19 There is Nothing Like Native Speech: A Comparison of Native and Very Advanced Non-Native Speech

Britt Erman & Margareta Lewis
Stockholm University

> "I've been here for 8 ½ years, my English should be more fluent than this. Yes … sometimes I really stumble on the words…on the words"

1. Introduction

The above quote shows that finding words can be hard even for someone who has lived and worked in the L2 community for a considerable time. Vocabulary is an area of L2 acquisition that has received increasing attention in the last couple of decades. The present study is part of the research program "High-level proficiency in L2 use"[1]. The program seeks to provide answers to questions pertaining to what characterizes the very advanced L2 user, and involves several language departments at a Swedish university. This study compares vocabulary of different frequencies in the oral production of two groups of speakers of English, one non-native Swedish group and one native English-speaking group as a control. The non-native Swedish group has lived and worked in the UK (London) for an average of 7.3 years. The main aim of the study is to establish the rate of high-frequency and low-frequency words in the spoken data of these two groups. The material is made up of a recorded semi-structured interview. In order to establish lexical variation the present study, in contrast to several earlier studies, includes results not only from frequencies of tokens but also frequencies of types and T/T

[1] Thanks for generous grants are due to The Bank of Sweden Tercentenary Foundation.

How to cite this book chapter:
Erman, B. and Lewis, M. 2015. There is Nothing Like Native Speech: A Comparison of Native and Very Advanced Non-Native Speech. In: Shaw, P., Erman, B., Melchers, G. and Sundkvist, P. (eds) *From Clerks to Corpora: essays on the English language yesterday and today*. Pp. 349–366. Stockholm: Stockholm University Press. DOI: http://dx.doi.org/10.16993/bab.s License: CC-BY.

ratios, (cf. Lindqvist 2010, Lindqvist et al. 2011; Bardel et al. 2012; Lindqvist et al. 2013; Forsberg Lundell and Lindqvist 2012). Including types in the study will give indications regarding variation, which is assumed to distinguish native from non-native speech. Lindqvist (2010) found in her study of L2 French that the advanced learners used more general words to refer to key objects in a video film clip compared to a native control group.

The interview is one of three tasks carried out with the same participants. The results from two earlier studies, one on vocabulary and one on multiword structures (formulaic language), both involving two other tasks, a role play (dialogic) and an online retelling task (monologic), showed that in the role play the results of the London Swedes (LS) were like the natives in both studies, whereas the retelling task revealed significant differences between the NS and LS groups (Erman & Lewis 2011; 2013; Erman et al. 2014). Some of the questions asked in the interview concerned the Swedish participants' knowledge of languages and in particular their knowledge of English. Questions relating to English included for example the age at which they started learning English at school (in Sweden), whether they found speaking English difficult when they arrived in England, and the extent to which they used English also at home when in the UK. It is worth noting that all the Swedish speakers used English at work, and most of them had English-speaking partners at the time of the recording. Reading through the transcribed interviews it became apparent that the interviewees had rather varied perceptions of their knowledge of English, as the extracts below show. However, the general impression from these extracts is that the interviewees believe that their English is quite good, some even to the extent that English has taken over at the expense of their mother tongue, Swedish.

- An easy ride when it comes to languages. Watched English TV a lot when little. Always speak English with my English partner.
- It's much more natural to use English when speaking about music. I just can't find the Swedish word…
- English was one of my worst subjects in Sweden. Wasn't good at English at first (was very shy) but then just started speaking to people.
- I was fluent when arriving in England.
- Sometimes I feel when I go back, I become so conscious about my Swedish. And obviously I can still speak Swedish...it's no problem, but ...
- … sometimes I could have difficulty of swinging back into ...into fluent Swedish. I mean, when it comes to the more advanced

Swedish, I think. Because, I think, my Swedish stopped developing when I was 22 and I came here. And...and here I don't ...I don't associate that much with Swedes.

- English is ...what I realize with English is [after living in France]... it got a lot more words than French. French is, I think, if you're good in French, you use grammar to show that you are educated.

In this last extract there is a hint that English is perceived as having a large vocabulary.

