# EVIDENCE OF LEXICAL PRIMING

# IN SPOKEN LIVERPOOL ENGLISH

Thesis submitted in accordance with
the requirements of the University of Liverpool
for the degree of Doctor of Philosophy
by

Michael Thomas Leonce Pace-Sigge

July 2010

**Abstract**

**Michael Pace-Sigge**

**Evidence of Lexical Priming in spoken Liverpool English**

This thesis is about two things. Firstly, drawing on Michael Hoey's *Lexical Priming,* it aims to extend the research represented in that book – into the roots of the concept of *priming* and into how far Hoey's claims are valid for spoken English corpora.

The thesis traces the development of the concept of priming, which was initially work done by computational analysts, psychologists and psycho-linguists, to present a clearer picture of what priming means and in how far the phenomenon of priming has been proven to be a salient model of how man's mind works. Moving on from that, I demonstrate how this model can be adapted to provide a model of language generation and use as Sinclair (2004) and Hoey (2003 *etc.)* have done, leading to the linguistic theory of *Lexical Priming*.

Secondly, throughout the thesis two speech communities are compared**:** a general community of English speakers throughout the UK and a specific community, namely the Liverpool English (Scouse) speakers of Liverpool, UK. In the course of this work, a socio-economic discussion highlights the notion of *Liverpool Exceptionalism* and, grounded in the theory of lexical priming, I aim to show through corpora-led research that this *Exceptionalism* manifests itself, linguistically, through (amongst other things) specific use of particular words and phrases. I thus research the lexical use of Liverpool speakers in direct comparison to the use by other UK English speakers. I explore the use of "*I*" and *people*, indefinite pronouns (*anybody, someone* etc.), discourse markers (like, *really, well, yeah* etc.*)* amongst other key items of spoken discourse where features of two varieties of English may systematically differ. The focus is on divergence found in their *collocation, colligation, semantic preference* and their *lexically driven grammatical patterns.*

Comparing casual spoken Liverpool English with the casual spoken (UK) English found in the Macmillan and BNC subcorpora, this study finds primings in the patterns of language use that appear in all three corpora. Beyond that, there are primings of language use that appear to be specific to the Liverpool English corpus.

With Scouse as the example under the microscope, this is an exploration into how speakers in different speech communities use the same language – but differently. It is not only the phonetic realisation, or the grammatical or lexical differences that define them as a separate speech group – it is the fact that they use the same lexicon in a distinct way. This means that lexical use, rather than just lexical stock, is a characterising feature of dialects.

# Acknowledgements

It is a long time since a sat in a postgraduate seminar at the University of Liverpool on a Wednesday afternoon in 2003 and saw and heard a paper that mentioned *Lexical Priming*. I was hooked there and then and decided to stay on as a postgraduate.

First of all I like to thanks those people who believed I could go all the way and actually present a PhD thesis when I still laughed this off - those who taught me in the 1st and 2nd year undergraduate, in particular Helena Kirby and Andrew Plowman, most of all though to the woman who probably believed I deserve having the best possible education and being able the highest reaches - my mother.

Very important, too, were all the fellow linguists who are a constant source of inspiration and keep me interested. First of all, Mike Scott, without whose *WordSmith* this would not have been possible. My thanks go to Linda Bawcom who helped editing chapter 3, Geoff Thompson who gave valuable material and advice for chapter 11 and, in particular, chapter 7. Likewise, I cannot thank Andrew Hamer enough for his interest and essential help on the socio-linguistic aspects beyond the call as a member of staff. Many thanks also go to Siobhan Chapman and Tony McEnery for their most valuable advice and direction for the final edit. All remaining mistakes are, of course, mine.

Most importantly for this thesis though is my original inspiration, the notion of *Lexical Priming*. Without Michael Hoey's theory I might not have been inspired enough do undertake this work. Crucially, it was Michael Hoey as an ever-patient, calmly directing and guiding supervisor who I like to thank most.

This work is dedicated to Zoë, my bestest.

# Table of Contents

# Figures and Tables

# Chapter 1 Introduction

## 1.1 Why this research

While there has been work on English accents for many centuries, the Liverpool English variant – Scouse – has only received attention since the 1970s. Previous surveys of English accents (Stanley Ellis 1974; Wells 1982:371 ff. and Trudgill 2000:71) are all agreed that Scouse is an *accent*. This is based on the fact that it differs mainly from Standard English in its realisation of sounds (particularly vowel sounds and the voiceless plosive consonants). Furthermore, Knowles (1978: 34) points out that Liverpool English is an accent but not a dialect on the grounds that "Liverpool English differs insufficiently in its grammar from Standard English". Likewise, it has only a small lexicon of words unique to the area.

A case can be made, however, for taking a different perspective on what counts as dialectal differences in order to explore whether Casual Spoken Liverpool English can be classified as a dialect. Dialectologists have traditionally concentrated on syntactic and morphological structures to describe a dialect. More recently, however, corpus linguistics has

suggested that **lexis** is a more complex phenomenon than traditional accounts of syntactical and morphological structures allow for, and some lexical features that have previously not been studied in a dialectal context may accordingly be relevant to a determination of difference.

Using corpora of Spoken English, I propose to research the complexity of the use of common **lexical items** and not rare or exclusive lexis by Liverpool speakers.

I hypothesise that, in casual spoken Liverpool English, it is not just the traditional criteria that identify a variety of language as a dialect. I am going to argue that a variety of English may also differ from other recognised varieties of English in respect of systematic variations in the use of **collocations, colligations** and **semantic preferences** (or **associations**[1]).

These terms can be briefly described this way:

- *Collocation* - the company a lexical item keeps. Collocation has been written about by Firth (1957); Halliday (1959); Sinclair (1991); Stubbs (1996); Partington (1998) and Hoey (2003a,b, c; 2005) amongst others.
- *Semantic preference* – the semantic field that a lexical item prefers. This term was coined by John Sinclair (1997). See also Hoey (2005).

[1] See chapter 3.4.1.4 for a detailed discussion of John Sinclair's term *semantic preference* and why Hoey (2005) uses *semantic association*.

- *Colligation* - the grammatical company a word keeps or avoids keeping and its preferred positioning and functions. See Firth (1957); Halliday (1959); Sinclair (1991); Hoey (2003a,b, c; 2005)
- *Lexically driven grammatical patterns* – extending the middle ground where grammar and lexis meet as revealed by corpus-driven research. – There were first discussed by Palmer & Hornby around 1933, written about by Hornby (1954) and taken up by Halliday & Hassan (1976), and later still by Hunston & Francis (2000).[2]

## 1.2 Potential value of this work

### *1.2.1 In respect of Dialectology*

If evidence of systematic differences in lexical use is indeed found between Scouse speakers and the speakers of other varieties, my research would extend the analytical tools of dialectology, in that I would have shown that dialects are as much distinguished by their collocational, colligational and semantic association uses as by their grammatical and lexicon differences. Indeed, if I were to find distinctive differences between Scouse and Casual Spoken Standard English, along the lines I have mentioned, it might even be necessary to re-define what counts as a dialect, in that it may not be only the grammatical or lexical differences

[2] These terms and the respective authors will be discussed in full detail in chapter 3.

that define a set of speakers as a separate speech group but the fact that they use the same lexicon in a distinct way.

My MA on Scouse lenition (Pace-Sigge: 2002) meant that I worked closely with a spoken corpus. I felt then that Liverpool English speakers seemed to use their lexicon in a way that was different both collocationally and colligationally from Spoken Standard English. In this thesis I intend therefore to re-visit the question of whether Liverpool English is an accent or a dialect.

The focus will be on whether lexical items that have so far not been described in the ways I have mentioned have preferences, which are distinctively different in Scouse from those in a corpus of Spoken Standard English.

Ultimately my goal is to discover which of the following descriptions best characterises Casual Spoken Liverpool English:

1. Casual Spoken Liverpool English is not a dialect at all, but clearly an accent – this would confirm the results of previous research. This would be shown if the collocations, semantic associations and colligations of Liverpool English prove not to be sufficiently numerous, or sufficiently frequent or sufficiently different from those of Standard English.
2. Casual Spoken Liverpool English is shown as markedly different in its use of lexical grammar (cf. Sinclair: 2000) in comparison to

Standard Spoken English. This would mean that, while Liverpool speakers use the same words, the same lexical stock, as other UK speakers, these are used in a different way. If this proved to be the case, we would have to classify Casual Spoken Liverpool English as either a separate dialect or part of another larger non-standard dialect.

Either way, if the 2nd outcome is achieved, it would have been shown that **lexical use**, rather than **just lexical stock,** is a characterising feature of dialects.

### *1.2.2 In respect to* Lexical Priming *in Spoken English*

Hoey's work (since 2002) has introduced the concept of *Lexical Priming* into the field of language studies. In this thesis I aim to map out the psychological development of the concept of *priming* and how Hoey came to find these principles salient for the use of competent language production. While Hoey has provided evidence of *lexical priming* based on corpora of written texts, the main focus of this thesis will rest, on trying to prove that *lexical priming* is a theory that is equally applicable to spoken (English) language. If *Lexical Priming* exists, I hypothesize that speakers in a geographically restricted area should be primed to reflect these primings in their speech. This means, they show patterns of

language use that show a preference amongst speakers in this area that is not shown by general UK English speakers.

## 1.3 The Casual Spoken Liverpool English Corpus: SCO and its comparators

For my initial research, I compare Casual Spoken Liverpool English (SCO) with Spoken English used by speakers throughout the UK. For this, I make use of a number of different corpora. (See 2.2 for a more detailed description). The most important of these is SCO, which was initially constructed for my MA and then, in a much expanded and fully transcribed form, for this research. SCO is based on recordings of over 50 informants. These people are Liverpudlians of a variety of age groups and of both sexes who live in the North, Centre & South of the city. The size of this corpus is 120 000 words.

Of the other corpora used in this research, the most important is the Casual Spoken English Corpus of Macmillan Dictionary Corpus (referred to as MAC), which contains 3.3 million words. For further reference, I also make use of data from the BNC Conversation Subcorpus (BNC/C), which contains 4.0 million words. In some cases, the Collins Bank of English (BoE) UKspoken with 9.2 million words will be a point of comparison as well.

# Chapter 2 Methodology

## 2.1 Building the Liverpool English Corpus (SCO)

This chapter describes the corpus linguistic research methods employed to discover whether and in what ways the Liverpool SCO corpus differs from other spoken English corpora.

### *2.1.1 General overview*

The Liverpool English Corpus (SCO) is based on casual spoken conversations collected by me between 2002 and 2005. It contains conversations held in a large variety of locations, by over 50 informants. These informants cover both sexes, and an age range from eight to 80. All informants live in areas across Liverpool. In the vast majority of cases, the informants are personally known to me (colleagues, friends, neighbours and relatives). All conversations are casual and informal – though the informants knew they were taped, there is little sign that this altered their speech[3].

