apparently native level it is often observed that learners’ writing remains vague or resembles native speaker speech written down more than it does native speaker writing (Cobb 2003). However, the question of what this overuse of *people* and *thing* looks like has not yet been fully answered. Are particular meanings of *people* and *thing* used more often by L2 learners of English than by native speakers? Can an investigation of learner use show whether there is a need for teachers of English as a second language to come up with new, more pedagogical explanations of the different meanings of these vocabulary items? The role of corpora in language teaching is not “to tell us what we should teach, but they can help us make better-informed decisions, and oblige us to motivate those decisions more carefully” (Gavioli and Aston 2001:239).

There are many studies now that show that the language or grammar described in English as a second language (ESL) textbooks does not correspond to the language used by native speakers (Barlow 2000). Since textbooks are pedagogical in nature, their content will never match exactly the language used by native speakers and it is clear that there are many factors involved in the content of textbooks such as sequencing, repetition, etc. However, the mismatch appears to be due to the combination of a lack of knowledge of language as it is used and an adherence to an ESL textbook tradition of topics to be taught and grammar points to be covered. The textbook descriptions are not necessarily wrong but the focus is often misplaced. The less frequent words or constructions may be highlighted and more frequent uses ignored (Barlow 2000). An analysis comparing non-native speaker and native speaker usage of the general nouns *people* and *thing* can therefore aid teachers of English as a second language in providing an overall view of how native speakers actually use these words, thus giving them a better idea of how they can teach their students to use these words more ‘native-like’.
1.2 Aim and Scope

The main aim of this investigation was to look at Polish and Swedish learners’ of English use of *people* and *thing* to find support for or refute Ringbom’s (1998) claim that non-native speakers use these two high generality vocabulary items more frequently than native speakers do. Furthermore, it is examined in what way non-native speakers and native speakers use these words, to see whether any discrepancies in usage can be identified. In order to accomplish this, the Polish (PICLE) and the Swedish (SWICLE) subparts of the International Corpus of Learner English (ICLE) were contrasted to comparable native speaker corpora. For each corpus 200 random tokens were chosen and analysed accordingly in an attempt to answer the following questions:

*Research Question 1.* Do Swedish and Polish learners of English as a second language overuse or underuse (in comparison with native speakers) the high generality vocabulary items *people* and *thing*?

*Research Question 2.* In what ways does their usage differ from that of native speakers and does this make their language production seem rather vague when compared to native speakers?

It was suggested, as proposed by Ringbom (1998), that most learners of English as a second language use the words *people* and *thing* more frequently than native speakers, since they lack the skills to use other, less vague words and therefore learners instead demonstrate a writing style more resembling spoken native language written down rather than traditional native writing. In addition, non-native speakers’ written production is often more informal and personal in comparison with native speakers’ more formal and academic writing style. Learners use features associated with high writer/reader involvement, such as *people* and *thing*, making their written production more interactive than is typical in native speaker written prose (Hunston 2002). It was therefore hypothesised that:

*Hypothesis 1.* Learner essays display a higher degree of personal involvement than native speaker essays of equivalent genre, since learner writing resembles native speaker speech written down more than it does native speaker writing.
Moreover, it was proposed that L2 learners produce these two vocabulary items in a restricted range of meanings compared to native speakers, and thus limiting the complexity of their production. The learners lack the knowledge of the various senses of *people* and *thing*, hence restricting the range of meanings they can make with these two words. As a result they end up overusing these vocabulary items of high generality in those meanings which are known to them. Consequently, it was also hypothesised that:

**Hypothesis 2.** L2 learners of English rely on the restricted, context-determined lexicon of spoken language rather than deploying the broader lexicon of native speaker writing.

When looking at the usage of *thing*, it was decided that the plural of the word, *things*, was not to be included in the analysis since its usage is quite different from that of the singular form.

1.3 Outline

The literature review provides the reader with necessary background information to set the study and the discussion presented in this paper in its context. In the methodology section, background information on the corpora used in this study is provided along with particulars of the data analysis technique. Results of the non-native speaker corpora and native speaker corpora comparison are presented and subsequently discussed, conclusions are drawn and implications for teaching suggested.

2. Literature Review

2.1 Corpora

Strictly speaking, a corpus by itself can do nothing at all, being nothing other than a store of used language. Corpus access software, however, can re-arrange that store so that observations of various kinds can be made. If a corpus represents, very roughly and partially, a speaker’s experience of language, the access software re-orders that experience so that it can be examined in ways that are usually impossible. A corpus does not contain new information about language, but the software offers us a new perspective on the familiar.

(Hunston 2002:3)

Linguists have always used the word *corpus* to describe a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or
tape recordings, which have been collected for linguistic study (Partington 2001). Crystal (Crystal 1991:86) defines a corpus as a “collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language”. More recently, the term has been reserved for collections of texts (or parts of text) that are stored and accessed electronically (Hunston 2002). Leech (1997), one of the leading figures in corpus linguistics, therefore slightly modified Crystal’s definition in saying that this collection of linguistic data should exist in an electronic form so it is possible for a computer to process the information since corpora consist of a large amount of data. In the later version, a corpus can therefore be seen as any text-only file, or set of text-only files, that can be loaded into a corpus access software (Barlow 2003). The benefits of storing texts electronically come from the accessibility that this format provides and from the potential for interchange and transmission of data (Barlow 2004). The use of a computer-based corpus software then enables the researcher to process data from the corpus in three ways: showing frequency, phraseology, and collocation (Hunston 2002).

2.1.1 Frequency

Frequency is an aspect of language of which we have very little intuitive awareness, but one that plays a major part in many linguistic applications which require a knowledge not only of what is possible in language but what is likely to occur (Granger 2002). By arranging the words in a corpus in order of their frequency one can compare these frequency lists between different corpora. Identifying possible differences between corpora in this way allows for issues to arise that can then be studied in more detail (Hunston 2002). Furthermore, by using more advanced software which counts not only words but also categories of linguistic items it is possible to look at grammatical features such as the past and the present tense. Biber et al. (1999) compared the frequency of these two tenses across four genres: conversation, fiction, news and academic. They found that in the genres of conversation and academic prose, the present tense is more frequent than the past tense. In fiction, the past tense was more frequent than the present tense, and in the news genre both tenses were equally frequent. However, those opposed to using corpus evidence in teaching argue that certain aspects of English are important even though they are not frequent, either because they carry a lot of information or because they have a resonance for a cultural group or even for an individual (Hunston 2002). As for the English language, for instance,
knowing the meaning of *thou*, in the tenth commandment “Thou shalt not kill”, is still important even though this particular word does not occur frequently in today’s English.

2.1.2 Phraseology

Corpus studies have demonstrated that routine phraseology is pervasive in language use (Stubbs 2004). Although native speakers can often recognise if a phraseological pattern is unusual, articulating the nature of the atypicality may be more difficult. Intuition is therefore a poor guide to this aspect of language (Barlow 2000). To understand language use more accurately it is necessary to back up native speaker intuition by evidence from a corpus. By using a corpus access software it is possible to search for a word or a phrase and then look at the so-called ‘concordance lines’. Concordance lines allow the user to observe regularities in the use of a word or phrase that cannot be done by using intuition (Hunston 2002). Therefore, phraseology can be used as an alternative view of phenomena that teachers of English are frequently called upon to explain. However, as pointed out by Stubbs (2004), a corpus can only tell us which phrases are the most frequent, and does not offer an explanation of why they are frequent, consequently one must also analyse the contexts in which the phrases occur. According to Stubbs, many phrases are very frequent since they are regular ways of expressing common meanings. He therefore argues that it is not surprising that phrases for place, time, and intention are amongst the most frequent in the language, since they make it possible for language users to reconstruct sequences of events in a comprehensible way. These frequently occurring phrases express relations which are essential for understanding connected discourse (Stubbs 2004).

Furthermore, Stubbs (2005) argues that some words are very frequent because they occur in frequent phrases. By searching the British National Corpus, he finds that the word *world* ranks among the top ten nouns, occurring more than 500 times per million words. According to Stubbs it is rather unlikely that people keep referring that often to the external world, and therefore investigates further in what contexts that *world* occurs. Although some of the phrases in which *world* occurs do refer externally to the world, as in “he traveled around the world”, it is made apparent when looking at a larger proportion of cases that the literal sense of *world* is not that common. Stubbs (2005) discovers that *world* occurs frequently in semi-fixed phrases, such as “Second World War”, and in evaluative phrasal constructions (e.g. “one of the world’s most gifted scientists”).