The aim of the present study is not to establish whether the London Swedes' own perceptions of their knowledge of English has a bearing on the results but to find out how the two groups differ in their use of vocabulary in this task, more specifically across two main frequency ranges to be explained below.

We start by accounting for earlier research on vocabulary with a focus on advanced L2 speakers' spoken production (2). After presentations of aims (2.1), and material and method (3), we discuss the notion of frequency in relation to L2 acquisition (4). A description of the 1–2000 frequency range (4.1) is followed by a display of the results from this range (4.2), a description of the frequency range beyond 2000 words as this is applied in the present study (4.3), and the results of this frequency range (4.4). Finally, since the 1–2000 frequency range also includes high-frequency words typical of spoken discourse, we introduce a selection of sequences involving words from the 1000 most frequent words functioning as pragmatic markers (4.5) and present results from their distribution across the participant groups (4.6). Apart from offering some general insights to be drawn from the results, section 5 discusses the main contribution of the study. Section 6 winds up by presenting some more voices from the London Swedes in light of the results.

2. Earlier research

Establishing methods that relate vocabulary knowledge to different proficiency levels in L2 production has in the last few years been a major concern (Daller et al. 2007; Milton 2007; Tidball and Treffers-Daller 2007; Lindqvist 2010; Lindqvist et al. 2011; Bardel et al. 2012; Lindqvist et al. 2013). One method used is the Lexical Frequency Profile developed by Laufer and Nation (1995). A basic assumption behind most studies of vocabulary in relation to frequency is that frequency of input will affect output, so that the more frequent a word is the more likely it is to appear in an L2 speaker's production (Cobb and Horst 2004; Vermeer 2004). There is also evidence to prove that frequency plays an important role in L2 acquisition, implying that high-frequency words are shared by more L2 users than low-frequency words (Tidball and Treffers-Daller 2007). The higher the percentage of words beyond the 2000 most frequent words is in an L2 user's production, the more advanced is this person's vocabulary (Laufer 1995).The proportion of low-frequency words is also commonly referred to as lexical richness in the literature. The results from studies of lexical richness have shown that the quantity of lemma tokens of different frequencies distinguishes not only native from non-native speakers but also L2 speakers at different proficiency levels (Bardel et al. 2012). Some advanced non-native speakers of L2 French have been shown to reach nativelike levels in their use of low-frequency lemma tokens. But if some of these were removed from the list containing many low-frequency words (i.e. the 'Off-list'; see section 3), such as thematic words occurring in teaching materials and words that are similar in L1 and L2 (and some others), no non-native speaker of either L2 French or L2 Italian reached nativelike levels (Bardel et al. 2012).

2.1 Aims and research questions

In the aforementioned studies of multiword structures (MWSs) and vocabulary with the same participants in two tasks (see Introduction) it was found that the London Swedes behaved like the natives in the role play, but differed significantly from the native speakers on both vocabulary and MWSs in the online retelling of a film clip that was unfamiliar to them. On the basis of these results it is hypothesized that the London Swedes, being immersed in an English-speaking community, will come close to the native speakers in the interview, since this task is connected to a situation that is believed to be familiar, notably answering questions about themselves. As mentioned, two main frequency ranges are examined: the first two thousand words (1–2000 frequency range, i.e. words of high frequency), and those outside the first two thousand words (the 2000+ frequency range, i.e. low-frequency words). Our main aim is to compare the LS group with the NS group with regard to T/T ratios, and quantity of types and tokens in these two frequency ranges.

Another aspect closely related to vocabulary is the use of pragmatic markers, which are assumed to vary with text type (Simon-Vandenbergen

2000). Based on this it is hypothesized that an interaction involving a description of self such as in an interview will generate a considerable number of pragmatic markers. Furthermore, pragmatic markers have been found to distinguish native from non-native speech (Altenberg 1997; Denke 2009; Fant & Hancock 2014). These facts lead to our second aim, which is to establish how the LS group compares with the NS group on a selection of frequent pragmatic markers.

3. Material and method

Table 1 provides some more information about the participants.