[3] Though Meyer (2004) is sceptical in how far corpora can be used for research into language variation, the SCO has been specifically designed to answer some of his criticisms . Furthermore,

The Liverpool corpus, which I will refer to as SCO throughout, contains two speakers taped in 1992, 15 speakers taped in 2002, a different set of 15 taped in 2003 and a further 20 people taped in 2004 and 2005. All informants describe themselves either as skilled working class or lower middle-class. Given that all empirical evidence highlights that Liverpool is the poorest city in England[4], with the lowest percentage of working-age people in gainful employment and the employment below-average amongst the higher managerial[5], it is fitting that the majority of the interviewees come from working-class and lower middle-class backgrounds:

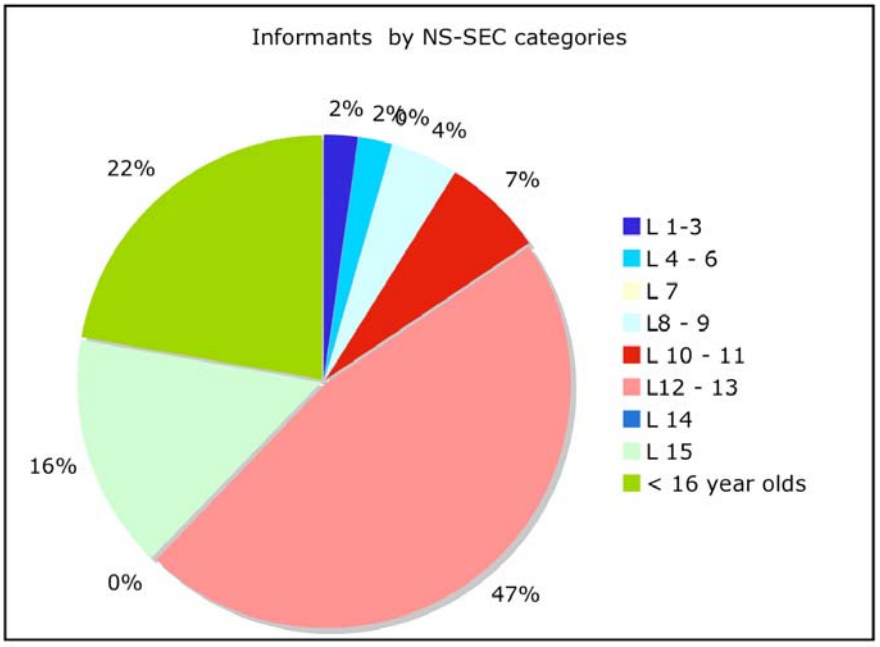



Figure 1: known economic background of 45 of the SCO informants.

There is an on-going debate as to what constitutes class. Both Sharon Ash (2002) and Ronald Macaulay (2005) give a comprehensive overview on the

some parts of this investigation are outside the framework Meyer considered.

4 *Indices of Deprivation*, a comparison of 354 local authorities in England in 2007 showed Liverpool as *most deprived*. Latest update to be found at (last accessed 10/09/2010): http://www.liverpool.gov.uk/Images/tcm21-64384.pdf

5 See chapter 4.1.1 for a full discussion of the socio-economic make-up of Liverpool.

various approaches used by sociolinguists since the 1960s. For this study, I use the NS-SEC criteria used by the UK Office of National Statistics (ONS).[6] The ONS classifies *Lower supervisory and technical occupations* as L10 & L11, *Semi-routine occupations* L12 and *Routine occupations* as L13. As Figure 1 shows, these are the occupations of the majority of the informants. Of the group of students[7] / under 16 year olds, their domestic background points to a similar class. See Appendix I.1 for a more detailed breakdown.

The SCO corpus contains a total of 119.079 words. Words that were inaudible (for example because the background noise inside a pub provided too much interference) have been marked as such. Longer periods of speech that are my own have not been transcribed – only the relevant utterance initiations and responses are kept, and these are not included in any calculations of frequency.

As in all corpora, variations in size, time of recording, choice of informants, etc, mean that one must be cautious in generalising from the SCO data. The SCO corpus does, however, highlight certain trends and features, which can be found in Liverpool speech.

[6] Full *National Statistics Socio-economic Classification User Manual* downloadable from http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=14066 (last accessed 10/06/09)

[7] Those listed as students, subsequently became the following: teacher; (small) shop-keeper; lower-level supervisor.

### 2.1.2 Method of SCO compilation

The SCO corpus records speech by informants who live in all parts of the city (South, Centre & North) and either come from Liverpool or have lived most of their lives in Liverpool. The total number of informants exceeds 50 – this means that no single person's idiosyncrasies are likely to greatly influence the resulting corpus[8].

Care was taken to include only informants who have lived most their lives in Liverpool or are firmly rooted in the city - see Figure 2:

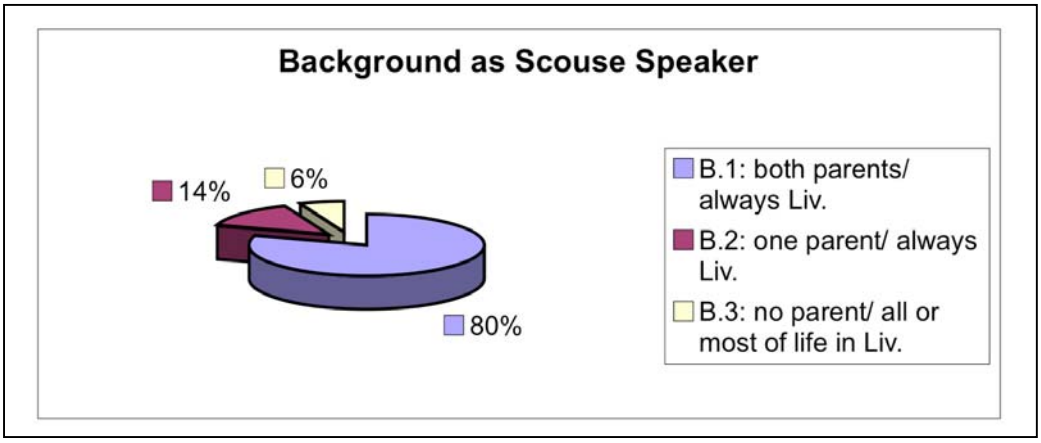



Figure 2: Background of the Informants, to determine whether they and / or their families have always lived in Liverpool.

Some Liverpudlians claim that they are able to tell which part of the city a speaker comes from, and which educational background they have[9]. However, Andrew Hamer [10], who has worked intensively on the characteristics of the Merseyside accent, has not been able to find any strong evidence for this (cf. Hamer: 1995 / 2009). Consequently,

[8] A complete breakdown of the informants can be found in Appendix I.
[9] Personal information and contributions by locals during an open lecture in Liverpool, 1995. Claims also include that listeners can determine whether the speakers are Catholic or Protestant; Everton or Liverpool supporters (!)
[10] lecturer, Liverpool University

informants hail from different parts of the Liverpool area. This excludes people from the Wirral where a different accent (despite the influx of former Liverpool dwellers) prevails. The Wirral historically has been part of Cheshire and, until the late 1990s, strongly tried to dissociate itself from Liverpool. On the other hand, the Liverpool area includes (New Town) Kirkby, which was specifically developed to house inhabitants from inner city Liverpool. People in Kirkby refer to "town" when they mean Liverpool. They see themselves still as part of it (though they are geographically removed).

My aim in creating the corpus was to record the speakers during casual conversation. Though ethical considerations determined that all participants knew they were being recorded, the recorded results appear to be sufficiently close to every-day conversation to justify transcription and analysis. In order to gain relatively unguarded, casual speech recordings, I never recorded total strangers. Instead, colleagues, family-members[11], friends and neighbours were recorded. Consequently, the speaker-listener relationship, and the normal development of the conversation as recorded, achieved a flow of speech that was not unduly influenced by self-consciousness.

A small, unobtrusive, handheld tape-recorder with in-built microphone was used so that the speaker was less inhibited by the fact of being taped. This felt to be more important than gaining the best possible clarity of recording.

[11] my in-laws and my daughter

As the focus of this study is on lexical clusters, there is no indication in the transcription of intonation or body language. Likewise, where there was overlapping speech, this was simply recorded as consecutive lines (apart from those cases where overlap made meaningful transcription impossible).

For comparison purposes two other corpora were used. I did not have the complete Macmillan Dictionary corpus. Instead, I used concordance lines for all the target words from the Macmillan Dictionary corpus (which will, from now on, be referred to as MAC), made available to me by Professor Michael Hoey. As a second comparator, I used the *Conversation* subcorpus of the British National Corpus (referred to throughout as BNC/C). These corpora are described in more detail in the following sections.

Likewise, I used concordance-lines of the target words when I checked the spoken subcorpus of the Bank of English (referred to throughout as BoE). At times the BoE acts as a corpus to cross-check findings – for example to confirm the validity of a marked divergence in results displayed when MAC and SCO corpora are compared. This facility was available to me through the Collins database accessible at the University of Liverpool library.

To create wordlists, to check for collocations and for keyword searches, full use of the WordSmith (Version 4, 2003 – Mike Scott) concordancer software has been made.[12] How the work was undertaken in detail will be described in section 2.5.

Comparisons were drawn by looking at sets of words and lexical items that occur frequently in casual speech in the two main corpora. For this thesis the focus was on individual words – and their collocations and colligations. Choosing a representative sample for unbiased comparison meant that the words selected had to match certain criteria.

- They had to be free-standing lexical items (words), not existing clusters.
- There had to be enough instances of the term for them to be relatively high-profile word in both corpora.
- They needed to be associated with both groups of speakers evenly.
- They needed to reflect functions that were performed by both speech communities.

I have the clear advantage of having recorded and transcribed the whole of SCO corpus myself. This meant that I gained valuable insights while transcribing and that I noticed peculiarities in the use of language. These,

[12] This software provides a function not unlike a zoom lens on a camera. With this, a single word, or even a cluster of words can be entered and searched so that the focus is on the concordance (zoom-in). Equally, rather than focussing, the software also enables the user to look at the wider picture and so at collocates for single words that can be found a number of words apart along the string. It also can highlight recurring clusters of words of a variety of lengths (zoom-out).

in turn, I was then able to check against occurrences in the UK spoken corpora, with the assistance of the computer.

To provide evidence of a clear local distinction in lexical patterning there would have to be found a marked divergence between the general UK (median) use and the specific Liverpool usage. This means that, while the same words (or items) – the same lexical stock – are available to Liverpool speakers as well as UK speakers, Liverpool speakers would have to be using them in a way that is different from the usage of most UK speakers. Given that the collective of English speakers utters millions of words over every hour and given the distortions that recordings and transcriptions can bring and adding to that the factors of time and corpora size, all the research undertaken here can only ever be seen as a snapshot of a larger whole.