Furthermore, Stubbs (2005) argues that some words are very frequent because they occur in frequent phrases. By searching the British National Corpus, he finds that the word *world* ranks among the top ten nouns, occurring more than 500 times per million words. According to Stubbs it is rather unlikely that people keep referring that often to the external world, and therefore investigates further in what contexts that *world* occurs. Although some of the phrases in which *world* occurs do refer externally to the world, as in “he traveled around the world”, it is made apparent when looking at a larger proportion of cases that the literal sense of *world* is not that common. Stubbs (2005) discovers that *world* occurs frequently in semi-fixed phrases, such as “Second World War”, and in evaluative phrasal constructions (e.g. “one of the world’s most gifted scientists”).
2.1.3 Collocations

The term ‘collocation’ was first coined in its modern linguistic sense by the British linguist J.R. Firth along with the famous explanatory slogan “You shall judge a word by the company it keeps” (Partington 1998:15). Collocations can be described as the statistical tendency of words to co-occur (Hunston 2002). Collocation can indicate pairs of lexical items, such as bread + butter and shed + tears, or the association between a lexical word and its frequent grammatical environment, then called ‘colligation’. Examples of colligations including the lexical word head are for instance: head of (e.g. “head of department”), head on (e.g. “meet something head on”) and head off (e.g. “head off towards somewhere”). A list of collocates of a given word can yield similar information to that provided by concordance lines, with the difference that more information can be processed more accurately by the statistical operations of the computer than can be dealt with by the human observer (Barlow 2004).

Searching the BNC we find that the most common collocates of thing are: is, to, and that. Looking at the usage of thing this would suggest that thing is often used to make the reader or the listener aware of something or draw their attention to something important, e.g., “the thing is, we cannot wait any longer before we…”, “the thing to remember…” or “the thing that makes deals happen is…”. The most common collocates of people are also grammatical, namely who, in, and are. “People who…” and “People are…” indicate that people is often used to talk about what people in general, for instance, do or think. The high frequency of “people in…” implies that people is also often used to refer to people living in, for instance, a certain country, i.e., “people in India believe that…”, however it could also be part of phrases such as “people in business” and “people in their mid-fifties” etc.

Researchers interested in what role collocations play in language learning have suggested that the storage capacity of human memory is vast, but that the speed for processing those memories is not. In order to make the processing more efficient we must learn short cuts, and that is why language consists of a relatively high number of prefabs. “We store a large number of complex items which we manipulate with comparatively simple operations” (Ladefoged 1972:282). This claim can be directly linked to the SLA theory of ‘connectionism’ which is concerned with the simple learning mechanisms which operate when complex language representations are processed by the learner (Ellis 1998). For teaching this would imply that
learning a language is not just about acquiring vocabulary and grammar, but also knowing which words tend to co-occur more frequently than others, and what prefabs that are commonly used.

2.2 Corpus Types

What the corpus consists of determines what we might say about the results of a search for a word or phrase. “In other words, a corpus has to represent something” (Barlow 2003:19). In designing a corpus, the size and representativeness of a corpus affect the validity and reliability of the research (Sinclair 1991). The representativeness of a corpus depends on the quality of the composition of the corpus. The composition should be determined by the purpose of the research (Chung 2003). Some commonly used corpus types are:

- Specialised corpus is used to investigate a particular type of language.
- General corpus may be used to produce reference materials for language learning or translation, and is often used as a baseline in comparison with more specialised corpora.
- Learner corpus used to identify in what respects learners differ from each other and from the language of native speakers, for which a comparable corpus of native-speaker texts is required.
- Parallel corpora parallel texts that are translations of each other, used by translators and by learners to find potential equivalent expressions in each language and to investigate differences between languages.

(Hunston 2002:14-15)

2.3 Learner Corpora

Corpus linguistics made its debut on the linguistic scene in the late 1950s and since then its major contribution has been in the field of variation studies (Granger 2003). “The diversification of corpora has given linguists a firm basis for comparing language varieties distinguished in terms of the medium (spoken vs. written), the field (general vs. specialised), and geographical status (World Englishes)” (Granger 2003:538). However, for many years second language learner varieties remained noticeably missing from corpus-based research. It was not until the 1990s that learner data started being collected and analysed (Granger 2003). Recent advances in software development now make it possible to analyse large databases of learner language, both from a ‘bottom-up’ perspective (to find patterns in data) and from a ‘top-down’ perspective (to test hypotheses) (Rutherford and Thomas 2001).
Learner corpus projects can be seen as a natural extension of the interest in language sampling. They were launched in the early nineties, partly to satisfy a need to verify or refute claims about transfer from the mother tongue to the foreign language (Horvarth 2001). A learner corpus is collected for a particular SLA or FLT purpose. Researchers may want to test or improve some aspect of SLA theory, for example by confirming or disconfirming theories about transfer from the learners’ first language (L1) or the order of acquisition of morphemes, or they may want to contribute to the production of better FLT tools and methods (Granger 2002). Design criteria are very important in the case of learner data because there is so much variation in English as a foreign language (EFL)/English as a second language (ESL). A random collection of heterogeneous learner data does not qualify as a learner corpus. Learner corpora should be compiled according to strict design criteria, some of which are the same as for native corpora, while others, relating to both the learner and the task, are specific to learner corpora. The usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables. Variables could include: learning context, mother tongue, level of proficiency, time limit, exam etc. (Granger 2002).

The most prominent figure within this field of corpus linguistics is without any doubt Sylviane Granger (Hunston 2002) who is also the founder of the most extensive and well-known compilation of learner corpora, the International Corpus of Learner English (ICLE) (Barlow 2005). The ICLE consists of 500-word non-technical argumentative essays produced by advanced learners of English as a foreign language. The ICLE database is also divided into different corpora, sorted according to the learners’ first language (Hunston 2002).

Research on learner corpora is most of the time contrastive in its nature and follows the procedures associated with ‘contrastive analysis’ (Barlow 2005). The essence of work on learner corpora is comparing corpora produced by various groups of non-native speakers, as well as looking at differences between non-native speaker corpora and native speaker corpora (Hunston 2002). Granger (1998) refers to this as a form of ‘contrastive interlanguage analysis’ where the native speaker corpora serve as a point of reference for the analysis of the non-native speaker corpora and hence provide evidence for the nature of ‘interlanguage’ (Barlow 2005). Interlanguage in second language acquisition can be defined as the linguistic system characterising the output of a non-native speaker at any stage prior to full acquisition of the target language (Barlow 2005). Focusing on error analysis and interlanguage, learner corpora, such as
ICLE, enable researchers and educators to directly analyse and compare the written output of second language learners (Horvarth 2001). Learner corpus research offers further refinement in identifying those forms which are problematic for learners (Meunier 2002). In addition, comparing different non-native speaker corpora, aspects of language use common for learners with different language backgrounds can be highlighted. When discrepancies among learners with different language background can be identified, the influence of, for instance, the learner’s L1 can be further investigated (Granger 2002). One of the biggest contributions of learner corpora analysis has been the directing of attention to observing learner language so that the notion of L1 transfer may be analysed under stricter forms (Horvarth 2001). Taking the learners’ mother tongue into account would aid teachers of second languages to provide more focused and appropriate teaching methods (Meunier 2002).

Studies conducted by Granger of the ICLE corpora “are quantitative rather than qualitative in nature, but there are interesting qualitative generalizations to be made” (Hunston 2002:207). For example, learner corpora often show a greater use of a smaller range of vocabulary items (Hunston 2002). Aijmer (2002) compared the frequency of modal words used by advanced Swedish learners of English with British native speakers. Using subparts of the ICLE corpora, which included collections of Swedish and British essays, she showed that Swedish learners use particular modal auxiliaries (will, would, must, have (got) to, should, might) and adverbs (probably, maybe, of course, certainly) significantly more frequently than British native speakers. Aijmer suggests that this may be due in part to the fact that Swedish makes use of combinations of modal verbs and adverbs to a greater extent than English. Another aspect of learner corpora which can be recognized is the high frequency of vocabulary items of high generality, such as people and thing (Ringbom 1998). Consequently the learners’ production can appear somewhat vague when compared to native speaker production (Hunston 2002). For this to be translated into pedagogic issues, the teacher could not simply say “Use thing less often” (Hunston 2002:208); instead the teacher would need to know the precise circumstances when native speakers would typically choose an alternative to thing, and what the alternative would be.

According to Cobb (2003:400) learner corpus research “amounts to a new paradigm, and a great deal of methodological pioneering remains to be done”. The relative youth of learner corpus research combined with the need for a wider range of learner corpora, covering a range of proficiency levels and a number of L1-L2 combinations, calls for cautiousness regarding the
generalisations we can make from analysing learner corpora. In order to achieve comparability of data across research studies and thus draw a more robust picture of advanced second language use, Granger and Tyson (1996) suggest that the following four variables should be controlled: type of learner (e.g., foreign vs. second language), stage of learner, text type, and the availability of a similar corpus of native speaker data. Cobb (2003:403) reports that it is characteristic of learner corpus methodology to extrapolate from cross-sectional to longitudinal data in order to address developmental issues with respect to learner interlanguage. In his own replication study, Cobb (2003) constructed such a graded corpus that consisted of similar essays written by different groups of ESL learners at beginning, intermediate and advanced levels of proficiency in order to answer developmental questions about their over- and underuse of particular high frequency lexemes and phrases. However, he readily recognizes the inherent disadvantages in this approach and suggests instead that “one would ideally [need to] have recourse to large writing samples from [the] same or equivalent learners over the years of their remaining studies” (Cobb 2003:401) in order to accurately elucidate developmental pathways. To fully understand the process of second language acquisition, it would be beneficial to carry out longitudinal studies of learner production so that the paths of language development can be understood better (Barlow 2005). The availability of digitized, longitudinal data for individual learners may push the efficacy of contrastive corpus analysis away from the primarily correlational role to which it has been relegated and toward a more reason- oriented, explanatory one, although it is unlikely in general that language learning rests on such one-to-one causal relationships as learner use preceded by expert speaker use (Belz 2004).