The method used involves sorting the transcribed texts into frequency ranges by using the Lexical Frequency Profile (LFP), which is accessible via LexTutor[2]. By feeding in the transcribed texts in this program we get not only different frequency lists (see below) but also the total number of words, which distributes as follows over the two groups (Table 2).

Lexical frequency profiles are available in LexTutor via the program Vocabprofile. In Vocabprofile all the words are registered alphabetically in terms of type and token frequency; this makes the data easily accessible and allows various kinds of analyses. The words have not been lemmatized, which means that type frequencies are indicated in terms of 'word forms'; for example, *museum, museums,* and *call, calls,*

Table 1. Participants.

Informants	Time with English	Average age
10 Native speakers	Life	32
10 London Swedes	9 years at school and an average of 7.3 years' residency in London	32

Table 2. Number of words over the native speakers (NS), and London Swedes (LS).

Tasks/Participants	NS	LS	Total
Interview	23061	25184	48245

[2] LexTutor is accessible at: www.lextutor.ca

called, calling are all registered as six separate types, while representing two lemmas. The LFP program maps the word forms onto their lemmatized forms (i.e. 'call' and 'museum' for the six word forms above) in four categories (or lists): the first most frequent 1000 words, the second most frequent 1000 words, and the Academic Word List (AWL; Coxhead 2000). The fourth category is a separate list, called the Off-list, comprising any word (or item see 4.2) outside the 2000 most frequent words and the words in the AWL list.

It should be mentioned that although some types, especially in the high-frequency 1–1000 list, are inflections of one and the same lemma as in the examples above the majority belong to different lemmas (see 4.1). The further we move away on a scale from high-frequency words towards low-frequency words the more likely it is that type equates lemma type, and is thus unique (see 4.1).

4. Analysis and results

As mentioned, the results are divided into two main groups, the 1–2000 words frequency range and words beyond 2000, the 2000+ frequency range. The words (i.e. tokens) in the first 2000 word span constitute the major part of the present material and cover between 88% and 90% of the texts (see Table 3 below). These figures are above the average for written text, which is 80% for the first 2000 words;[3] this discrepancy may be explained by the rather informal character of the text type studied here, and by the fact that the present material constitutes spoken production.

In the present study the 2000+ frequency range is made up of the words in the AWL list and a pruned version of the words in the Off-list (see 4.3). The words in AWL make up the smallest proportion of the words for both groups, covering between 1% and 2% of the texts. It is common in the literature for calculations only to include number of lemma tokens (Bardel and Gudmundson 2012; Lindqvist et al. 2013; Forsberg, Lundell and Lindqvist 2012), but, as mentioned, in the present study it was relevant also to include the number of types. For instance, on some measurements the LS group is nativelike on the number of tokens, whereas they are non-nativelike on the number of types, which is an indication that this group recycles their types more often, implying less diversity. Although her own study only includes

[3] See http://www.lextutor.ca/research/Cobb

lemma tokens, Lindqvist (2010: 415) emphasizes the importance of also including types in studies on vocabulary.

4.1 Description of the 1-2000 frequency range

The 1–2000 frequency range apparently holds the most frequent content words and among the first thousand words we find many grammatical words needed to ensure structure and coherence, such as determiners, pronouns, conjunctions, etc. Words of high frequency by necessity come out in different word forms (i.e. types in LexTutor), some of which are based on the same lemma. In order to provide more exact relations between LexTutor (LT) type and lemma type we lemmatized all the LT types to find out the proportions over the frequency ranges. In the 1–2000 frequency range it was found that the proportions of different lemmas to LT types in the two groups are: NS 79.4% and LS 79.5%, and for the AWL lists: NS 92.6% and LS 93.0%. At the other end of the scale are the Off-list words where it was found that for NS 98.1% and LS 97.6% of the LT types belong to different lemmas. In other words, the vast majority of word forms (LT types) in the interview belong to different lemmas with average percentages for NS 86.7% and LS 85.8%.

Lexical frequency profiles with their focus on words are obviously independent of syntax and text type. It is not within the scope of the present study to evaluate the vocabulary produced, i.e. either to establish whether the words are syntactically, semantically or pragmatically appropriate, or their functions.