## 2.2 MAC corpus as comparator

The Macmillan English Dictionary (in co-operation with Bloomsbury publishing) came out in 2002 and was therefore the most-up-to-date material to work with when I started on this thesis in September 2003. The Macmillan Dictionary is corpus-based and I have had access to concordance lines from their corpus as it stood at 2002. That corpus is based on casual speech subcorpora that are mainly UK English but also contain a (small) element of US spoken English material. The resulting corpus material consists of 3.3 million words and will be referred to as

*MAC*. The MAC corpus was, however, withdrawn for copyright reasons, as a result of a changed contractual relationship with Bloomsbury publishers with whom the original corpus had been created. It was therefore no longer available to me for the final chapter and some of the comparisons are based on the current (Jan 2009) Macmillan spoken corpus (MAC:MED) with a word total of 8,336,253 words.[13] While the exact details of the MAC corpus are no longer retrievable, MAC was made up out of the same elements, albeit proportionally smaller, than the MAC:MED.

## 2.3 Comparing SCO with other Spoken English corpora

In this section, I will briefly show the characteristics of the spoken English corpora employed as comparators in this study.

The *British National Corpus* (BNC) is a widely used English corpus which contains a spoken and a written English section. The *BNC Spoken Conversation* sub-folder (referred to as BNC/C hereafter) is a natural

[13] MAC:MED is made up out of the following elements:

| | |
|---|---|
| UK Political meetings – 464.093 words<br>UK University / School Teaching – 808.847 words<br>UK Business talks – 557.176 words<br>UK Club Meetings; Sports Talk – 506.015 words<br>UK Conversation Female (informal) – 1.810.769 words<br>UK Conversation Male (informal) – 1.679.469 words | US Broadcasting Speech – 73.035 Words<br>US Speech: University and Press (including White House Press Briefings) – 2.436.849 words<br><br>TOTAL: 8.336.253 words |

comparator for my purposes. This subcorpus contains 4,022,428 words. The material in this folder is available for research in text format files and these have been used for further investigation. Further details about the BNC can be found in *Aston* and *Burnard* (1998) as well as the BNC website:[14]

http://www.natcorp.ox.ac.uk/docs/userManual/design.xml.ID=spodes

The Collins Cobuild *Bank of English* (BoE) is available through subscription. The search functions available allow search by a single word (for example "*I*") and the BoE server then displays all the concordance lines found as well as a record of the number of all concordance lines. Further investigation (e.g. "*I*" plus *context word*) is also possible. To investigate the concordance in more detail (with other software for example), Collins allows the retrieval and download as text file of concordance lines.

The BoE offers a variety of subcorpora and their spoken English subcorpus is UKspoken, which contains 9,272,579 words. Any reference to BoE in this thesis refers to this subcorpus. The material in this corpus contains conversations recorded during job interviews, speeches, and exchanges in educational settings. If we classify "informal speech" as ad-hoc and not pre-planned and say that the setting does not permit large discrepancies of relative speaker power (as is the case in, for example, a job interview), this means that the BoE does not exclusively contain informal spoken exchanges.

[14] Last accessed 09/03/2009

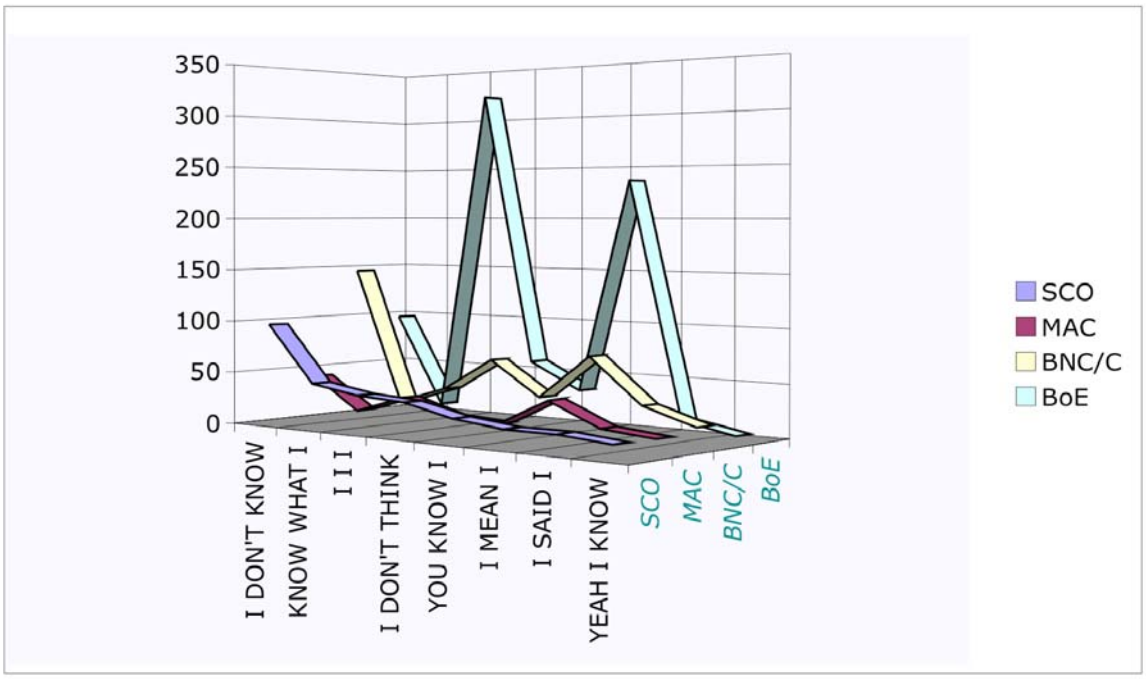



Figure 3: Highest Frequency "I" clusters compared in 4 Spoken corpora – SCO, MAC, BNC/C and BoE. Occurrence per 100.000 relative to total of words in corpus.

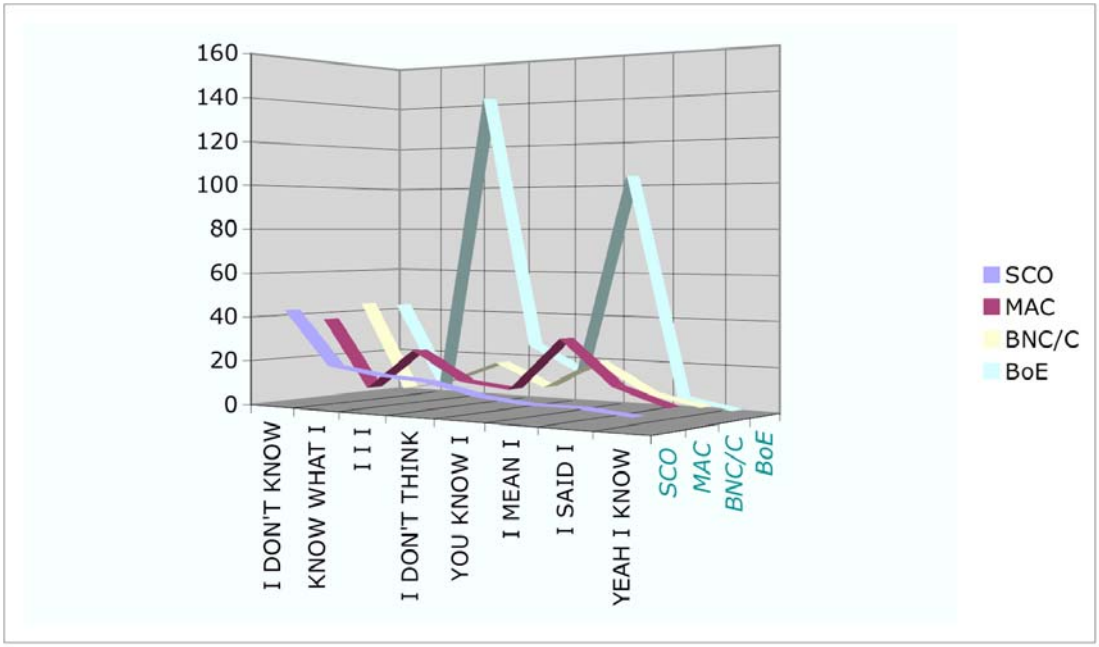



Figure 4: Highest Frequency "I" clusters compared in 4 Spoken corpora – SCO, MAC, BNC/C and BoE. Occurrence per 100.000 relative to total of "I" usage.

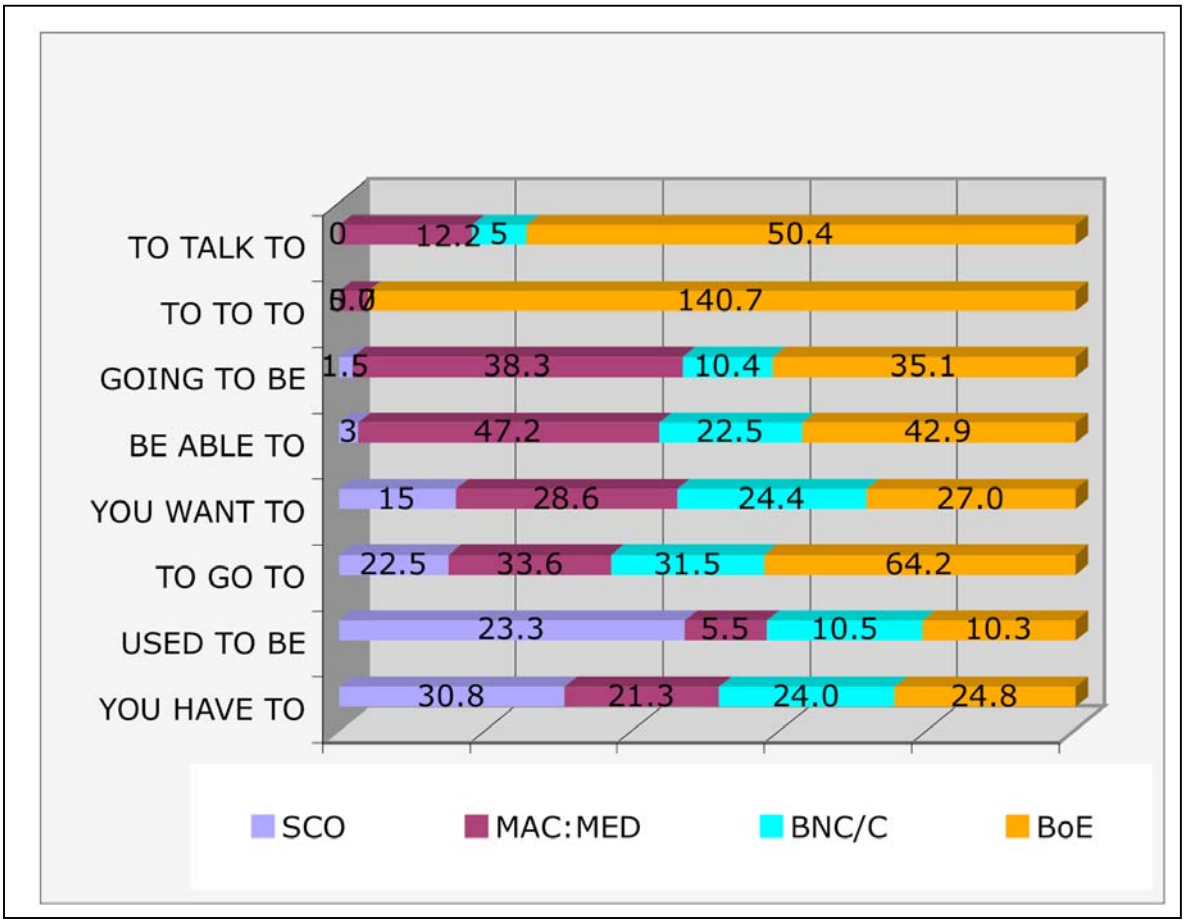



Figure 5: The 2 most frequent "TO" clusters and their occurrence in other corpora. Bottom to top: SCO; MAC:MED; BNC/C and BoE

## 2.4 Choosing the comparator corpus

Given the availability of three comparator corpora, each with its own strengths and potential weaknesses as comparator, it seems necessary to determine which would be the most appropriate for present purposes. Therefore, all four of the corpora introduced above were examined to see whether any one of them seemed to fall outside a general pattern.