2.4 Learner Corpora and Language Teaching

Mark (1998) points out that traditionally language teaching approaches have dealt mainly with three factors: describing the target language, characterizing the learners (motivation, learning styles, aptitude etc.), and instruction (through task, syllabus and curriculum). However, Mark argues that “it simply goes against common sense to base instruction on limited learner data and to ignore, in all aspects of pedagogy from task to curriculum level, knowledge of learner language”.

Granger and Tribble (1998) provide examples of material using a learner corpus and a comparative native speaker corpus. In these examples likely sources of error are first identified by comparing the two corpora in statistical terms, and then comparable concordance lines are
used to encourage the learners to be aware of the differences between native and non-native usage. This could then be directly linked to the theory of ‘noticing’, or ‘consciousness/awareness raising’, advocated by scholars such as Schmidt. In his ‘Noticing Hypothesis’ Schmidt (1994) points out the vital importance for noticing in language learning. Furthermore, he claims that this noticing may in turn stimulate the processes of language acquisition. However, some linguists have fundamental objections to this type of comparison because they believe that interlanguage should be studied in its own right and not as somehow deficient as compared to the native norm (Granger 2002). One way of dealing with this, as mentioned by Hunston (2002), could be to compare, for example, Swedish schoolchildren’s usage with expert Swedish speakers of English rather than with British speakers of English.

Another interesting way of using learner data was adopted by Seidlhofer (2002) in a teaching experiment, where she had learners write a summary of a text and a short personal reflection on it. What the learners had produced she later used as the primary objects of analysis, and thus getting the learners to work with and on their own output. Seidlhofer concludes the experiment by reporting that it was particularly successful because the learners were forced to be more active and responsible for their own learning than what they normally would be. In addition, this method employed by Seidlhofer (2002) shows that teachers can collect their own small corpora, to supplement larger learner corpora such as ICLE (Granger 2002). This material can then either be used as suggested by Seidlhofer (above) or for form-focused instruction.

There is great potential for using corpus-based material in language teaching. Corpus-based investigations can help both teachers and students reveal complexities and patterns in languages that tend to be missed in traditional intuition-based analyses (Barlow 2000). Consequently teachers will also increase their own knowledge of the language they are teaching. Corpus data can serve as a powerful tool with which learners can discover the foreign language on their own and the role of the teacher becomes that of materials-provider. In addition, by incorporating corpus-based material in language teaching teachers can feel more confident that the language presented to the learners is directly relevant to the language used outside the classroom, since corpora consist of authentic language use (Hunston 2002).
3. Data Description

3.1 The ICLE Corpus

A computer learner corpus (CLC) is an electronic collection of authentic texts produced by foreign or second language learners. Although all corpora need to be assembled according to explicit design criteria (Atkins and Clear 1992) extra care has to be taken in collecting the data for learner corpora given the large number of variables affecting the learning/acquisition process. In this study the ICLE corpus was used, which is a very richly documented corpus. More than 20 task and learner variables have been recorded for each of the texts in the corpus through a detailed profile questionnaire completed by all learners. As shown in Figure 1, some of these variables (medium, genre, average length, learner proficiency level) were used as corpus design criteria and are therefore shared by all texts in the corpus whereas others (gender, mother tongue background, essay topic) differ from text to text.

![Figure 1: Shared and variable features of the ICLE corpus](image)

(Granger 2003:539)

**Figure 1 Shared and variable features of the ICLE corpus**

All the variables have been stored in a database and can be used by researchers as queries to compile subcorpora that match certain criteria, thus allowing for interesting comparisons (e.g., female vs. male learners, Polish- vs. Swedish-speaking learners).
3.1.1 Learner Variables

The learners who have contributed data to the ICLE have a great deal in common. They are all about 20 years old and study English in a non-English-speaking country, thus making them EFL rather than ESL learners. All learners are university undergraduates specialising in English in their second, third, or fourth year, and their level can be roughly described as advanced, although individual learners and learner groups differ in proficiency. In spite of their similarities in terms of age, L2 status, and proficiency level, the learners display some crucial differences, the most important one being mother tongue. The ICLE corpus consists of 11 different mother tongue backgrounds: Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. Another variable with a potentially significant impact on learner output is the amount of time learners have spent in an English-speaking country. ICLE learners differ considerably in this respect: 40% have never stayed in an English-speaking country whereas some 30% have lived in an English-speaking environment for 3 months or more. Yet another relevant variable is gender. The corpus consists of data from both male and female learners, although the latter clearly make up the majority (80%).

3.1.2 Task Variables

The ICLE corpora have many task attributes in common. They contain exclusively written productions of a specific genre, namely, essay writing, and represent general English rather than English for specific purposes. The essays are, on average, 700 words in length. The topics differ tremendously, even though the majority of the essays (85%) are argumentative. The rest of them are literary essays. The essays also include certain dissimilarities in task settings. Recorded variables consist of whether there was a time limit for writing, whether the essay was part of an exam, and whether the learners were allowed to use language reference tools such as grammars or dictionaries (Granger 2003).

3.1.3 Size and Representativeness

The ICLE corpus consists of 3,640 essays, totaling 2.5 million words. Each of the 11 subcorpora contains around 330 essays totaling approximately 200,000 words. In comparison with large corpora, such as the British National Corpus (BNC), the ICLE is very small. However, when it comes to learner language, size cannot simply be assessed in terms of the number of words. Equally important is the number of learners, and, in this respect, the ICLE, which comprises
writing by well over 3,000 learners, makes up a solid empirical basis for SLA and FLT research (Barlow 2005). However, because of its limited number of words, the ICLE cannot be used for all types of linguistic investigation. It is well suited for the analysis of high-frequency phenomena at all linguistic levels (morphology, grammar, lexis, discourse) but is not suitable for the investigation of, for instance, infrequent linguistic items (Granger 2003). Table 1 (below) shows a summary of the various sizes of the corpora used in this study.

Table 1 Size of the corpora used in this study

<table>
<thead>
<tr>
<th>Corpus Size</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>PICLE</td>
<td>330,000</td>
</tr>
<tr>
<td>SWICLE</td>
<td>205,000</td>
</tr>
<tr>
<td>Newspaper Corpus</td>
<td>94,000</td>
</tr>
<tr>
<td>BNC – University Essays</td>
<td>55,717</td>
</tr>
<tr>
<td>BNC - School Essays</td>
<td>146,530</td>
</tr>
</tbody>
</table>

3.1.4 PICLE & SWICLE

The two learner corpora that were used in this study are called ‘PICLE’ and ‘SWICLE’, which are the Polish and Swedish subcorpora of ICLE. PICLE consists of about 500 essays produced by advanced learners of English as a foreign language, totaling approximately 330,000 words. This Polish section of ICLE was accessed online through a web-based search engine (Kaszubski 2006). SWICLE comprises predominantly argumentative essays written by Swedish university students. The corpus includes 355 essays totalling around 205,000 words. The average size of the essays is about 560 words. SWICLE was searched using the ICLE CD-ROM version produced by Granger et al. (2002).

3.2 The Native Speaker Corpora

For the corpus comparison two different native speaker corpora were used: the online version of the 100 million words British National Corpus (BNC) and a newspaper corpus consisting of UK and US quality press editorials and popular science book excerpts containing approximately 94,000 words. The BNC is a collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written (Hunston 2002). Within the BNC, a sub-corpus consisting of essays written
by university students was chosen to make it possible to compare the usage of the word within
the same genre. The BNC was searched using an online corpus concordancer (Davies 2006).

Furthermore, as pointed out by Hunston (2002), it also worth comparing the learner
corpora to other type of registers, in this case journalism in form of the newspaper corpus, to
provide a more overall picture of the learner’s language production compared to native speakers.
The newspaper corpus was retrieved from the same webpage as the PICLE (Kaszubski 2006). In
addition, it was also decided to look at school essays written by native speakers in order to
compare these to the essays written by university students, and by doing so aiming to identify
some trends for the use of people and thing among native speakers as they get more proficient
and relate that to second language learners. The corpus containing the essays written by school
pupils was also retrieved from the online version of the BNC (Davies 2006).

4. Methodology
The method most frequently employed so far to analyse learner corpora is called ‘contrastive
interlanguage analysis’, which means that the researcher carries out either a comparison of non-
native speaker data with native speaker data (NNS vs. NS) or a comparison between different
types of non-native speaker data (NNS vs. NNS) (Granger 1996). The first type of comparison
makes it possible to uncover the patterns of use distinguishing non-native speaker data from
native speaker data. In this study a non-native speaker and native speaker comparison was made
where, as described above, the non-native speaker data came from PICLE and SWICLE, and the
native speaker data were collected from the BNC and a newspaper corpus. By making such a
comparison it is possible to examine: qualitative differences (misuse) and quantitative differences
(over- and underuse). For the purposes of this study the latter category was employed in order to
investigate Ringbom’s (1998) claim that non-native speakers overuse the high generality
vocabulary items people and thing.