We start by accounting for the T/T ratios, and types and tokens per hundred words pertaining to the 1–2000 frequency range (Table 3) followed by a corresponding account of the results from the 2000+ frequency range (Table 4). Finally, we present and discuss results from searches targeting specific sequences (*you know, I think, sort of*) – which are among the 50 most frequent collocations according to Shin and Nation (2008) – and their distribution over the two groups.

The NS group functions as benchmark, and the threshold for significance is set at $p < .05$.[4]

[4] The chi-square test has been used throughout the study. We wish to thank Nils-Lennart for drawing our attention to this website: http://www.quantpsy.org/chisq/chisq.htm.

Table 3. T/T ratios, types and tokens/100 words in 1–2000 range in the Interview.

Interview	Type/Token	T/T ratios	p	Type/100 wds	p	Token/100 wds	p
NS	1416/20470	0.07		6.1		88.76	
LS	1366/22712	0.06	.000	5.4	.001	90.20	0.23

4.2 Results for the 1–2000 frequency range

Words belonging in the 1–2000 frequency range cover a large part of the texts as can be seen in the number of tokens per 100 words (Table 3). We also observe that the results in Table 3 are all based on LT results, since lemma types and LT types per 100 words yielded the same result, both showing that the difference between the NS and LS groups is highly significant (for lemma types per 100 words *p* <. 000). For this reason, *types* refers to LT types throughout the study.

Our hypothesis that the LS group would be nativelike on measurements pertaining to this task given its everyday character is only partly supported. While the LS group is nativelike on tokens per hundred words, they use significantly fewer types compared to the NS group. This result gives support for the inclusion of types in vocabulary studies. The highly significant difference in T/T ratio in the LS group compared to the NS group indicates that they recycle more words in this frequency range.

4.3 Description of the 2000+ range

In the present study the 2000+ frequency range is composed of a pruned version of the words in the Off-list combined with the words in the AWL list. The LexTutor Off-list is a heterogeneous group of items, low-frequency words as well as very informal high-frequency words and voiced pausing. In order to avoid a situation where words, because they are outside the frequency bands of the first 2000 words, would unduly be considered advanced or low-frequency, the Off-list was scrutinized and certain items were removed (cf. Lindqvist 2010; Lindqvist et al. 2013). As a consequence, all the items in the LexTutor Off-list that were deemed as not being part of a language's vocabulary, such as voiced pausing and word fragments, were removed. Indeed, equating the Off-list words with lexical richness can be misleading (Lindqvist 2010: 415).

The following types of items in the Off-lists have also been removed: **names** (of people, regions, places, continents, countries (including languages and nationalities, many of which are similar in Swedish and English, therefore more readily accessible; cf. Horst and Collins 2006; Milton 2007; Lindqvist et al. 2013)), **feedback** words (*yea, yeah, ok, huh, mm*), **foreign words** (*cher*), **contractions** (*wanna, gonna, gotta, coz*), **swear words** (*fucking*), **slang words** (*kids, guys, crap, ass*), and **voiced pausing** (*eh*, *uh/uhm/um(m))*, and, finally, **fragments** of words (*Thur*, *archi,* etc.).

Table 4 below shows the results for T/T ratios and types and tokens /100 words in the 2000+ word range.

4.4 Results for the 2000+ range

While the LS group is nativelike on T/T ratios, they significantly differ from the NS group on types and tokens per 100 words in the 2000+ frequency range (Table 4).

Our hypothesis that the LS group would be nativelike also in the 2000+ frequency range in view of the everyday character of this task was not confirmed by the results. The number of tokens per 100 words is significantly lower compared to the NS group, and the difference between the groups in the number of types per 100 words is highly significant, the *p*-value being close to zero. One possible explanation for this result is that the NSs use more specific vocabulary compared to the NNSs, which is in line with the results from several earlier studies (Ovtcharov et al. 2006; Lindqvist 2010; Erman & Lewis 2011).

It is worth noting that a comparison of T/T ratios between the three tasks targeting the LS and NS groups, i.e. the interview in the present study and the role play and the retelling task in Erman & Lewis (2013), shows that the interview is the task that demonstrates the highest T/T

Table 4. T/T ratios, types and tokens/100 words in 2000+ range (incl. AWL) in the Interview.