As the main comparison is between SCO and other English spoken corpora, the main comparator needs to demonstrate a high level of typicality of English across the UK. This typicality can be, amongst other things, tested by the degree of congruency it has with similar corpora. All three comparators are general corpora and therefore differences amongst them are because of their construction and not because they aim to describe different varieties. It is at those points where all three corpora agree that we find the safest point of comparison.

Figures 3 and 4 show how the highest-occurring 3-word (3w) clusters in the four corpora discussed compare. "I" is the highest occurring word in all four spoken corpora and there is a high degree of overlap in the highest occurring "I" clusters in all four. In Figure 3 the comparison shows that the occurrence of the core clusters per 100.000 (100k) words is relative to the word-total of each corpus. Figure 4 makes the same comparison, only this time the comparison relative to the total occurrence of "I" in each corpus. Looking at Figures 3 – 5 we see, however, that BoE

quite often stands out in its results. Therefore, BoE can only be taken as a further source of comparison (i.e. to confirm salient features that appear in MAC and BNC/C but not SCO). It cannot, however, function as the main comparator to SCO.

While Figures 3 and 4 show the uses of "I" in MAC, Figure 5 highlights the research done with MAC:MED corpus. It reveals that MAC and MAC:MED present a middle way between all the corpora. Given that the MAC corpus was the most recent spoken corpus available as research on this thesis started, MAC comes out as the most trustworthy corpus to be used as a comparator.

That said, relevant results, where clear difference in use between SCO and MAC is found, will also be compared to occurrence patterns in the BNC/C throughout.

## 2.5 WordSmith concordancing

In the cases of BoE, MAC and MAC:MED there was no access to the full (sub)-corpora. Consequently, direct comparisons between lexical behaviour patterns that concern the whole of the corpus have only been made between SCO and the BNC/C. All corpora, however, allowed access to full concordance lines and direct comparisons were made with the assistance of Michael Scott's Wordsmith 4.0. This software produces full wordlists, concordances (including listing clusters, patterns, etc. and their respective frequencies), and comparisons of both keywords in

context (KWIC) function and, beyond that, key-phrases of any two different corpora. All results presented here have been calculated with the use of WordSmith 4.0[15].

With SCO and BNC/C, the full corpus was concordanced with the key terms researched as search words. With BoE, MAC and MAC:MED, concordance lines based on the key terms were used. The following steps were undertaken to be able to compare SCO material successfully:

Initially, a wordlist for the SCO corpus was created. This was used as an indication which words can be seen as sufficiently high frequency for any calculations. Next, unsuitable words were discarded. These included non-language elements (i.e. *inaudible*), corpus-specific names (*Liverpool, Al,* etc.). Amongst the remaining high-frequency terms, suitable points of comparison were selected and these became the core-terms then investigated.

A first step in the investigation was to compare relative proportional occurrence of each term in the respective corpora. Step two was then to create concordances of those keywords in the respective corpora. Once WordSmith has produced concordance lines, more detailed information is available. The first comparison between SCO core term data and the comparator data analysed concerned the top collocations – those words that can be found up to five words to the left or the right of the core word.

[15] Wordsmith 4.0 was the latest available version in 2003 and, so not to have dissimilar results by using different parameters, it has been used throughout. More details on the software can be found here: http://www.lexically.net/wordsmith/index.html (last accessed 14/03/2009)

The resulting comparison showed whether the proportional frequency of occurrence was broadly similar or showed deviation.

The next line of enquiry focused on clusters around the core term. These were mainly 2-word (2w) and 3-word (3w) clusters as longer clusters tend to occur at very low total frequencies in SCO. Longer clusters were discussed where they were recorded with sufficient frequency levels in SCO.

The final research chapter (Chapter 11) makes use of other WordSmith facilities like *Keyword Search* (see below) and the construction of frequency lists of the most frequent clusters in a corpus. SCO and BNC/C keywords were initially compared to gain a broad overview, which key terms might be worth discussing.

Scott, in WordSmith 4.0, describes keywords as such: *Key words are those whose frequency is unusually high in comparison with some norm.* (see Appendix (II.1)). Scott also indicates how keywords (and therefore, keyness) are calculated:

> The "key words" are calculated by comparing the frequency of each word in the wordlist of the text you're interested in with the frequency of the same word in the reference wordlist. All words that appear in the smaller list are considered, unless they are in a stop list. (Scott: 2003 Cf. Appendix (II.2))

The focus of the research was, however, to compare the keyness of clusters of words between BNC/C and SCO (and vice versa). Taking the clusters that were noticeably more key in SCO than in BNC/C, the core

words that appeared in a number of key clusters were selected to form the basis of the cluster - occurrence comparison.

## 2.6 Unchallengeable Claims

Macaulay: 2005, studying the use of *you know* in Scotland, and using a spoken corpus in many ways (not at least size) similar to SCO, comments as follows:

> Quantitative studies of discourse features are still at a very preliminary stage. No doubt, improved methods of creating machine-readable corpora of speech recorded under a variety of circumstances (Sinclair, 1995) will provide more accurate information on many aspects of discourse. In the meantime, small-scale projects such as [these], while they cannot provide evidence on which to make unchallengeable claims (sic), can perhaps provide pointers for future research. From the figures presented (…), some tentative conclusions can be drawn, though their significance may not extend beyond the boundaries of Scotland. (Macaulay 2005: 765)

What he says applies with equal strength to this investigation into Liverpool English, which also employs a specialist corpus (SCO) that is small by any comparison.

One further, important, point needs to be raised. In 2.1.1, I noted that the SCO corpus mainly consists of material recorded from working class and lower middle class informants. It might be argued that it is the dissimilarity of class background between SCO and the (assumed to be proportionally more middle-class speakers consisting) MAC and BNC/C

that accounts for the differences found. This may be true.[16] I must point out, however, that Liverpool is a very poor city by any standard. There are proportionally more low-income, routine workers than people with executive power in Liverpool and this means that the proportional frequency of lower class speech pattern is higher than in, say, the English of Edinburgh. In other words, the working and lower middle classes are the predominant social groups within this particular geographical area and shape the area's speech pattern.

[16] It will certainly be very interesting to compare SCO with other spoken corpora which consists of lower-class speakers only. Unfortunately, there are very few of these and they are hard to come by. Only this kind of comparison could help in deciding whether SCO reflects a class rather than a geographical variation of English.

# Chapter 3 The Theoretical Backbone

## 3.1. The concept of Priming in the context of language use

In order to find out whether Corpus Linguistic techniques provide the kinds of answers we are looking for as a first step I like to clarify how this approach works. Both dialectology and corpus linguistics focus on naturally occurring speech intensely, investigating patterns of language usage. This chapter will highlight how corpus linguistics techniques, and in particular the theory of *lexical priming*, are being used to investigate the evidence from the corpora available.

It will be shown how the theory was developed out of the material that corpus linguistic research brought up to provide a model of language generation and use. Moving on from that, I shall examine work done by computational analysts, psychologists and psycho-linguists to present a clearer picture of what *priming* means and how far the theory of *priming* has been accepted to be a proven model of how the mind works as regards language.

## 3.2 Lexical Priming

This chapter is solely concerned with the theoretical background, the backbone, on which the corpus-linguistic aspects of this thesis hinge. One of the main motivations behind the thesis lies in researching how far Michael Hoey's theory of lexical priming can be verified when looking at a spoken variant.

This chapter gives an overview of this theory, starting with the roots that appear to go back to the 1920s, via the series of new definitions of colligation and the impact of computation that led to the rise of corpus linguistics, to the publication of papers and the book *Lexical Priming*, which in turn led to new research initiatives – this thesis being one of them. The reception of the theory and its future development will be briefly shown here, too.

### *3.2.1 Where Lexical Priming came from*

Some ideas need incubation time and new people, new techniques, and new technologies to finally make the impact. The computational machine is a case in point. Babbage[17] could make one – but only the IBM / Apple / Microsoft–led electronic revolution of the 80s and 90s of the last millennium made IT impossible to live without. It is of little surprise,

[17] Charles *Babbage*, FRS (26 December 1791 – 18 October 1871) conceived a machine capable of computations but his plans could not be turned into a functioning machine at the time. In 1991 such a *Difference Machine* based on Babbage plans was successfully constructed by the British National Museum of Science and Industry.

then, that some ideas were developed out of nearly forgotten research. Like Babbage, who could conceive but not build a computer, early linguists could conceive the ideas that would find new importance in what we now call corpus linguistics. It is a bit like the invention of early aircraft. As long as man remained on the ground, only geographical fixed elevated points (like trees for short distances and mountains for longer distances) could give an impression of what things looked like from above. Today, an outline is available to anyone – not least because Google Earth[18] makes satellite pictures accessible. The same experience is true when millions, or indeed, billions of words from different sources can be collated and used for concordances which allows for a much finer grained vision of language.

Vastly expanded computer power has made corpus linguistics an influential force. Today it is hard to imagine that it was a tedious, complicated, and time-consuming process to even assemble a small corpus in the 1960s. Then, computers had to be fed by punch-cards and machines the size of large rooms had, compared to today, laughably weak computing powers. Even with very crude methods, and small memories, though, general tendencies could be highlighted. Something that early collections of words in context and intuitions about language use were unable to do.