4.1 General Nouns
Both people and thing belong to a group of words often referred to as ‘general nouns’. Halliday
and Hasan (1976:274) define general nouns as “a small set of nouns having generalized reference
within the major noun classes, those such as ‘human noun’, ‘place noun’, ‘fact noun’ and the
like”. People belongs to the ‘human noun’ class together with other nouns such as person and
man. Thing is normally put into the same class as object, and that noun class is commonly known
as the ‘inanimate concrete count’. Halliday and Hasan (1976) claim that general nouns play a significant part in verbal interaction, and are also an important source of cohesion in the spoken language. Moreover, Partington (2001) argues that some references by general nouns are so vast or vague that it is not possible to relate them to any particular part of the surrounding text, and that, in fact, nouns such as thing is often used to avoid being too specific.

4.2 Meanings of people and thing

A corpus is not the best place to look if you simply want to know the definition of a word. Dictionaries and encyclopedias are designed to describe conceptual or denotational meanings, arranging the different senses of a word in some kind of order (Partington 2001). Presented below in Table 2 and Table 3 is an overview of the different meanings of the two words people and thing used for the analysis in this study. Meaning is in this study used to refer to the different senses of a word as outlined in a dictionary, in this case, The Free Online Dictionary (Farlex 2006). In the structuring of the different meanings of people and thing, another dictionary, The Longman Dictionary of Contemporary English (2005), was also used to make sure that the categories were reasonably reliable.

Table 2 Overview of the different meanings of people

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Humans considered as a group or in indefinite numbers</td>
<td>People were dancing in the street</td>
</tr>
<tr>
<td>B</td>
<td>A body of persons living in the same country under one national government; a nationality</td>
<td>The Polish people</td>
</tr>
<tr>
<td>C</td>
<td>Persons with regard to their residence, class, profession, or group</td>
<td>City people</td>
</tr>
<tr>
<td>D</td>
<td>The mass of ordinary persons</td>
<td>People are often afraid of what is not yet known to them</td>
</tr>
<tr>
<td>E</td>
<td>The citizens of a political unit, such as a nation or state</td>
<td>The people of India</td>
</tr>
<tr>
<td>F</td>
<td>Persons subordinate to or loyal to a ruler, superior, or employer</td>
<td>The queen showed great compassion for her people</td>
</tr>
<tr>
<td>G</td>
<td>Family, relatives, or ancestors</td>
<td>Our people have lived here for generations</td>
</tr>
<tr>
<td>H</td>
<td>Animals or other beings distinct from human beings</td>
<td>Rabbits and squirrels are the furry little people of the woods</td>
</tr>
</tbody>
</table>

Source: The Free Online Dictionary (Farlex 2006)
Table 3 Overview of the different meanings of *thing*

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 An entity, an idea, or a quality perceived, known, or thought to have its own existence</td>
<td>The thing I like about her is…</td>
<td></td>
</tr>
<tr>
<td>2 The real or concrete substance of an entity, an entity existing in space and time, or an inanimate object</td>
<td>What is that thing doing over there?</td>
<td></td>
</tr>
<tr>
<td>3 A creature</td>
<td>The poor little thing</td>
<td></td>
</tr>
<tr>
<td>4 An individual object</td>
<td>There wasn’t a thing in sight</td>
<td></td>
</tr>
<tr>
<td>5 An object or entity that is not or cannot be named specifically</td>
<td>What is this thing for?</td>
<td></td>
</tr>
<tr>
<td>6 A thought, a notion, or an utterance</td>
<td>What a rotten thing to say!</td>
<td></td>
</tr>
<tr>
<td>7 A piece of information</td>
<td>He wouldn’t tell me a thing about the project</td>
<td></td>
</tr>
<tr>
<td>8 A means to an end</td>
<td>Just the thing to increase sales</td>
<td></td>
</tr>
<tr>
<td>9 An end or objective</td>
<td>In blackjack, the thing is to get nearest to 21 without going over</td>
<td></td>
</tr>
<tr>
<td>10 A turn of events, a circumstance</td>
<td>The accident was a terrible thing</td>
<td></td>
</tr>
<tr>
<td>11 A particular state of affairs; a situation</td>
<td>Let’s deal with this thing promptly</td>
<td></td>
</tr>
<tr>
<td>12 A persistent illogical feeling, as a desire or an aversion; an obsession</td>
<td>He has a thing about seafood</td>
<td></td>
</tr>
<tr>
<td>13 The latest fad or fashion; the rage</td>
<td>Drag racing was the thing then</td>
<td></td>
</tr>
<tr>
<td>14 An activity uniquely suitable and satisfying to one</td>
<td>Let him do his own thing</td>
<td></td>
</tr>
</tbody>
</table>

Source: *The Free Online Dictionary* (Farlex 2006)

### 4.3 Data Collection and Coding

The two words *people* and *thing* were searched for using the ‘simple word search’ and the ‘co-text size’ was set to left 60 and right 60 in order to capture enough context to define in which meaning the particular word was used. However, in some difficult cases the whole paragraph in which the word originally occurred had to be looked at. The search resulted in a large number of tokens which was not surprising since both words are rather common. Due to the restriction of the web-based search engine, only 50 tokens could be shown at a time, so therefore four different searches for each corpus were performed to make the total number of concordance lines analysed 200. However, in some searches the same tokens appeared again, so to increase validity additional searches were done to replace the recurring tokens with new ones. Furthermore, the first 100 tokens used in the analysis were sorted ‘first right’ and the next 100 hits were sorted ‘first left’ to give an enhanced indication of the usage of the words. By examining the
concordance lines it was possible to observe regularities in the use of the words that cannot be done by using intuition (Hunston 2002).

Firstly, the frequencies of the two words were analysed, which were calculated by how many instances of the words occurred per 100,000 words (see Table 4 and 5 below). Secondly, the different meanings of the words within the 200 tokens were identified and frequencies for each of these meanings were also calculated (see Table 6 and 7 below). Both the frequencies of the words and the different meanings were then compared between the two non-native speaker corpora and the three different native speaker corpora.

As for the coding of the learner data and the native speaker data it was especially difficult to make a distinction between meaning A and D of people (see Table 2). The instances where people was preceded by a determiner such as some or these, or an adjective, it fell under the first category, as demonstrated by (3) produced by one of the Polish learners:

(3) Probably in this case the Orwell's sentence that some people are more equal than others will always be true. (PICLE)

When people, as in (4), was not preceded by any determiner or adjective it was put in category D.

(4) People think in stereotypes. (SWICLE)

This method of distinguishing between meaning A and meaning D was based on the perception that meaning A refers to a, to some extent, limited group of people, which can, for instance, be identified as exemplified in (3) above, by a determiner such as some or these. Meaning D, however, is considered in this study to refer to an unlimited number of people. It is used to express what people in general is, for instance, perceived to behave or think about something, as illustrated in the example above.

In addition, meanings G and H did not occur in any of the 200 random tokens looked at in the various corpora used and were subsequently excluded from the results and analysis. When looking at the usage of thing, it was decided that the plural of the word, things, was not to be included in the analysis since its usage is quite different from that of the singular form. As a result, all the meanings of thing taken from the dictionary which refer to the plural, things, was therefore excluded from the coding categories presented above in Table 3. While coding the instances of thing it was at times hard to differentiate between the frequent meanings of 1 and 6
(see Table 3). It was chosen to go with meaning 6 when *thing* was preceded by an adjective expressing a certain feeling about something, as illustrated in (5) below by one of the Swedish learners:

(5) Aims in life are different: for some people the most essential *thing* is a happy family with at least two children; for some, it is a political or business career.  

(SWICLE)

When *thing* was preceded by a quantifier of some kind such as *one, only or first* it was put into category 1, as exemplified in (6).

(6) There is especially one *thing* worth mentioning, namely our addiction to television.  

(PICLE)

Finally, no instances of meaning 5 of *thing* could be identified in any of the corpora and it was therefore excluded from the analysis. As for the non-native speaker data this was not very surprising since the learners had plenty of time to search for an exact term instead of *thing* or to re-write the whole sentence when writing their essays. Furthermore, this meaning of *thing* is probably mostly used in spoken production by native speakers, when the speaker does not prioritise finding an alternative to *thing* but is more concerned with maintaining the flow of the conversation.

5. Results

In this section the results of the non-native speaker corpora and native speaker corpora comparison are presented. Table 4 and Figures 2 and 3 show the frequencies of the general nouns *people* and *thing* in the five different corpora used for the analysis.