Interview	Type/Token	T/T ratios	p	Type/100 wds	p	Token/100 wds	p
NS	627/1077	0.58		2.7		4.7	
LS	537/1041	0.52	0.09	2.1	.000	4.1	.005

ratio in this frequency range. This is apparently the task where these speakers display the most diversity.

Summing up, while the non-natives reached nativelike levels in number of tokens in the 1–2000 frequency range, it is in the number of types in both frequency ranges that differences between natives and non-natives become visible. In light of the fact that the 1–2000 frequency range covers between 80% and 90% of all spoken texts, and to judge by the results of the present study, variation in this frequency range obviously is a nativelike feature, which distinguishes native and advanced non-native speakers. It is proposed in the present study that reaching a nativelike level in types in the first 2000 frequency range should be included in what is considered advanced vocabulary. In other words, showing variation among the 2000 most common words should be a skill worth aiming for also for advanced non-native speakers.

On the basis of the results presented in this study it seems reasonable to suggest that a contributing factor to divergences shown between the LS and NS groups is the difference in exposure, which has an effect also on types of high-frequency words as well as in the range of productive vocabulary at large.

4.5 Combinations of high-frequency words

LexTutor provides not only statistics, and alphabetical lists of words item per item in the frequency lists, but also the entire texts with each word marked for frequency and identifiable in the text. Depending on the query one can do either a search in the texts proper or in the word lists. If we are interested in specific *combinations* of words we apply the search command to the entire texts.

Since our results show that there are significant differences between the NS and LS groups in both frequency ranges, a sub-study involving particular, frequent combinations of high-frequency words, the majority functioning as pragmatic markers, was carried out.

The use of pragmatic markers has been shown to distinguish NN and N speakers of English (Denke 2009), and very advanced NN and N speakers of French and Spanish (Hancock & Kirchmeyer 2009; Hancock 2012; Fant & Hancock 2014). English pragmatic markers which have been shown to be used differently by NN and N speakers include *you know* (Denke 2009) and *sort of* (De Cock 2004). Denke (2009) found that not only is the pragmatic marker *you know* significantly more frequent in NS than in NNS speech, but the marker is also used differently, the NS speakers using the marker to organize discourse, i.e. as a discourse marker, and the NNS group as an editing marker in connection with stalling and repair. De Cock (2004) found that pragmatic markers of vagueness (*sort of, kind of*) are underrepresented in NNS compared to NS speech. Another English pragmatic marker, which, along with *you know* and *sort of*, belongs to the 50 most common 'collocations' in the 10 million word spoken part of the British National Corpus is *I think* (Shin and Nation 2008). This pragmatic marker has been shown to be overused by NN speakers in both speech (Altenberg 1997) and writing (Aijmer 2001). The results from the study of these collocations with a potential function as pragmatic markers will be shown below. It should be noted that this study is purely quantitative.

4.6 Results for pragmatic markers

The results show that in total figures the NS group has twice the number of pragmatic markers compared to the LS group (525 vs. 226). Numerically the LS group comes the closest to the NS group in their use of *you know*. Although the difference is statistically significant (*p* <.03), it is close to the threshold (*p* <.05).

The difference between the LS and NS groups for *sort of* is highly significant, the LS group using approximately one sixth (1/6) of the number used by the NS group. The significantly higher figure for *I think* in the LS group confirms results from earlier studies showing that there is a general tendency for non-natives to overuse this marker in both speech (Altenberg 1997; de Cock et al. 1998) and writing (Granger 1998; Ringbom 1998; Aijmer 2001). *I think* is a versatile marker and can signal a tentative attitude as well as authoritative deliberation

Table 5. Collocations (pragmatic markers) over the NS and LS groups in the interview.