Nevertheless, even as early as in the 1920s, Harold Palmer started what would become a cornerstone of British Applied Linguistics, Palmer

[18] http://earth.google.co.uk/

devised lists of the most frequently used words and phrases, constructed what he later termed Pattern Grammar (which was then refined by AS Hornby in 1954[19] and taken up by Hunston/Francis in 1996) and gave a detailed study of collocations to the Carnegie Conference in 1934.[20]
It seemed then that traditional grammar was to tumble:

> The traditional categories of grammatical description are survivals of a medieval scholastic instrument. They have been used to deal both with the forms and meanings of linguistic constituents in the vaguest of socio-philosophical terms, and judged by modern standards they have been found wanting in both enterprises. (…)
> Is there any more reason to perpetuate them than medieval alchemy?
>
> (Firth 1937: 154)

This was published over 70 years ago. Firth was a positivist, a believer of English as both a world-language and that "the English language is the greatest social force in the world" (Firth 1937: 156). "A language is not merely a community of sounds or even of grammar and dictionary. It is also a community of usage and idiom..." (Firth 1937: 155). Like much else said by Firth so many years ago, it seems to affirm work done much later; it appears to have sown a seed for John Sinclair's corpus work as well as a lot of empirical research into language use is based on corpora of naturally occurring language.

With such corpora, patterns became not just visible but also viable for fundamental research purposes. The concept of pattern grammar

[19] Hunston / Francis (1999) see Hornby's 1954 book *A Guide to Pattern and Usage of English* as their forerunner.
[20] Cf. Richard Smith: 1999 - see Appendix IV

consequently came prominently out of the work on the Collins Cobuild dictionary, which was the first corpus-based dictionary. It was the review of repeated patterns in preparation for that dictionary that led to the discovery that the lexis is not best described as made up of interchangeable blocks in a fixed structure that is called grammar.

John Sinclair describes it concisely in *The search for units of meaning* (1996):

> At present the only available measure of significance (*of a language pattern*) is to compare the frequency of a linguistic event against the likelihood that it has come about by chance. Since language is well known to be highly organized, and each new corpus study reveals new patterns of organization, a relationship to chance is not likely to be very revealing. A complete freedom of choice, then, of a single word is rare. So is complete determination. As in ethics, freedom and determinism are two conflicting principles of organization which between them produce a rich continuum.
>
> (Sinclair [1996] 2004: 29)

To filter language and exclude chance – regardless which corpus is being used – a practice needs to be established for comparing usage and finding patterns that highlight this organization of language. The accepted solution is to create concordances.

It is from the analysis of concordance lines that further areas of research stem. Nelson (2000) discusses the tripartite backbone of concordance work and places the importance of such work in seeing language as the means of communicating as follows:

> …collocation, semantic prosody and colligation are not totally separate concepts, but are, rather, interdependent and together create a network of meaning.
>
> (Nelson 2000:122)

Below, I will try and give a historical overview of the meanings attached to these three terms. The order I adopt is borrowed from John Sinclair's theory of their stages of removal from the actual word in abstraction.
In these sections it is shown in what way "Lexical Priming" can serve as an explanation for their existence.

#### 3.2.1.2 Collocation

Collocation is a noun whose use dates back to 1605 (Merriam-Webster) and indicates the following: "the act or result of placing or arranging together; specifically: a noticeable arrangement or conjoining of linguistic elements (as words)".
Michael Hoey[21] points out that the term collocation, widely attributed to Firth (1957), was already being used by the eighteenth century explorer of language change and language families, Sir William Jones. For all that, it was Firth that brought its use into the mainstream[22].

[21] personal communication.
[22] Xiao & McEnery (2006: 105) give a shorter overview and say: "Collocation has been studied for at least five decades. The word collocation was first used as a technical term by Firth (1957) when he said 'I propose to bring forward as a technical term, meaning by collocation, and apply the test of collocability' (Firth 1957: 194)."

> … the concept of collocation, that is, syntagmatic relations between words as such, not between categories. As Firth (1957) puts it: "you shall know a word by the company it keeps … The habitual collocations [of words] are simply the mere word accompaniment." (Stubbs 1996: 35)

Firth's diligent and hugely influential student, Halliday, uses the term collocation liberally in his 1959 work *The Language of the Chinese "Secret History of the Mongols".* This work would become seminal for Hoey.

Michael Hoey updated the definition to make it more specific:

> The statistical definition of collocation is that it is the relationship a lexical item has with items that appear with greater than random probability in its (textual) context. (Hoey 1991: 6f)

This is clearer and closer to the mathematical definition of the term, as it excludes co-occurrence – the instances where words happened to occur in close proximity of each other but at random and without the formulation of a pattern.

Sinclair, like Kjellmer (1984), Stubbs (1996) and Biber *et al.* (1998) as well as many other corpus linguists, describes collocation as a phenomenon observable in language and made visible in concordances. Sinclair and Stubbs keep on pointing out that more often than not, concordances make collocations visible that would not have been found by simply relying on intuition[23].

[23] Cf. Sinclair (1991: 112) *The commonest meanings of the commonest words are not the meanings supplied by introspection.* Or Stubbs (1995:381) *Often, a corpus will reveal a use of a word which is obvious once it has been seen, but which did not occur to one's intuition.* Also, Louw (1993) is an

The traditional dictionary definition given above is mirrored by the synonyms that Roget's Thesaurus suggests: arrangement; assemblage; location and phrase. The latter is of particular importance, as it hints at the fact that certain frozen collocations can form a phrase – or idiom. Sinclair narrows the definition of the term even further. He points out that that the "idiom principle" grows out of "frozen collocations":

> Tending towards idiomacity is the *phraseological tendency,* where words tend to go together and make meanings by their combinations. Here is collocation, and other features of idiomaticity. (Italics in original – MPS), (Sinclair [1996] 2004: 29)

Collocations are more than words appearing together in one context. Once a statistically high frequency of use is established, this can be seen as more than just a chunk of words but rather as a meaningful cluster that has "idiomaticity".

Hoey initially accepted collocation as a term to describe what Sinclair, he and the others found. It was part of the linguistic landscape of the day – and he was employing the term in that way still in 1997.

The next step for Hoey was to ask how collocation comes into being. This is where the pervasive use of collocation starts to become interesting. It is those linguists who are concerned with how the mind works – psycholinguists – who actually highlight why there are collocations and not mere co-occurrences of words. Wray (2002a) points out that

---

influential article which shows how concordance data on frequent collocation provide observable evidence of pragmatic meanings.

collocation is a fluid version of formulaicity and highlights that formulaic blocks appear as part of first language acquisition.

This brings a psychological dimension into the discussion. As discussed above, psychologists had constructed experiments over the past decades that prove that human minds connect some words more closely than others. Our brains appear to link knowledge of how words collocate with each other with the possibility of cohesion between any two lexical items. Halliday and Hasan (1976) speak of lexical items that are in one way or typically associated with each other.

While Hoey (2005) quotes Leech (1974) and Partington (1998) to give psychological reasons why speakers would collocate, it needs to be said that this is also highlighted by Halliday and Hasan, using wording oddly prescient of what Hoey would write in 2005:

> Without our being aware of it, each occurrence of a lexical item carries with it its own textual history, a particular collocational environment that has been building up in the course of the creation of the text and that will provide the context within which the item will be incarnated on this particular occasion. (Halliday & Hasan 1976: 289)

This is echoed by Hoey:

> The importance of collocation for a theory of the lexicon lies in the fact that at least some sentences … are made up of interlocking collocations such that they could be said to reproduce, albeit with important variations, stretches of earlier sentences. It could be argued that such sentences owe their existence to the collocations they manifest.
>
> (Hoey 2005: 5)

Michael Hoey turns Halliday & Hasan's argument on its head. It is not the creation of a text that makes us collocate. We carry, without being aware of it a template in our heads to collocate certain words, and these subconsciously recognisable collocates create the sense of cohesion for the reader:

> We can only account for collocation if we assume that every word is mentally **primed** for collocational use. As a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context. (author's highlighting - MPS), (Hoey 2005: 8)

This brings the issue of collocation full circle – from an observed phenomenon that is statistically more than random and therefore displaying a pattern that is formed through the exposure to a word in its specific context. It shows that what we call intuition clearly works on two levels. If we are asked to name collocations "intuitively" our mind seems to map language differently, as we come up with what we think are strong collocates, whether or not these may be statistically of a low frequency. Our subconscious intuition however produces collocates without thinking – and these are recorded as our typical language use. Collocation, therefore, follows a psychological, subconscious process.

### 3.2.1.3 Colligation

As the root of the term discloses, this is another Latin word. According to the OED it is

> **colligation** noun. ORIGIN mid 16th cent. (in the literal Latin sense): from Latin ***colligat- 'bound together,'*** from the verb ***colligare***, from ***col- 'together'*** + ***ligare 'bind.'***
> The current sense dates from the 1960s.

Interestingly, colligation seems to have been a psychological concept that was first described by German philosophers and psychologists in the mid 19th century. In 1895 (in the English translation) Oswald Külpe[24] describes how emotion fuses things together:

> (…) feeling and sensation, according to Külpe, are the elements of conscious processes from which all other mental products are formed, either by 'fusion,' in which the constituent elements suffer loss of distinctness, or by ' colligation,' in which the combined elements gain in distinctness. (Külpe, quoted by Angell 1896: 419)

Yet more detailed is E. B. Titchener's[25] description of how the German philosopher Wundt defines colligation (I gather Titchener has translated this from the original German of *Beiträge zur Theorie der Sinneswahrnehmung*, 1862, as he makes no reference to an English edition):

> The mind takes cognisance of this paired relationship by an unconscious act of colligation, a form of induction by simple enumeration. Since A has, a thousand times

[24] I was only able to find quotations of the original text.
[25] ditto.

> over, been followed immediately by a, and B by b, the mind argues that A will in the future always be attended by a, and B by b; or, in general, that an objective sensation will always be attended by a subjective sensation. We are still far removed from perception; the combinations Aa, Bb, leave the component sensations A, a, B, b, just what they were; but we have, in the act of colligation, taken the first step toward perception. (Wundt 1862 quoted in Titchener 1922: 351)

I quote this at length to make clear the link between the use of the term in linguistics and the use of the term colligation amongst early psychologists. While the definition of colligation is much broader, as with the concept of Lexical Priming it indicates that we are dealing with a psychological concept. Its early definition appears, at the end of the 20th century, to be brought back into use again and tightened up for a new purpose.

The OED definition, however, seems to be less specific and more Firthian in its definition:

> In Linguistics: be or cause to be juxtaposed or grouped in a syntactic relation:
> [intrans.] the two grammatical items are said to colligate | [trans.] pronouns are regularly colligated with verbal forms.

While traditional grammar used prototypical concepts of colligation like pre-nominal slots (cf. Bache: 1978), corpus linguists could now underpin claims with empirical data.