Table 4 Frequency of *people* and *thing* in the five different corpora

<table>
<thead>
<tr>
<th></th>
<th>SWICLE</th>
<th>PICLE</th>
<th>Newspapers</th>
<th>BNC University Essays</th>
<th>BNC School Essays</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>people</em></td>
<td>704</td>
<td>802</td>
<td>145</td>
<td>174</td>
<td>209</td>
</tr>
<tr>
<td><em>thing</em></td>
<td>94</td>
<td>50</td>
<td>33</td>
<td>4</td>
<td>35</td>
</tr>
</tbody>
</table>
Table 4 and Figures 2 and 3 show that the noun *people* was used almost four times more frequently by the learners compared to the native speaker school pupils whereas the difference in the usage of *thing* was not as considerable, specifically for the Polish learners. Furthermore, an even greater difference in the usage of *people* in comparison to the university students could be identified, and when compared to the newspaper corpus, the word appeared nearly six times more in PICLE. The frequency of *thing* in the newspaper corpus was close to that of the native speaker school pupils, however, what stands out in the frequency count is the considerably lower figure identified in the native speaker university students’ essays. The Polish university students used *thing* more than ten times as frequently when compared to their native speaker counterparts. The statistical significance of the findings presented in Table 4 was tested using the Chi-squared test. The result showed that the differences between the learner corpora and the native speaker corpora were statistically significant ($p \leq 0.05$).
Table 5 and 6 show the frequencies of the different meanings (see section 4.2 for an explanation of these) of the word *people* and *thing* in the five different corpora.

Table 5 Frequency of the different meanings of *people* in the five different corpora

<table>
<thead>
<tr>
<th>Meaning</th>
<th>SWICLE</th>
<th>PICLE</th>
<th>Newspapers</th>
<th>BNC University Essays</th>
<th>BNC School Essays</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75</td>
<td>80</td>
<td>55</td>
<td>56</td>
<td>50</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>8</td>
<td>25</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>C</td>
<td>25</td>
<td>20</td>
<td>30</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>D</td>
<td>90</td>
<td>92</td>
<td>55</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>

Figure 4 Frequency of the different meanings of *people* in the five different corpora
Table 6 Frequency of the different meanings of thing in the five different corpora

<table>
<thead>
<tr>
<th>Meaning</th>
<th>SWICLE</th>
<th>PICLE</th>
<th>Newspapers</th>
<th>BNC University Essays</th>
<th>BNC School Essays</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>115</td>
<td>90</td>
<td>40</td>
<td>29</td>
<td>52</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>15</td>
<td>11</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>52</td>
<td>30</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
<td>20</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 5 Frequency of the different meanings of people in the five different corpora

Table 5 and Figure 4 show that meaning D of people, in the sense of ‘the mass of ordinary people’, was the most frequent for all these corpora except for the newspaper corpus where meaning A, ‘humans considered as a group or in indefinite numbers’, attained the same percentage. Comparing the non-natives’ and the natives’ usage of people it is made apparent by looking at the tables that the learners use this noun considerably more in meanings A and D, and
although the natives do make use of these two meanings more than the others, their usage is more diversified. In fact, meanings E and F of *people* did not appear at all in the sample of the learner corpus which was investigated, however, in all three native speaker corpora it appeared a number of times.

Examining Table 6 and Figure 5 it is found that meaning 1 of *thing* was the most frequent and meaning 6 the second most frequent in all five corpora. Furthermore, the learners only used *thing* in 9 of the 16 possible meanings of the word whereas the native speaker data only lacked one of its meanings. The learners used meanings 1 and 6 almost twice as frequently compared to the native speakers, except for the school pupils who actually used meaning 6 somewhat more frequently than the learners. Overall, it is clear that there were very small differences between the Polish and the Swedish learners’ usage of these two nouns, and that a discrepancy between the natives and the non-natives could clearly be identified. Performing a chi-square test of the distribution of the different meanings in Table 5 and 6 shows that there was a significant difference (p=<0.05) between the learners’ and native speakers’ production of these two nouns.

To investigate how the usage of these two nouns differ between native and non-native speakers further, the collocates of *people* and *thing* were identified, and can be seen in Table 7.

**Table 7 The most common collocates of people and thing in the five different corpora**

<table>
<thead>
<tr>
<th></th>
<th>SWICLE</th>
<th>PICLE</th>
<th>Newspapers</th>
<th>BNC University Essays</th>
<th>BNC School Essays</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td><strong>1st Left</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>that 10</td>
<td>some 10</td>
<td>the 10</td>
<td>the 15</td>
<td>the 15</td>
</tr>
<tr>
<td></td>
<td>make(s) 9</td>
<td>that 9</td>
<td>many 7</td>
<td>of 6</td>
<td>of 10</td>
</tr>
<tr>
<td></td>
<td>of 8</td>
<td>many 8</td>
<td>elderly 5</td>
<td>elderly 4</td>
<td>retired 5</td>
</tr>
<tr>
<td></td>
<td>any 8</td>
<td>make(s) 8</td>
<td>most 5</td>
<td>many 4</td>
<td>ordinary 5</td>
</tr>
<tr>
<td></td>
<td>more 8</td>
<td>more 7</td>
<td>young 4</td>
<td>retired 4</td>
<td>local 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>People</td>
<td><strong>1st Right</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>in 9</td>
<td>who 16</td>
<td>have 7</td>
<td>to 6</td>
<td>who 6</td>
</tr>
<tr>
<td></td>
<td>have 8</td>
<td>are 9</td>
<td>of 7</td>
<td>are 6</td>
<td>in 6</td>
</tr>
<tr>
<td></td>
<td>to 8</td>
<td>in 8</td>
<td>who 6</td>
<td>of 5</td>
<td>of 5</td>
</tr>
<tr>
<td></td>
<td>are 7</td>
<td>use 7</td>
<td>are 4</td>
<td>who 2</td>
<td>living 4</td>
</tr>
<tr>
<td></td>
<td>who 7</td>
<td>living 7</td>
<td>in 4</td>
<td>have 2</td>
<td>to 3</td>
</tr>
<tr>
<td>Thing</td>
<td><strong>1st Left</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>another 16</td>
<td>only 12</td>
<td>the 12</td>
<td>one 6</td>
<td>first 4</td>
</tr>
<tr>
<td></td>
<td>important 14</td>
<td>important 10</td>
<td>one 6</td>
<td>the 4</td>
<td>important 3</td>
</tr>
<tr>
<td></td>
<td>same 10</td>
<td>one 10</td>
<td>good 5</td>
<td>only 4</td>
<td>same 3</td>
</tr>
<tr>
<td></td>
<td>one 8</td>
<td>another 8</td>
<td>another 4</td>
<td>first 4</td>
<td>only 3</td>
</tr>
<tr>
<td></td>
<td>good 7</td>
<td>same 8</td>
<td>important 4</td>
<td>important 4</td>
<td>the 3</td>
</tr>
<tr>
<td>Thing</td>
<td><strong>1st Right</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>to 12</td>
<td>is 18</td>
<td>that 12</td>
<td>that 10</td>
<td>we 4</td>
</tr>
<tr>
<td></td>
<td>that 11</td>
<td>that 10</td>
<td>is 6</td>
<td>which 8</td>
<td>which 3</td>
</tr>
<tr>
<td></td>
<td>is 9</td>
<td>they 8</td>
<td>for 6</td>
<td>he 8</td>
<td>that 3</td>
</tr>
<tr>
<td></td>
<td>we 8</td>
<td>to 8</td>
<td>of 5</td>
<td>to 8</td>
<td>to 3</td>
</tr>
<tr>
<td></td>
<td>as 8</td>
<td>as 8</td>
<td>as 4</td>
<td>for 6</td>
<td>like 2</td>
</tr>
</tbody>
</table>
Presented in Table 7 (above) are the most common collocates of *people* and *thing* in the five different corpora divided into ‘first right’ and ‘first left’. Sorting the concordance lines in this way makes it possible to look at the most frequent words which either precede or succeed the word that is being investigated.

By examining Table 7 it seems as if the learners mainly rely on a small number of words compared to the native speakers since in the learner data there were more tokens of the same words whereas the native data displayed a greater range of words associated with *people* and *thing*. Furthermore, *the* is the most frequent first left collocation in all three native corpora, while it does not even make into the five most frequent in neither of the learner corpora. This could possibly be explained by the fact that meaning E, as in “the People of India”, was not found to be very frequent in the learners’ essays.

Yet another aspect of the production of these Polish and Swedish learners of English that was found to occur regularly was that they often start a sentence with *people*. In SWICLE *people* was used sentence-initially 18 times and in PICLE the number was even higher, namely 20, compared to the newspapers corpus’ 4 times and the two BNC corpora’s 6 times.

6. Discussion

6.1 Frequencies

When comparing the L2 learners’ (Swedish and Polish) usage of the two nouns *people* and *thing* in written production with native speakers’ it is clear that non-native speakers use these particular words much more frequently, thus giving support to Ringbom’s (1998) statement and therefore also answering the first of the research questions. The data also showed that younger native speakers seem to produce *people* and *thing* more frequently than the somewhat older native speakers; it could therefore be argued that as native speakers become more proficient and aware of stylistic variation in their language production they start using these general nouns less frequently and start employing other words, for instance replacing *people* with *citizens*, *individuals*, or *inhabitants*, and *thing* with *aspect*, *object*, *issue*, *matter* or *device*, thus making their production more explicit and less vague.