Sequences Groups	you know	p	sort of	p	I think	p	Total
NS	162		235		128		525
/100 wds	0.7		1.0		0.55		2.27
LS	138		36		203		226
/100 wds	0.55	0.03	0.14	.000	0.8	.000	0.9

(Simon-Vandenbergen 2000; Aijmer 2001), but, as mentioned, the present study does not take qualitative aspects of these markers into account. It is worth noting that the *p*-values for *sort of* and *I think* are close to zero. This result strongly diverges from the LS results for *you know* which in comparison differ marginally from the NS group. One tentative explanation for the overuse of *I think* is that there are formally similar phrases in Swedish ('jag tycker', 'jag tror', 'jag tänker') with partly overlapping meanings and functions with the English phrase. The formal similarity and shared semantics between the English phrase and the three Swedish phrases may thus explain an overuse on the part of the Swedish L2 English users. This contrasts with the underuse of *sort of* which has no formal correspondence in Swedish. Swedish uses other downtoning items.

In sum, results from earlier studies of *sort of* being significantly underrepresented and *I think* significantly overrepresented in non-native compared to native speech have been confirmed in the present study. Nevertheless, it is worth noting that the significant overuse of *I think* in the LS group compared to the NS group does not compensate for a significant underuse by the LS group of all three pragmatic markers when collapsed compared to NS group (*p*-value < .000).

5. Conclusion and discussion

As is clear from our results, our hypothesis, that the LS group living and working in the L2 country would be nativelike on both frequency ranges studied in view of the fact that the participants are invited to talk about themselves, was in the main contradicted by the results. In only two out of six measurements (one for each frequency range) did the LS group score like the NS group. More specifically, they produced a nativelike number of tokens per 100 words in the high frequency range (1–2000), and were nativelike on the T/T ratio in the frequency range beyond 2000 (2000+). The most interesting result cutting across the two frequency ranges is that the LS group produced significantly fewer types compared to the NS group. However, the result for high-frequency tokens (the 1–2000 frequency range) for the LS group is in line with the general assumption that frequency plays an important role in L2 acquisition (Tidball and Treffers-Daller 2007).

The most important insight gained from the results of the present study is that when studying vocabulary it is important to analyze tokens as well as types, since they may yield divergent results. In other words, it is with regard to types that there is room for further development for

L2 users in the high-frequency as well as the low-frequency range. The results of this study suggest that displaying variation in the first 2000 frequency range is as much a native feature as showing variation in the beyond 2000 frequency range.

Furthermore, results from many earlier studies suggesting that the use of pragmatic markers is one area that distinguishes NSs and NNSs are supported in the present study, notably through significantly fewer occurrences of *you know* and *sort of*, and significantly more occurrences of *I think* in the LS group compared to the NS group. The quantity as well as proportion of pragmatic markers is thus what distinguishes the two groups. It is also worth noting that the significant overuse of *I think* in the LS group compared to the NS group does not compensate for a significant underuse by the LS group of all three pragmatic markers when collapsed compared to NS group (p-value < .000). One plausible explanation for this result is that although the pragmatic markers are known by the NN speakers, they have not become routinized. This would be in accordance with Bialystok (1993) who sees state of knowledge and control of knowledge as two separate processes. In other words, although the LS group obviously knows these pragmatic markers and may know when to use them, they might not have automatic control of them, which in turn can be explained by constraints related to real-time task performance.

The overall results suggest that native speakers have more immediate access not only to high-frequency and low-frequency words, but also to productive vocabulary more generally, including pragmatic markers. This can only be explained by differences in exposure and degree of more or less immediate access to items relevant for the situation.

6. Winding up

Against the backdrop of the results let us contemplate some more voices from the London Swedes regarding their beliefs about their knowledge of English. According to one of them, British English is difficult because of the rate at which it is spoken, and it is worth noting that this view persists after several years in the country.

> It's sometimes difficult to actively participate in the social environment. They...it's uh...British English is difficult, I think. It's spoken very, very fast... very quickly and you really have to ...to listen to understand. Uh...and sometimes you just don't understand what ...what they're talking about. I was taken aback by that.

We observe that understanding rapid speech can be an obstacle even at high levels of proficiency. From the introduction we recall the words of another participant concluding that English comes more naturally when speaking about certain topics (repeated here):

> It's much more natural to use English when speaking about music. I just can't find the Swedish word…

And below is one more extract along similar lines:

> And I would say that my English now has come to a point…and… uh… and at work, some topics at work, I feel more confident in English.