Based on the work of language use in context by Malinowski, Firth and his colleagues[26] make use of the term. Firth describes colligation as such:

> Colligation represents the syntactic juxtaposition of two or more grammatical categories. Colligation is derived from the concept of collocation which is the means of stating the 'meaning' of the word according to the habitual company it keeps; there is however no necessary relationship between colligation and collocation.
>
> (Firth quoted in Bursill-Hall 1960: 247)

This term has been brought, by Sinclair and Hoey, into wider circulation, while credit must go to Halliday for keeping the notion of colligation a live one – in particularly in the context of language learning. From the 1960s on, however, linguists were concentrating more on other theories and, until the recent rise of corpus linguistics, the concept of colligation (though not to the same extent collocation) lay dormant.

Indeed, trawling through all the related literature, mention of colligation is only fleeting (see e.g. Greenbaum 1988) but it appears to be not in use as a major concept. It is not clear who brought the term back into the discussion. Lia (2004) makes a reference to a work by Bahns in 1993[27].

[26] The "London Linguists". These include A.E. Sharp who speaks of colligation and also H.F. Simon, who uses terms like colligates in connection with standard Chinese (Some Remarks on the Structure of the Verb Complex in Standard Chinese. London: SOAS. 1958). Simon makes reference to both Firth & Halliday.

[27] However, Bahns never refers to *colligation* as claimed. Furthermore, he speaks of *grammatical collocation*, a concept different from *lexical collocation* and, crucially, different from colligation altogether.

We find that John Sinclair in his 1991 book *Corpus Concordance Collocation* concentrates on just these three themes. In his paper *Trust the Text* from 1990 he discusses the issue of delexicalisation. To my mind, this work is a first step towards the way of defining colligation the way he does later:

> The meaning of words chosen together is different from their independent meanings. They are partly delexicalised. This is the necessary correlate of co-selection.
>
> (…)
>
> We are given to understand in grammar that adjectives add something to the noun, or restrict the noun … That is no doubt true in some cases, but in the everyday use of adjectives there is often evidence rather of co-selection and shared meaning with the noun.
>
> (Sinclair [1990] 2004: 20)

Sinclair gives here the example of adjectives, in traditional English grammar seen as an independent item from the noun. Sinclair highlights that this is probably only true in a minority of cases. From the co-selection given above, the Lexical Item – a unit larger than the word – as well as the Idiom Principle is an obvious extension. It can be assumed that the next development from here is colligation, the way Sinclair defines it (in contrast to Firth's definition). Starting with the publication of his papers *The search for units of meaning* and The *Lexical Item* in 1996/97, Sinclair starts devoting more time to defining and working with the concept of colligation. This appears first when he discusses the lexical item "naked eye":

> … you can see with the naked eye … / just visible to the naked eye …
>
> The other prepositions are *by, from, as, upon & than* (…). The word class 'preposition' is thus an inherent component of the phrase, accounting for over 90 % of all cases.
>
> What we have done, …, is to change our criterion from collocation to *colligation*, the co-occurrence of grammatical choices (Firth 1957b) to account for greater variation.
>
> (Sinclair [1996] 2004: 32)

Though he does not make it explicit here, he actually diverges from Firth in linking the grammatical choice very clearly to a lexical necessity and therefore moves away from the split of lexis versus grammar that Firth still upheld.

Indeed, Sinclair puts colligation squarely in the middle of a continuum:

**word ⇒ collocation ⇒ colligation ⇒ semantic preference ⇒ lexical item** [28]

In *The Lexical Item* (1997) Sinclair spells out more succinctly what the hurdles are to move from a traditional view of grammar to the lexis-based axiom – and how disparate parts can fit together.

> … the tradition of linguistic theory has been massively biased in favour of the *paradigmatic* rather than the *syntagmatic* dimension. Text is essentially perceived as a series of relatively independent choices of one item after another, and the patterns of combination have been seriously undervalued.
>
> (…)
>
> Word gives information through its being chosen (paradigmatic) and at the same time it is part of the realisation of a larger item (syntagmatic) (Sinclair [1997] 2004: 140f.)

[28] A discussion of "semantic preference" follows in 3.2.1.4. Hoey would extend this (Hoey: 2005) by adding a further step: *NESTING*. The concept of nesting, implies a less linear, more cluster-like relationship where collocations and colligations of the same sets of words can form different relationships.

These two approaches can be combined:

> .. the two axes of patterning, the paradigmatic and the syntagmatic, are related; the relationship is … quantifiable. The three categories that relate words together on either dimension are *collocation, colligation* and *semantic preference.*
>
> (Sinclair [1997] 2004: 141)

This links in with what Sinclair has described in *The search for units of meaning*.

Susan Hunston (2001) refers back to this when she defines the term colligation:

> "Colligation" is a term coined by Firth but little used since then. (…) If we take seriously Sinclair's assertion that there is no longer sense in distinguishing between lexis and grammar (1991:3), then the distinction between collocation and colligation to a large extent disappears. On the other hand, the term "colligation" is helpful in drawing attention to the fact that the evidence of many instances of naturally-occurring language can be used to explain behaviour that is traditionally associated with grammar.
>
> (Hunston 2001: 15)

Nelson (2000), in his unpublished PhD thesis, totally sidesteps Sinclair when talking about colligation, and repeatedly refers to Hoey (1997[29]). Consequently he quotes:

> Hoey (1997) further divided colligation itself into two main classes:
>
> *Textual position*: The notion that a lexical item may have a strong tendency to occur in a

[29] Nelson refers to the paper that laid the foundations for LP: Hoey, M. (1997). From Concordance to Text Structure: New Uses for Computer Corpora. In: Melia, J. & Lewandoska, B. (eds) *Proceedings of PALC 97*. Lodz: Lodz University Press.

> certain textual position rather than others, e.g. at the beginning or end of a text.
>
> *Grammatical context*: A lexical item will tend to 'co-occur with a particular grammatical category of items' (1997:4). The implication of this is that when a word has more than one sense, each sense is found in a different grammatical context, with sense and a specific grammatical context in a direct relationship. (Nelson: 2000. p. 148)

This highlights two important points of Hoey's work: that words can be found in a physical location (*textual position*) as well as in a grammatical context to disambiguate their meaning. This goes beyond the mere collocation of words – and Hoey (1997) suggests that it therefore makes little sense to treat lexical and grammatical relationships as the same – or to give them the same name. This led to the 'Drinking Problem' hypotheses[30]:

> a) Where it can be shown that a common sense of a word favours common colligations, then the rare sense of the word will avoid those colligations.
>
> b) Where two senses of a word are approximately as common (or as rare) as each other then both will avoid colligational patterns of each other.
>
> c) Where either a) or b) do not apply, the effect will be humour, ambiguity (momentary or permanent), or a new combining of the two senses.[31] (Hoey1997: 12)

This shows that a word, if it is to be used unambiguously, will prefer its restricted colligations. Since 1997, Hoey has added other kinds of association. At the same time, however, the door for creative use of language – one of the main features of language *per se* – is still left open

[30] See http://www.natcorp.ox.ac.uk/archive/reports/birm_sem.html for an explanation of these "Whimsically termed hypotheses" (Hoey 2005: 82)

[31] Point c) echoes, most probably with intent, Louw (1993) – see 3.4.1.4.

while its consequences are described. At the same time it becomes obvious that Sinclair and Hoey have developed, independently of each other, and both building on the works of Firth and Halliday, a closely resembling definition of their use of colligation.

Building on Hoey[32], Susan Hunston (2001), highlights in her conclusion:

> In Hoey's terms, the paper has attempted to illustrate how colligation - the grammatical behaviour of a word in its various senses – links together not only those concerns traditionally treated as "lexis" and "grammar", but also those concerns traditionally discussed as "text". *It has also demonstrated one half of the phenomenon of repetition (cohesion being the other half), that is, that the phraseology of an individual text repeats the phraseology of innumerable other texts, and derives meaning from this repetition.* (my italics – MP-S) (Hunston 2001: 31)

Hunston, in her evaluation of Hoey's work, does two things. First of all, her own research proves the viability of Hoey's ideas with regards to colligation. Secondly, in saying that "phraseology of an individual text repeats the phraseology of innumerable other texts, and derives meaning from this repetition" she already foreshadows one of the key planks of the Lexical Priming theory – namely that meaning lies in sequences of words and this meaning is created through repetition.

Stubbs (1996) in his discussion of co-selection and lexico-grammar suggests why both Hoey and Sinclair came to the same conclusion as to what colligation should be:

[32] Hunston actually refers to the Pit Corder Lecture Michael Hoey delivered at BAAL 1998

> Quirk *et al.* (1985) imply by omission that such exposition [of the clausal object in sentences] is possible for any verb. But corpus data show (Francis, 1993) that two verb lemmas, FIND and MAKE, account for the vast majority (over 98 per cent) of such structures. Such strong probabilistic relations between lexis and syntax should find a place in grammar. (Stubbs 1996: 40)

This shows that intensive corpus work made visible strong correlations of place and grammatical context of a word. As Hoey (2005: 43) says – "colligation *may* simply be an idea whose time has come".

Finally, in 2005, Hoey gives a tighter definition of the use of colligation in conjunction with Lexical Priming:

> 1 the grammatical company a word or word sequence keeps (or avoids keeping) either within its own group or at a higher rank;
> 2 the grammatical functions preferred or avoided by the group in which the word or word sequence participates;
> 3 the place in a sequence that a word or word sequence prefers (or avoids).
>
> (Hoey 2005: 43)

This does not preclude the creative openness given in the 'Drinking Problem' hypotheses. It is important to note, though, that Hoey extends colligational properties beyond a single word – he speaks of word sequences, a concept close to Sinclair's Lexical Item. These sequences are often (though not always) appearing in the form of collocational clusters. Hoey (1997) defines colligation as the *grammatical company* a word or sequence *either prefers or avoids*. Preference, it is important to note, does

not mean total prescription – the company a word or a cluster of words prefers can form a highly probable pattern of occurrence.

### 3.2.1.4 Semantic Prosody, Preference and Association

As mentioned earlier, the Language of the Chinese "Secret History of the Mongols" enabled Halliday to use techniques now familiar to corpus linguists – mainly counting keywords and highlighting occurrence patterns. More interesting still, (particular in the light of what is going to be discussed in 3.2.3.3) is Halliday's work on paragraph initial key words.[33] In my view, this lays the groundwork for his later research into cohesion patterns in text. Halliday notes that the original text is graphically divided into "chapter and "paragraph". Below the level of the paragraph, the Mongolian language has the "word" and then the "character" (Halliday 1959: 29). Pointing out paragraph-initial patterns, Halliday finds the following:

> *Certain pieces, defined by position in the paragraph, display features marking them off statistically from the pieces as a whole.* If we take the final piece of each paragraph and compare the frequency of occurrence of certain elements, commonly found as piece-final, in these 282 pieces with their frequency in the [total of] 5,386 pieces of the whole work, we find striking differences. (my highlighting – M P-S), (Halliday1959: 23)

Halliday goes on to give more detailed percentages. Hoey himself confirms that Halliday has strongly influenced his thinking in this area[34].