When conducting a comparative search in the learner corpora and the native speaker corpora of *citizens* it was found that this particular word was used considerably less frequently by the non-natives compared to the natives. In (7) this particular native speaker first uses *people* and
then alternates the word by using *citizens* instead, where a less proficient learner probably would have used the same word again.

(7) A territorial army containing ordinary *people, citizens* enlisting in their own force, and of their own free will. (BNC – School Essays)

Making a similar comparative search for *aspect*, one finds that in the native corpora this word is almost twice as frequent when compared to the learner corpora. In (8), a learner would then typically have written “another amazing thing…” which is supported by the results of the collocations analysis in Table 8 where *another* was one of the most frequent words associated with *thing*.

(8) Another amazing aspect of Tamburlaine’s character are his skills as an orator. (BNC – University Essays)

Applied to learners of English as a second language, this would imply that as they also become more proficient they will show a similar development. In fact, the second extension of Ringbom’s study involves investigating whether the overuse of basic vocabulary decreases over time, and if so how much and how fast. For this to be proven empirically, learner corpora would have to be divided into different proficiency levels, as recommended by Granger and Tyson (1996), and thereafter analysed accordingly. One would ideally have recourse to large writing samples from the same or equivalent learners over the years of their remaining studies, and indeed such an evolution in corpus building is currently underway. In the meantime, corpora of learner writing at three levels can be used to experiment with methods and provide some indication of what might be found. Extrapolation from cross-sectional to longitudinal data is a characteristic of learner corpus methodology, as it was in earlier interlanguage studies. Cobb (2003) conducted such a study but concluded by recognizing the inherent disadvantages in this approach and suggested, as did Barlow (2005), future researchers to carry out longitudinal studies of the learners’ production of the nouns *people* and *thing* to better understand the paths of development.

Another possible explanation to this overuse is linked to Halliday and Hasan’s (1976) claim that general nouns, such as *people* and *thing*, play a significant part in verbal interaction, and are also an important source of cohesion in the spoken language. It is a well-known fact that learners of English as a second language at the earlier stages of the acquisition process often fail to make a distinction between spoken and written production. Even at apparently native level it is
often observed that learners’ writing remains vague or resembles native speaker speech written down more than it does native speaker writing (Cobb 2003). By using too many spoken features in their writing the learners’ production may be perceived as less academic and more vague compared to traditional writing (Hunston 2002). This type of writing which more resembles spoken production written down than traditional writing is something that is also true for people learning their native language. At an early stage children learning to write in their first language, write similarly to the way they speak that language, but as they get older and more experienced as writers their written production becomes more formal and one can clearly identify a differentiation between spoken and written language. This is made apparent in Table 4 by the fact that the younger native speakers produced more instances of both people and thing compared to the older native speakers which partly confirms the hypothesis that these learners do display a higher degree of personal involvement since their writing resembles native speech more than traditional writing.

To investigate this hypothesis further the spoken sub-corpus of the BNC was searched for frequencies of people and thing. As predicted, thing was found to be substantially more frequent in the native spoken corpora compared to the native written corpora, attaining a frequency of 113 per 100,000 words which is close to the frequency identified for SWICLE, but more than twice as frequent to PICLE. This then lends support to the hypothesis that the learner essays do demonstrate a writing style more resembling native spoken language written down than traditional native writing. Furthermore, people was found to be used almost twice as frequently in the native spoken corpus compared to the native written corpora. However, the frequency was still significantly (p=<0.05) lower in comparison with the learner corpora, reaching a frequency of 412 per 100,000 words. Although this is merely half of the tokens which were found in PICLE, it still suggests that the learners’ written production displays features associated with native spoken language. However, it should be pointed out that it can also simply be a matter of pure overuse of the noun by the learners (especially in the generic meaning which is discussed below) or a combination of the two. Nevertheless, these features of the learners’ writing make their written production seem rather vague when compared to their native speakers’ counterpart.

When comparing the two non-native speaker corpora it was found that Swedish learners use thing almost twice as often as the Polish learners. This allows for some interesting observations, since when discrepancies among learners with different language backgrounds can
be identified, the influence of, for instance, the learner’s L1 can be further investigated (Granger 2002). In the case of Swedish learners overusing *thing* it could possibly be explained by the fact that in Swedish the word *sak* often functions as equivalent to *thing* and is used more extensively than *thing* is in English. In fact, when comparing the frequency of *sak* in a Swedish newspaper corpus (SVD 2000) with *thing* in the newspaper corpus used in this study, it was found that *sak* was used 48.57 times per 100,000 words compared to 32.91 for *thing* in the newspaper corpus. Therefore, one can argue that part of the Swedish learners’ problem of overusing *thing* seems to be due to negative L1 transfer, when they translate *sak* directly into *thing* not taking into consideration that native speakers of English often use another word instead, such as *aspect*, which, as mentioned earlier was found to be generally underused by the Swedish learners compared to the native speakers.

Moreover, the Polish equivalent of *thing*, *rzecz*, was searched for in the 93 million word Polish native corpus PELCRA. The frequency was surprisingly found to be lower than the native newspaper corpus, reaching a frequency of 21.13 per 100,000 words. This could then help explain the fact that the Polish learners used *thing* nearly half as frequently compared to the Swedish learners since *rzecz* does not seem to be as common in Polish as *sak* is in Swedish. The Polish learners have therefore been affected by a positive L1 transfer as opposed to the Swedish learners negative L1 transfer (Mitchell and Myles 2004). Taking the learners’ mother tongue into account like this would aid teachers of second languages to provide more focused and appropriate teaching methods (Meunier 2002). However, for teachers who have students with a wide range of L1 backgrounds this can be experienced as rather time-consuming. Therefore they might have to focus on problems with L1 transfer that are known to be common for most people learning English as a second language, as well as other common second language learning issues, such as making the learners aware of the distinction between spoken and written English language.

As for the usage of *people*, the discrepancy between the two learner corpora was not as substantial; the Polish learners only used the word somewhat more frequently. However, some interesting observations of L1 transfer could still be made. *Folk*, the closest Swedish equivalent of *people*, occurs 29 times per 100,000 words according to the Swedish newspaper corpus which is considerably lower than what was found for *people* in the English newspaper corpus. This does not then explain the overuse of *people* by the Swedish learners but partly confirms that neither a positive nor negative L1 transfer could be identified. The Polish equivalent of *people*, *ludzie*,
attained a frequency of 39 tokens per 100,000 words in the Polish native corpora, which also to some extent refutes any positive or negative L1 transfer. When analysing the usage of people, it seems as if both the Swedish and Polish learners are simply overusing the word without any, as far as this study can identify, L1 transfer having a positive or negative affect on it.

This initial analysis of the frequencies of people and thing in the five different corpora highlighted other issues which required a more in-depth analysis. As previously pointed out by Hunston “identifying possible differences between corpora when it comes to frequency allows for issues to arise that can then be studied in more detail” (2002:62). Below the most frequent collocates and meanings of people and thing are investigated in further detail in order to examine in what way the learners’ usage differ from the native speakers’ and whether/how this makes their production vaguer compared to traditional writing, which is the second research question.

6.2 Collocates and Meanings

As shown in Table 5, the generic meaning of people, namely meaning D, as demonstrated by (7) is significantly more common in the learner corpus compared to the newspaper corpus.

(7) Nowadays people are addicted to television.               (SWICLE)

Looking at the two essay corpora we find the discrepancy not as substantial, suggesting that within the genre of essay writing, the word people is more frequent than in the journalism register. Therefore, in essay writing it appears as if native speakers also use people more in the generic meaning, but still considerably less frequently than non-native speakers since an unambiguous divergence still could be identified in this study (p=<0.05). Another set of findings imply that native speakers often use an adjective before people to indicate that these people belong to a certain group, something that does not occur as often in the learners’ production. By looking at the most frequent collocates of people in Table 8 (above) it is found that common adjectives used by the native speakers together with the noun people include elderly, retired and, as demonstrated in (8), young:

(8) Success would encourage young people to stay in the inner city instead of moving.             (SWICLE)

By pointing out a specific group of people using, for instance, an adjective, the native speakers’ written production does not appear as vague as the non-native speakers’ production which relies
more on the generic meaning of the word, as can be seen in Table 8, where most of the common collocates associated with people in the two learner corpora are grammatical, and not, content words. However, one grammatical word that is noticeably missing from the most common collocates in SWICLE and PICLE is the. For the Polish learners this can be explained by the fact that in Polish, the corresponding word for people, ludzie, is not preceded by the definite article. Consequently, the People of Poland, is in Polish polscy ludzie, directly translated back into English as something like Polish people. In Swedish, the definite article in the same phrase is put after folk, the equivalent of people, producing the phrase folket av Polen. To sum up, the Polish learners end up overusing people in the generic meaning when they translate the Polish phrase for the People of Poland into English and the Swedish learners overuse becomes a fact when they disregard that the definite article in English is produced by putting the in front of, for instance, people. This then is another case of negative L1 transfer.