As a linguist it is easy to agree with these two speakers, linguistics being one of the many domains dominated by English.

> I think for certain things my English is better and for other things my Italian. For the job English is far better, but if I'm talking, I think, emotionally or generally, then I would be more comfortable in Italian than...than in English, I think.

Some acknowledge that English is difficult, but also that practice helps, as in this quote from the introduction, repeated here.

> English was one of my worst subjects in Sweden. Wasn't good at English at first (was very shy) but then just started speaking to people.

This view is shared by another speaker who, like the former speaker, eventually realized that participating in conversations is essential in everyday life.

> I always struggled with languages. That was never my strong subject in school. I'm a physicist. I'm a mathematician and... and I can't learn anything by heart. I need to have, you know, I need to understand why it is this way. But I realized that if I don't say anything, this is gonna be really, really boring and a bit useless so .. and I just kind of started speaking.

A couple of speakers comment on their Swedish accent when speaking English. In the second extract the speaker apparently considers her Swedish accent part of her identity.

> When I speak Swedish, it sounds like I'm singing. But when I speak English, I think my voice sounds really...uh ... monotone, do you see

> that? I heard some people say they thought I was Irish which was for me very, very strange.
>
> I have a Swedish accent but, yeah, and I don't think I'm trying to get away from that.

Whether or not you have an accent is of no importance according to another speaker.

> In England no one cares if you have got an accent.

Finally, two of the ten Swedes comment on English vocabulary and the limitations they often sense when speaking the language, which provides a clear link to the results of the present study. In fact, these quotes neatly summarize the overall results from the present study.

> Actually, I was surprised. I often find myself using English expressions but with Swedish words. And that's also funny because if you migrate away from what you're used to, you casually speak to someone about something else, you realize how poor your vocabulary actually is.
>
> I've been here for 8 ½ years, my English should be more fluent than this. Yes, sometimes I really stumble on the words...on the words. (We recall this quote from the beginning of the article.)

Of the three vocabulary tasks administered to these two groups (the interview in the present study, a role play and a retelling task (Erman & Lewis 2013), the interview was the task in which the London Swedes were the furthest away from the native group. Furthermore, this was the task in which the natives showed the most diversity in regard to low-frequency as well as high-frequency words. The fact that the native speakers distinguish themselves from the non-native speakers by having significantly higher numbers of types in both frequency ranges in this task may be explained by the interviewee being able to talk freely about anything that comes to mind in answering the questions asked by the interviewer. All in all, the results have shown that the native speakers had more immediate access to words across the board. Indeed, the most important insight gained through this study is that it is *frequent* word *types*, i.e. those within the 1–2000 range, that require practice in order to approach the quantity of native speakers. Infrequent words are presumably less important for general communication. Finally, the results of this study should encourage more research involving word types as well as tokens, and on larger corpora of different types of spoken production.

References

Aijmer, K. (ed.). (2001). *A Wealth of English: Studies in Honour of Göran Kjellmer.* Gothenburg Studies in English 81. Acta Universitatis Gothoburgensis, Gothenburg University.

———. (2001). *I think* as a marker of discourse style in argumentative student writing. K. Aijmer (ed.), 247–257.

Altenberg, B. (1997). Exploring the Swedish component of the International Corpus of Learner English. B. Lewandowska-Tomaszczyk & P. J. Melia (eds) *Proceedings of PALC'97: Practical Applications in Language Corpora.* Lódz: Lódz University Press, 119–132.

Bardel, C., Gudmundson, A. & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production: vocabulary acquisition and use in L2 French and Italian. N. Abrahamsson & K. Hyltenstam (eds) *High-level L2 Acquisition, Learning and Use*. Thematic issue of *Studies in Second Language Acquisition,* 34:2, 269–290.

Bialystok, E. (1993). Symbolic representation and attentional control in pragmatic competence. G. Kasper & S. Blum-Kulka (eds) *Interlanguage Pragmatics*. Oxford: Oxford University Press, 43–59.