[33] Hoey returned to this subject in the AHRC funded *textual priming project*: http://www.lexical priming.org/textual-priming-project/ (last accessed 11/05/09)
[34] Personal communication.

This becomes obvious when the following quotes on paragraph initial position below are compared:

> In general, any two lexical items having similar patterns of collocation – that is, tending to appear in similar contexts – will generate a cohesive force if they occur in adjacent pairs. (Halliday & Hasan 1976: 286)

Pieces of a jigsaw start fitting together. Halliday & Hasan' s book *Cohesion in English* proved to be a milestone in text-linguistic research. For them, cohesion and collocation are closely linked, and this enabled Hoey to build on their work and push the limits even further:

> More radically, (…) for example, a lexical item may have a preference or aversion to appearing in paragraph initial position. (Hoey 2002: 3)

The next step towards the development of the Lexical Priming Theory was, I believe, the rejection of sentence grammar. (cf. Winter: 1982; Brazil: 1995). Eugene Winter, to whom Michael Hoey was research assistant in the early 1970s, is widely quoted in connection with Lexical Priming. Winter (1977) speaks of clause relations, a concept discussed by Winter in 1971, 1974, 1977 and 1979, and revised later. This led to Winter's thoughts about clause operations:

> The notion of lexical choice means the selection of items from the open-ended vocabularies of nouns, verbs, adjectives and adverbs as head as well as their pre- and postmodifying structures. Lexical selection at its most simple generally means selecting lexical items as constrained by the autonomous grammar of the constituents of clause and its grouping elements. (Winter 1982: 37)

Hoey had this in mind as he formulated his "second claim" (Hoey 2003: 401) that "every lexical item is primed to occur as part of a textual

semantic relation". Hoey also says that Winter's work on clause operations may have also influenced his stance on colligation.[35] Winter and Halliday worked together[36] and this claim can be seen as a subsequent extension of Halliday's findings described above.

As parts of the larger theory come together, these individual influences stand out as the foundations for what was to come.

According to Partington (1998), Sinclair (1987) proposed that a word may carry meaning in association with others. For this, he borrowed a term from phonology (used by Firth in 1957): *prosody*. Partington describes prosody as follows:

> Often a favourable or unfavourable connotation is not contained in a single item, but is expressed by that item in association with others, with its collocates. A clear example is the word *commit,* which, ..., collocates with items of an unpleasant nature.
>
> (Partington 1997: 66)

This defines the issue that Sinclair discusses in *The units of meaning*. Words have little or no meaning by themselves, yet in "association with others"[37] a positive or a negative meaning is communicated. Consequently, certain word combinations are preferred while others would be seen as unusual (dispreferred or, as Hoey (2005; 2008a,b) would say "breaking the priming"). One can say *to commit a murder* while one

[35] personal communication.

[36] Halliday, like most others, refers to *semantic relations* rather than *clause relations.* I take it Winter wants to highlight that the relation is more tightly defined and related to the syntax of the sentence.

[37] Note the link of Partington's (1998) "association with others, with its collocates" and the term chosen by Hoey (2005) *semantic association*.

avoids saying **to commit charitable works*. This "bad company" or "good company" that a word keeps Sinclair calls **semantic prosody.**

A first detailed study of the uses of prosody was undertaken by Louw in 1993 and this has subsequently become the point of reference, for in it,

> Louw investigates how writers sometimes diverge from "the expected profiles of semantic prosodies", that is, how they upset these normal collocational patterns.
>
> (Partington1997: 68)

In Louw's own words, computing technology brought prosodies out into the open:

> Semantic prosodies have, in large measure and for thousands of years, remained hidden from our perception and inaccessible to our intuition. … At present, (computer held) corpora are just large enough to allow us to extract profiles of semantic prosodies from them.
>
> (Louw 1993, quoted in Partington 1997: 69)

From Louw's work on how normal collocational patterns are "upset" by writers there is a link to Hoey's 'Drinking Problem' Hypothesis, which is an example of a breach of an expected colligational pattern can be used for humorous reasons.

Yet another definition of Semantic Prosody is given by O'Keefe *et al.*:

> …words as well as having typical collocates (for example *blonde* typically collocates with *hair* but not with *car*), tend to occur in particular environments, in a way that their meaning, especially their connotative and attitudinal meanings, seem to spread over several words. (O'Keefe et al. 2007: 14)

However, it can be said that this only highlights the difficulty of giving a clear-cut definition of the term as this appears to blur the boundaries

between Semantic Prosody, Semantic Preference and Semantic Association.

John Sinclair agrees with Bill Louw's formulation of "prosodies having remained hidden to the lexicographer's *naked eye*[38]". This is seen by Sinclair (in *The units of meaning)* as a semantic feature that can be illuminated by a single occurrence of any corpus (as long as this has the selected semantic feature):

> Whatever the word-class, whatever the collocation, almost all of the instances with a proposition at N-2 have a word or phrase to do with visibility either at N-3 or nearby. This new criterion is another step removed from the actual words in the text, just as colligation is one step more abstract than collocation. (Sinclair [1997] 2004: 32)

Sinclair also points out that *"… this feature is relevant in the same way to both syntagmatic and paradigmatic phenomena."* (Sinclair [1997] 2004: 142)

Xiao & McEnery (2006) point out the closeness of use of the terms semantic *prosody* and *semantic preference* and highlight that the concept can easily be applied to languages other than English:

> Our contrastive analysis shows that semantic prosody and semantic preference are as observable in Chinese as they are in English. As the semantic prosodies of near synonyms and the semantic preferences of their collocates are different, near synonyms are normally not interchangeable in either language. (Xiao & McEnery 2006: 124f.)

[38] A word sequence which Sinclair discussed to explain his concept of Semantic Preference.

Michael Stubbs has done intensive work on what he terms "the varying levels of structure of prosody" (Stubbs 1996; 2001(a); 2001(b); 2006; 2008a) and expands on the work by Sinclair. Stubbs draws our attention to the fact that Sinclair's definition of semantic prosody is bound to language use and draws a bridge to speech-act-theory:

> Austin argues that all utterances have an illocutionary force and Sinclair argues that all extended lexical units have a semantic prosody (which is a way of modelling the reason for speaking). Searle (1995) has developed a … concept of agency, but, since he uses no data on language use, he can only discuss speech act forces based on introspection. It is only corpora which can provide data for studying prosodies from the bottom up, and therefore show how we could do real 'ordinary language philosophy'.
>
> (Stubbs 2006: 26)

He opens up the prospect that language philosophy can be grounded in empirical facts.

Stubbs, furthermore, reasons that "semantic prosodies have pragmatic and textural functions." He declares: "For this reason, I prefer the term 'discourse prosody'" (Stubbs 2008a: 178). Like Michael Hoey (see below) he appears to have found limitations in the earlier definitions of the term and explains the structure as follows:

| | | | | |
|---|---|---|---|---|
| 1 | collocation | lexis | tokens | co-occurring word forms |
| 2 | colligation | syntax | classes | co-occurring grammatical classes |
| 3 | semantic preference | semantics | topics | lexical field, similarity of meaning |
| 4 | discourse prosody | pragmatics | motivation | communicative purpose |

Adapted from Stubbs: 2008a, p.179

With this, Stubbs shows that prosody and preference are in an 'increasingly abstract' field. These terms no longer describe simple phenomena of co-occurrence that first-level concordance analysis would show. While semantic preference is looking at the word-field that is common with the node (or target) term and therefore looks at something familiar to traditional linguistics, "discourse prosody" is cultural; it expresses the background and attitude of the user. Hence the term *motivation* used by Stubbs.

The terminology has been problematicised by Whitsitt (2006)[39], and, as one consequence, Hunston 'revisited' the concept in 2007 to come up with yet another term:

> … my own suggestion would be that the term 'semantic prosody' is best restricted to Sinclair's use of it to refer to the discourse function of a unit of meaning, something that is resistant to precise articulation and that may well not be definable as simply 'positive' or 'negative'. I would suggest that a different term, such as 'semantic preference' or perhaps 'attitudinal preference', is used to refer to the frequent co-occurrence of a lexical item with items expressing a particular evaluative meaning. On the other hand, as 'prosody' and 'preference' are both metaphors, more transparent terminology in both cases might be less open to confusion. (Hunston 2007: 266)

With this, Hunston defines *semantic prosody* as a discourse function while *semantic preference* has to do with terms found as a frequent co-occurrent that expresses a form of evaluation. Yet it is the non-specificity

[39] The concept of *semantic prosody* remains widely debated and, at the time of writing, the discussion carries on – with both Louw and Whittsitt continuing to write about it.

of either term that makes it hard to use one or the other to describe phenomena found in language.

Whitsitt's criticism of the use of the term *semantic prosody* by Louw, which seems in one way aligning prosody with connotation and in another way with metaphor apparently paves the way for Hoey's redefinition of the term:

> Hoey suggests an alternative, and he does so with a change in metaphors. As is known, a change in metaphors can indicate a shift in paradigms of thought, and Hoey's introduction of the metaphor of "priming" does offer an alternative (…). It seems that Hoey might be thinking more in terms of priming something which has been, as he puts it, "loaded" (2003:1). What needs to be stressed, however, is the very significant point Hoey makes that our expectations, which may even explain why we have collocations, is not sustained by linguistic or semantic principles. (Whitsitt 2005: 298)

Michael Hoey, developing on this, chooses a different approach from that of Louw and Sinclair and is, by his own admission, closer to Stubbs' definition.

Instead of splitting up the less-direct, implied-meaning qualities into smaller defined groups, he groups semantic preference and semantic prosody under the umbrella term of **semantic association.** Hoey argues that

> My reason for not using Sinclair's term (Semantic Preference – MP-S) is that one of central features of priming is that it leads to a psychological preference on the part of the language user; to talk of both the user and the word having preferences would on occasion lead to confusion. (Hoey 2005: 24)

This connects neatly with Whitsitt's description. Indeed, by focussing the psychological component of word choice, the selection of the term 'association' is probably very fitting. His definition therefore is:

> (**semantic association)** exists when a word or word sequence is associated in the mind of a language user with a semantic set or class, some members of which are also collocates for that user. (Hoey 2005: 24)

It is a definition that is remarkably open and reflects Hoey's thinking that the language first of all resides in the individual user.

Dominic Stewart (2010), discusses *Semantic Prosody and Lexical Priming* in great detail and states:

> Hoey illustrates that from its point of departure a word takes wing beyond recall, and that priming gains much of its strength from its ability to go beyond the phrase, sentence and textual chunk. It is my view that we can take these characteristics of priming and apply them, to a degree, to the various descriptions of semantic prosody. (…) Indeed, Hoey's notions of semantic and pragmatic association are (…) more nuanced. (Stewart 2010: 156)

This illustrates that Stewart sees forms of prosody as intrinsically linked to priming and that, furthermore, Hoey links together the various forms of *semantic prosody* that Stewart discusses in his book in a satisfactory manner.