When it comes to collocates one might wonder why linguists are so interested in collocates? One reason comes from the idea that collocations are the building blocks of language and are, in some sense, fundamental units of language in use and that frequency of occurrence has a direct impact on the organisation of grammar (Bybee and Hopper 2001). In addition, the frequency of context words surrounding a keyword is used to provide clues as to the nature of the search word. The accumulation of contextual patterns provides good empirical evidence on various aspects of the meaning of a word (Barlow 2004). Trawling through the collocates, one also discovers that the learners seem to use a low number of collocates with a relatively high frequency which is commonly found for learner corpora which often show a greater use of a smaller range of vocabulary items (Hunston 2002). Previous studies conducted by Granger and others of the various subparts of the ICLE corpora have been quantitative rather than qualitative in nature, but interesting qualitative generalisations have also been made (Hunston 2002:207). For instance, an analysis of the collocates of thing suggests that the learners use a lot of prefabricated chunks such as an important thing is… and another thing is… In Stubbs’ (2004) view these phrases are very frequent since they are regular ways of expressing common meanings, in this case that of introducing a new topic into the text. Furthermore, Stubbs argues that frequently occurring phrases such as these express relations which are essential for understanding connected discourse (2004). This can possibly be explained by the SLA theory of connectionism where it is believed that second language learners often store language as prefabs.
which they tend to overuse and rarely modify. In general languages, as stated by Ladefoged (1972), consist of a relatively high number of prefabs, which makes learning a language easier and the actual language use more efficient. However, according to the number of collocates associated with the two nouns investigated in this study, learners of English as a second language seem to use a more restricted range of prefabs more frequently than native speakers. By repeatedly employing the same prefabs the learners’ written production displays a lack of complexity (Hunston 2002). When the learners keep producing “an important thing” native speakers might instead replace the adjective important with vital, imperative or as in (9) essential and thus increase the complexity of their production.

(9) Pollution is yet another essential thing to keep in mind for future generations. (SWICLE)

To support this hypothesis, the usage of essential and imperative were compared between SWICLE and the native university students. By making this comparison we find that essential and imperative are used by the Swedish learners on average 8.2 and 0.98 per 100,000 words whereas the natives’ frequency of these words are 16.2 and 3.6 per 100,000 words. Consequently the native speakers use these synonyms of important twice to three times as frequently compared to the non-native speakers. This type of underuse by the learners makes the learners’ production less varied and complex and is often demonstrated by the number of prefabricated chunks such as ...one important thing. An additional phrase that seems to be very frequent in the learner data is another thing and another /adjective (i.e. important)/ thing. When comparing the number of those particular phrases in SWICLE and the BNC university essays we find 24 tokens in the learner corpus and only 6 in the native speaker corpus. This also then substantiates the hypothesis that the learners make use of a set of prefabricated phrases much more often than the native speakers.

Although the frequencies of the above collocations (even if we group the data) would be too low to guarantee accuracy, and some might be topic-prone, the fact that transpires from Table 8 is that formal argumentative native English writing features a wider range of collocations co-occurring with the nouns people and thing. Informing Polish and Swedish learners explicitly about this stylistic tendency and providing appropriate exemplification would be beneficial for their development as EFL writers.

Yet another aspect of these L2 learners’ production that occurs frequently is that they often start a sentence with people in the generic sense (see Table 9 above), which to some extent contributes
to the vagueness of their written language. The learner data consists of more sentences such as (10) and (11) in comparison with the native speaker data.

(10) *People* are stuck in between a society in change and the church, a conservative world.                   (PICLE)

(11) *People* want their leaders to be perfect, while they allow themselves to do both this and that.               (SWICLE)

In contrast, native speakers often seem to make use of some sort of determiner, such as *these*, *many* or *most*, when they use *people* at the beginning of a sentence. This could possibly be to avoid sounding like they are talking about everyone or *people* in general. By letting a determiner precede *people* they make this more apparent to the reader, as demonstrated in (12) and (13).

(12) Many *people* even in Britain think that this is cruel.  
(13) Most *people* were looking for a government.      (Newspapers)

Therefore teachers should encourage their students to be a bit more specific when using *people*, thereby making their production seem less vague and to not risk making bold statements about what *people* in general are believed to think or do. However, for the Swedish learners this can also be related to how it is common for Swedish EFL teachers to stress different ways of translating the generic Swedish word *man* into English since *people* is one of the words often suggested along with *you* and *they*.

Moreover, when it comes to the usage of *people* the overuse is more concentrated to the generic meaning and the restricted range of meanings being used by the learners. The learners produce many sentences such as *People think this...* and *People are...* etc whereas the native speakers’ production displays a greater range of meanings where those types of sentences are complemented by others, e.g. *The people of Russia....* and *Local people who....* The fact that neither *the* nor *of* make the five most frequent collocates before or after *people* when they do so in all three native speaker corpora clearly substantiates this discrepancy.

Looking at the correlation between meaning A (i.e. *...as many of these people have read...*) and meaning D (see above for an example) of *people* it is found that within the newspaper corpus these are used with the same frequency, something that is not the case for the learner corpus, or the two essay based native speaker corpora. This could perhaps be explained by the fact that in essay writing *people* is used more frequently in the generic sense, by both non-native speakers and native speakers, thus indicating that this has not so much to do with whether
you are a native speaker or a non-native speaker, but more to do with what genre the written production belongs to. It would be of interest to investigate whether *people* in the generic sense is used even more frequently in spoken production by native speakers, since this could then help prove the hypothesis “that learners use features which are more typical of spoken English than of academic written English” (Hunston 2002:207). Particularly since a preliminary comparison of *people* and *thing* between the learner corpora and the spoken sub-corpus of the BNC, as discussed earlier (see page 30), showed similarities in frequencies.

When it comes to the usage of the different meanings of *thing* it is made apparent by looking at Table 8 that for the Swedish and the Polish learners and the younger native speakers the usage is concentrated around meanings 1 and 6, although the native school pupils display a greater range of meanings than the non-native speakers. Conversely, the data from the newspaper corpus and the native university students show that their use of the different meanings of *thing* is more multifaceted.

The results from this study clearly show that Polish and Swedish learners of English as a second language use *people* and *thing* in the generic sense more frequently than native speakers. Nevertheless, since the difference in usage between the younger native speaker school pupils and the Polish and the Swedish non-native speakers was not significant we have to consider the frequencies of the different meanings further before confirming the hypothesis that the non-native speakers’ production appear more vague when compared to native speakers’ production as a result of using the word *people* and *thing* instead of other more unambiguous words.

The analysis of the learner corpus resulted in no instances of meaning E and F, whereas in the three native speaker corpora a number of instances were found. The following examples of meaning E, …*for the people of the United Kingdom to*…, and of meaning F, …*of Christ in His people*… taken from the newspaper corpus demonstrate the type of meanings that did not appear in the learner corpora. Even though meaning C (i.e., *the exodus of middle-class people from most*) was used with approximately the same frequency across the five corpora, meaning B (e.g., ….and the Russian people had many…) was used considerably less frequently by the learners compared to the native speakers. Taking the usage of all these various meanings into account, it would therefore seem that what makes the learners’ production appear vaguer than the native speakers’ is that the learners do not use *people* in as many varied meanings as native speakers. So in contrast to what was predicted, it is not essentially the fact that the learners fail to
use other words instead of *people* that makes their production seem vague, instead it is due to their failure to use the word in meanings other than the generic one. Although this study shows that native speakers also use *people* rather frequently in the generic sense, they also use it in more specific meanings (i.e. meaning E and F), as shown by the examples above, and thus making their production more like traditional academic writing. Even though the same can be said for *thing* it does seem as if the overuse of *thing* mainly depends on the proficiency of the writer, hence the low frequency of *thing* used by university native speakers. In more formal academic English, *thing* is used considerably less often since then the word is substituted by a more precise and less general one.

The fact that SWICLE and PICLE were not compared to native speaker corpora consisting of essays on the same topic is one of the limitations of this study since some of the findings, as pointed out above, can be restricted to specific topics or genres. The second limitation of this study is that learner corpora are not a complete reference of what the learners actually know. The topics for the essays of course have a considerable influence on what kind of data the learner corpora end up consisting of. The various meanings of the word *people* which the native speakers produced and which the learners did not, could have some relation to the nature of the essay topics, meaning that the learners might know how to use these other meanings, but they just did not get the chance to show it. Another shortcoming is the fact that the corpora used for this study are relatively small in size which means that the findings should not, therefore, be generalized from without caution. This is, of course, ultimately true for all studies based on data observation: corpus-based research involves observing samples of language production in order to predict how the language may behave elsewhere (Partington 2001). In this sense, studies such as this one have no findings; they merely offer predictions to be tested against further data. “Like all language learners, the corpus user develops predictions from limited experience, predictions which have to be continually tested and revised in the light of new experiences of the language in use” (Partington 2001:84). A further limitation is that the author does not have any knowledge of the Polish language and although a Polish speaker was consulted in an effort to overcome this, no extensive conclusions regarding possible L1 transfer from the Polish learners’ mother tongue could be drawn. Further investigations into the usage of the word *people* and *thing* by Polish learners of English as a second language would preferably include this since Horvarth (2001) claimed that confirming or denying L1 transfer into the target language is why learner corpora
started being analysed in the first place. In addition, the web-based search engine of the PICLE which was used for this study only generates 50 hits each time, which also limits the generalisations we can make from the results of this study. However, in an effort to overcome this limitation, four separate searches were added together.