Cobb, H., & Horst M. (2004). Is there room for an academic word list in French? P. Bogaards, & B. Laufer (eds) *Vocabulary in a Second Language: Selection, Acquisition, and Testing*. Amsterdam: Benjamins, 15–38.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly,* 34:2, 213–238.

Daller, H., van Hout, R. & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics,* 24:2, 197–222.

Daller, H., Milton, J., & Treffers-Daller, J. (eds) (2007). *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

De Cock, S., Granger, S., Leech, G. & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. S. Granger (ed.) *Learner English on Computer*. London & New York: Longman, 67–79.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English language and literatures (BELL) New Series,* 2, 225–246.

Denke, A. (2009). *Nativelike Performance: A Corpus Study of Pragmatic Markers, Repair and Repetition in Native and Non-native English Speech*. Saarbrücken: VDM Verlag.

Erman, B. & Lewis, M. (2011). Multiword structures in the speech of non-native and native speakers of English. Paper presented at EUROSLA 21, 21st annual conference of the European second language association, 8–10 September.

Erman, B. & Lewis, M. (2013). Vocabulary in advanced L2 English speech. N.-L. Johannesson, G. Melchers & B. Björkman (eds) *Of Butterflies and Birds, Dialects and Genres*. Stockholm Studies in English 104. Acta Universitatis Stockholmiensis, 93–108.

Erman, B., Denke, A., Fant, L. & Forsberg Lundell, F. (2014). Nativelike expression in the speech of long-residency L2 users: A study of multiword structures in L2 English, French and Spanish. *International Journal of Applied Linguistics* 24, doi: 10.1111/ijal.12061 2014.

Fant, L. & Hancock, V. (2014). Marqueurs discursifs connectifs chez des locuteurs de L2 très avancés: le cas de *alors* et *donc* en français et de *entonces* en espagnol. M. Borreguero Zuloaga & S. Gómez-Jordana Ferary (eds) *Marqueurs du discours dans les langues romanes: une approche contrastive*. Limoges: Lambert Lucas, 317–335.

Forsberg Lundell, F. & Lindqvist, C. (2012). Vocabulary development in advanced L2 French: do formulaic sequences and lexical richness develop at the same rate? *Language, Interaction, Acquisition (LIA)*, 3:1, 73–92.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. A.P. Cowie (ed.) *Phraseology, Theory, Analysis and Applications*, 145–160.

Hancock, V. & Kirchmeyer, N. (2009). Étude du marqueur polyfonctionnel *vraiment*. *L'information grammaticale*, 120, 14–23.

Hancock, V. (2012). Pragmatic use of temporal adverbs in L1 and L2 French: Functions and syntactic positions of textual markers in a spoken corpus. C. Lindqvist & C. Bardel (eds) *The Acquisition of French as a Second Language: New Developmental Perspectives*. Special issue of *Language, Interaction and Acquisition*, 3:1, 29–51.

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. L. Eubank, L. Selinker, & M. Sharwood Smith (eds) *The Current State of Interlanguage: Studies in Honor of William E. Rutherford*. Amsterdam: Benjamins, 265–272.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307–322.

Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avancé de français. *La revue Canadienne des Langues Vivantes/The Canadian Modern Language Review*, 66:3, 393–420.

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *International Review of Applied Linguistics (IRAL)*, 49, 221–240.

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. *Eurosla Monographs Series,* 2, 109–126.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, 24:4, 459–488.

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. H. Daller, J. Milton, & J. Treffers-Daller (eds) *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press, 133–149.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: a cross-linguistic approach. S. Granger (ed.) *Learner English on Computer.* London & New York: Longman.

Shin, D. & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, 62:4, 339–348.

Simon-Vandenbergen, A.-M. (2000). The functions of *I think* in political discourse. *International Journal of Applied Linguistics*, 10:1, 41–63.

Tidball, F., & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech. H. Daller, J. Milton, & J. Treffers-Daller (eds) *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press, 133–149.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. P. Bogaards, & B. Laufer (eds) *Vocabulary in a Second Language: Selection, Acquisition, and Testing*. Absterdam: Benjamins, 15–38.