7. Conclusions

The distinctions between native and learner English that have been tested in this study have often been put forward by linguists, educators, and teachers, but until now there has been no systematic way to test or refine these hypotheses. Learner corpus analysis makes this possible, with promising results, mainly because contrastive learner corpus analysis is most useful as a step between intuition and hypothesis. Learner corpus analysis can establish the plausibility, for example, that overuse of words and phrases is a source of perceived non-nativeness in advanced learner writing.

This study has shown that Polish and Swedish learners of English as a second language use the high generality items *people* and *thing* more frequently than native speakers. However, this study has also shown that in the genre of essay writing, *people* is more frequently used than in other registers, such as journalism. It has also been suggested that younger native speakers use *people* in the generic sense almost as frequently as non-native speakers but as their language proficiency increases it gets used less frequently in this sense. In addition, it has been shown that *thing* is used considerably less frequently in formal academic writing compared to both the learners and the younger native speakers. However, it was also observed that *thing* was used almost twice as frequently by the Swedish learners, which could possibly be explained by L1 transfer, where the Swedish word *sak* is excessively translated into *thing* when in fact English native speakers use other words when Swedes use *sak*. Finally, what makes the learners’ production seem vaguer compared to native speakers’ is that they mainly use *people* in a generic sense, whereas native speakers’ usage is more varied in meaning, and that *thing* is used in sentences where the native speakers use a less generic and more specific word. In this essay it is proposed that one possible explanation for this could be related to the well known fact that second language learners rely more on spoken production in their writing in comparison to native speakers, similar to that of young native speakers’ written production. Nevertheless “advanced learners need to deepen their knowledge of words and phrases: to understand not just the main
meanings of a word, but to understand a range of meanings and to know how contexts in terms of discourse and genre provide clues to the appropriate meaning” (Barlow 2000:114).

In 1994 Granger published a seminal paper in which she claimed that “one should not exaggerate the impact of native corpora on foreign language teaching [and that] having access to comprehensive frequency lists may well help course designers compile better lexical syllabi, but it will not give them access to learners’ actual lexical problems” (Granger 1994:25). Learner corpora were intended to fill that gap. The release of a learner corpus such as the ICLE marks the beginning of a new stage in the evolution of learner corpus research. The time has come to use the resource on a wider scale in both SLA and ELT. On a more theoretical level, the ICLE data can be used alongside other data types of a more experimental nature to give SLA theories a more solid empirical foundation, in particular as regards the important question of L1 transfer. On a practical level, the ICLE can help produce more learner-aware pedagogical material designed for advanced EFL learners in general or focused on the needs of one national learner population.

This study has suggested that the use of *people* and *thing* declines as both the native and non-native speakers become more proficient, but to answer this question fully, future studies will have to include longitudinal data of both learner and native data. Nevertheless, this study provides some indication of what might be found in such a study. Furthermore it has identified some of the issues that teachers of English as a second language should be aware of when helping their students to avoid using the general nouns *people* and *thing* in a non-native like manner.

8. Implications for Teaching

The link between learner corpora and pedagogy has always been presumed to be a clear-cut one. It is based on the idea that corpus evidence suggests “which language items and processes are most likely to be encountered by language users, and which therefore may deserve more investment of time in instruction” (Kennedy 1998:281). Through contrastive analysis, learner corpora studies usually share the following aims. Firstly they investigate, using empirical data, how particular word or structure in a foreign language is incorrectly used by learners of one cultural group compared to native users of that language. Secondly they correlate the findings of the data with explanations from the learners’ L1. Finally they recommend pedagogical measures to help teachers and learners become more conscious of these errors in order to reduce them. Pedagogical recommendations commonly suggested are corpus-driven activities based on data.
from the target language (Tan 2005). The results of this study suggest, which is in line with Hunston’s (2002) recommendation, that there is no point in teachers simply telling their students that they should use *people* and *thing* less often. Instead teachers should point out and try to teach their students the various contexts in which the words can be used. It is also suggested that since the learners’ production seems to be more similar to that expected in speech than that of academic written English (Granger and Rayson 1998), teachers should make their students aware of this fact, so that the students could possibly avoid writing as they speak. In the case of the word *people*, that means not using it too much in the generic sense, and learning how to use in other more specific meanings. When it comes to overuse of *thing*, however, it is more a matter of teaching them how to use other more specific words or variation of the word such as *aspect*, *feature* or *part*.

By using learner corpora for students with different language backgrounds teachers could identify what aspect of their mother tongue results in negative transfer into the target language. This study has suggested that Swedish learners may be overusing *thing* because its equivalent in Swedish, *sak*, is used in a wider range of meanings compared to *thing* in English. When it comes to the Polish learners’ overuse of *people*, it has been proposed that they end up overusing *people* in the generic meaning due to a negative L1 transfer since the corresponding word for *people* in Polish, *ludzie*, is never preceded by the definite article. Consequently, for teachers of English as a foreign language who possess knowledge of their students’ L1, learner corpora could help them identify possible negative L1 transfer and establish more focused and appropriate teaching methods (Meunier 2002). One method involves using parallel concordances which, as pointed out by Frankenberg-Garcia (2004) however, should only be used when L2 problems can be traced back to the learners L1.

Parallel corpora enable us to access so many comparable facts of linguistic performance that it is easy to lose sight of the language contrasts that really matter, and to overburden learners with contrasts that bear no relation, and can even be detrimental, to their learning processes. Detecting negative transfer and other forms of crosslinguistic influence can help inform teachers which parallel concordances are likely to be pedagogically relevant to their students.

(Frankenberg-Garcia 2004:217)

Using parallel concordances of *people* and *thing* could enable the Swedish and the Polish learners to see how the meanings they formulate in their L1 can be expressed in English in a more native-like manner.
Another good way of getting at the overuse of the word *people* and *thing* by the learners would be for teachers to collect essays over time written by their students to see how the usage of these words change as they get more proficient (Knoch 2004). By identifying potential overuse or signs that the students are not using the word in enough different meanings or where another less ambiguous word is required, the teachers can make their student aware of this and find ways of helping them overcome this problem.

Vocabulary instruction in advanced courses typically consists of piling more and more low frequency vocabulary into learners’ heads for passive use in reading comprehension (Cobb 2003). The findings presented in this study suggest that some attention should also be paid to two other areas. Firstly some effort might usefully be devoted to the diversification of general nouns such as *people* and *thing* into more nuanced and only slightly less frequent items such as *individuals* and *aspect* which allow for both greater differentiation and more native-like production. Secondly, vocabulary courses should include the teaching of vocabulary for productive use since learners typically show their vocabulary knowledge to be far greater than what comes through in their writing. In a study involving vocabulary testing of a group English as a second language learners Cobb et al. (2001) found substantial passive vocabulary knowledge at levels well beyond their presupposed proficiency.

Moreover, if we can assume that advanced learners are reading widely in a variety of academic and other text types, then it seems clear that reading alone is not enough to expose the main features that distinguish the various genres, such as the low writer-reader visibility that typifies an argumentative text. There is then a case for mixing these learners’ reading with focused awareness raising of the formal features of different genres, perhaps accompanied by an occasional reversion to the outmoded practice of writing from models (Cobb 2003).

Granger and Tribble (1998) provide examples of material using a learner corpus and a comparative native speaker corpus. In these examples likely sources of error are first identified by comparing the two corpora in statistical terms, and then comparable concordance lines are used to encourage the learners to be aware of the differences between native and non-native usage. This could then be directly linked to the theory of ‘noticing’, or ‘consciousness/awareness raising’, advocated by scholars such as Schmidt. In his ‘Noticing Hypothesis’ Schmidt (1994) points out the vital importance for noticing in language learning. Furthermore, he claims that this noticing may in turn stimulate the processes of language acquisition. However, some linguists
have fundamental objections to this type of comparison because they believe that interlanguage should be studied in its own right and not as somehow deficient as compared to the native norm (Granger 2002). One way of dealing with this, as mentioned by Hunston (2002), could be to compare, for example, Swedish schoolchildren’s usage with expert Swedish speakers of English rather than with British speakers of English. Researchers using language corpora have indirectly given the impression that learner language is by default inauthentic, unnatural and does not exhibit native-like language behaviour, as a result of labels like overuse, underuse, and misuse (Tan 2005). Tan (2005) instead argues for a more realistic definition of authentic language use: one which focuses on the genuine use of a language to reflect the internal culture of the people in their local environment. This link between language and culture is particular interesting when you look at the growing role of English as a lingua franca and should therefore be on the agenda for future research into SLA and EFL/ESL pedagogy.

Finally, it is important to remember that corpus-based material should be viewed as a complementary method for language teaching, and not as a substitute for all other methods (Granger 2003). However, further research in corpus linguistics will undoubtedly continue to influence how languages are taught in the classroom, and how language teaching materials and dictionaries are designed.
9. References