### 3.2.2 *A brief description of* Lexical Priming

Lexical Priming is a theory that has been developed by Michael Hoey since the mid 1990s. Though it was not yet referred to as *lexical priming*, Hoey's work on *bonding* already provided a framework for what would become the *Lexical Priming* Theory[40]:

> What we are now contemplating, (…), is the possibility of finding bonding across texts written between three and fourteen years apart, solely because of the mental concordances of the authors retained records of the texts they had read, which in turn were written in the light of *their* author's mental concordances, which (perhaps) included sentences drawn from a common primary source (author's highlights)
>
> (Hoey 1995: 90)

These *mental concordances* would later be seen by Hoey as having been created through the process of priming.

According to Hoey[41] publication and development of the theory started with the talk[42] given at PALC (University of Lodz, April 12-14 1997), Poland. The following year, Hoey delivered the Pit Corder lecture at the Annual Meeting of the British Association for Applied Linguistics. One direct response was Hunston (2001) *Colligation, lexis, pattern and text*. This paper has been discussed in detail above. Hunston combines Hoey's ideas with the work she has done on Pattern Grammar. This paper is still mainly concerned with colligation:

[40] This is a genuine find by me. Hoey thought that *Lexical Priming* had been a new departure. (Personal communication). I show, however, that the basic idea is a development of the notion of *bonding* worked on by Hoey (1991; 1995)
[41] Personal communication
[42] Later published as Hoey: 1997

> Cohesion and colligation are themselves connected, as each depends upon repetition. Cohesion depends on repetition within the text (as Hoey draws on his own work in Hoey 1983;1991 here), while colligation depends on repetition between the text and other texts … (Hunston 2001: 14f)

The first texts referring to the process of *priming* appeared during 2002/2003 in a number of papers by Hoey. While each one of these was drawing on and building from its predecessor, each paper highlighted a different angle of the theory. First there was *Lexis as Choice* (2002). During ICAME 2002 in Götheburg, Sweden, Hoey was still referring to *Textual Colligation* and it had the subtitle *A special kind of priming*. Then, in 2003, priming actually appeared in the title: *Lexical Priming and the properties of text* and *Why grammar is beyond belief*. All these in turn led to the publication of the monograph *Lexical Priming* in 2005, which discusses the issue in-depth.

Priming itself will not be discussed in this section, as a separate part is reserved for that. That *priming* – a subconscious forming of the ability to relate entities to each other – and language structure based on how words link up with each other (*collocate)*, go together, has most succinctly been put by Michael Stubbs:

> Examples of collocation show that there is much in language use which is automatic and unconscious. This means that introspection about lexical meaning is often unreliable or at least incomplete, and also that, in terms of its automaticity, lexis is much like syntax. (Stubbs 2001b: 89 also in 2006: 26f.)

This highlights some more important points. Stubbs, like Sinclair and Hoey, find increasing evidence that it is the lexis that structures the grammatical structure (rather than the other way around). Lexical Priming neither operates in nor follows a fully pre-determined universal pattern, as Hoey is the first to admit:

> … grammars exist as a product of our primings. Each of us, presumably to different extents and with different outcomes and different degrees of regularity, constructs a grammar – leaky, inconsistent, incomplete – out of the primings we have for the sounds, words, phrases and so on that we encounter. This grammar, or perhaps one should say grammars, may in turn be used to regulate and remark on our linguistic choices.
>
> (Hoey 2008b: 7)

This particular summary of Hoey's theory points to one crucial quality of priming: it is something that exists within the individual first of all. However, as social beings and as integral part of all our animate and inanimate surroundings, we are touched, influenced and formed by what we are exposed to. Language is no exception. Would this contradict the validity of the theory, given that every speaker would identify her/himself first as a native speaker of a (or a set of) specific language(s)? In short, the answer is no, as the sum of individual primings create the "leaky" fuzzy total of any given form of communication. Should an individual priming or grammar fall too much out of the boundaries of acceptability, communication would no longer be effective. At the same time, primings are the product of encounters with other people, who themselves have

been through the process of having encountered for themselves what the "norms" of effective communication are.

Early on in his book, Hoey (2005) draws our attention to the hypotheses on which his Lexical Priming theory is based:

**Priming hypotheses**

Every word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word. If one of the effects of the initial Priming is that regular word sequences are constructed, these are also in turn Primed. More specifically:

1 Every word is primed to occur with particular other words; these are its collocates.

2 Every word is primed to occur with particular semantic sets; these are its semantic associations.

3 Every word is primed to occur in association with particular pragmatic functions; these are its pragmatic associations.

4 Every word is primed to occur in (or avoid) certain grammatical positions, and to occur in (or avoid) certain grammatical functions; these are its colligations.

5 Co-hyponyms and synonyms differ with respect to their collocations, semantic associations and colligations.

6 When a word is polysemous, the collocations, semantic associations and colligations of one sense of the word differ from those of its other senses.

7 Every word is primed for use in one or more grammatical roles; these are its grammatical categories.

8 Every word is primed to participate in, or avoid, particular types of cohesive relation in a discourse; these are its textual collocations.

9 Every word is primed to occur in particular semantic relations in the discourse; these are its textual semantic associations.

10 Every word is primed to occur in, or avoid, certain positions within the discourse; these are its textual colligations.

Very importantly, all these claims are in the first place constrained by domain and/ or genre.

(Hoey 2005: 13)

As can be seen, Hoey tries to cover every occasion, location and opportunity in which a word or word-sequence could be employed. The very apparent repetition of the term 'word' (or "keyness" should one compare *Lexical Priming* to similar texts) indicates how the lexis is seen

as the centre of this theory. Equally, the consistent use of the cluster "is primed" indicates how Hoey might prime the readers themselves.

It is Biber (2009) who found in his research that spoken language is more formulaic than (academic) written language:

> Conversation
>
> – Most lexical bundles are sequences rather than frames
>
> – Both variable and fixed slots are usually function words
>
> – Content words are highly restricted Biber (2009)

By contrast, Biber points out that in academic writing "high frequency patterns tend to be frames". This means that a fixed colligational structure allows for a greater lexical variation. In spoken language, however, *formulaic chunks* are far more prevalent. This is explained by the ad-hoc nature of spoken language production:

> Psycholinguistic implications -
>
> - In speech, lexical sequences -- including content words -- stored and used as chunks
> - In writing, frames stored separately from content words
> - Many content words select a single frame
> - But frames associated with a large set of possible content words
> - Other (most?) content words are not associated strongly with a frame
>
> Biber (2009)

Finding "lexical sequences stored and used as chunks" in spoken language use provides a link between Biber's (2009) research and Hoey's

(2005) claims. It also provides a further reason why it is essential to use speech to test the validity of *lexical priming* theory.

In this thesis, I will take the Lexical Priming hypotheses as the background to my research. If lexical priming is a valid theory, it should be applicable not just to the written word as found in the Guardian corpus (Hoey 2005) but should also be applicable to language as **spoken** by any given **speech community**:

> "..A word's likely primings for a particular set of members of a speech community must be limited to the genre(s) and domain(s) from which the evidence has been drawn. For this reason, indeed, specialised corpora may be more revealing than general corpora." (Hoey 2008: 9f.)

The corpus on which my research is based is just this kind of corpus: specialised, drawn from a specific speech community and limited in its genre.

### 3.2.3 Lexical Priming issues

Reviewers of the book *Lexical Priming* (Hoey: 2005) have noted the failure to mention either Harold Palmers' work on collocation (a link I try to make in this chapter) or any mention of Alison Wray and her work on psychological explanations for language acquisition. Wray looks at the mental storage of chunks, while Hoey focuses on the individual word and

its primings for the individual. Wray herself indicates her debt to the work of Nick Ellis – whose work Hoey was not really aware of until they heard each others' presentations in 2006 (see 3.5).

Furthermore, Hoey appears to limit priming to too narrow an area in language. In this thesis, I point out that in spoken language, the use and length of pauses and (some) hesitancy markers indicate primed speech behaviour. Salim (forthcoming)[43] describes the evidence she found that punctuation marks follow a primed pattern. She also found that, in religious texts, whenever *God* is mentioned in the Qur'an the words it is nesting in are very similar regardless of the form of address used (Lord; the Almighty; etc.). Yurchak (2006) describes how official Soviet texts became fossilized in form and through constant re-use of formulas to a point where content no longer mattered. This could be seen as a form as hyper-priming. Hoey (2005) fails to mention that such forms of overuse can, on occasion, lead to a breakdown of communicative competence.

It must be said that Lexical *Priming* gives very little space to the psychological research that has been undertaken to describe and prove the existence of *priming*. In the second half of this chapter, where I focus on the research done into *Artificial Intelligence*, *psychological* and *psycholinguistic* (theoretical and laboratory-based) research undertaken into investigating and defining *priming*, I try and rectify this[44].

[43] unpublished PhD thesis, University of Liverpool. 2010.
[44] See also Hoey (2008b) who in this later publication shows greater awareness of this.

## 3.3 Priming

As section 3.2 shows, Hoey's theory of *lexical priming* is firmly grounded in corpus linguistic work done prior to his development of the theory. Yet while Hoey could be called assiduous as to his corpus linguistic pedigree, his book shows far too little regard for earlier research into the (psychological) concept of *priming* itself.

In this section, therefore, *priming* will be defined and the historical background to *priming* research will be given.

The *Sage Handbook of Social Psychology* provides the following characterisation of priming:

> Another factor that influences the accessibility of information in memory is priming. The activation of stored knowledge through experiences in the immediate context can make prime-relevant information more accessible in memory, and such recent construct activation can influence inferences, evaluations, and decisions on subsequent tasks (Bargh and Pietromonaco, 1982; Bargh et al., 1986; Devine, 1989; Higgins et al., 1977, 1985; Sherman et al., 1990; Srull and Wyer, 1979). A second factor that influences the accessibility of information in memory is the frequency with which a construct has been primed (Bargh and Pietromonaco, 1982; Srull and Wyer, 1979).
>
> Traits, attitudes, or stereotypes that have been frequently activated in past experience are more available in memory than those that have been less frequently primed. Such frequency of activation, if it occurs on a regular and continuing basis, can result in certain constructs becoming chronically accessible, such that no external priming in the immediate context is necessary to make them highly accessible (Higgins et al., 1982). Moreover, because people differ in the kinds of experiences they have that would

> generate such routine construct activation, individuals quite naturally differ in the particular constructs that are chronically accessible (Bargh et al., 1986; Markus, 1977).
>
> (Sherman *et al*: 2003: 55)

This entry highlights all the relevant aspects of the notion of *priming*. Sherman *et al.* describe how the human brain does not *access memory* in a *random* way, since information can be accessed all the easier when it can be linked to other known information. This link is made all the better the more (often) a person absorbs the same (or slight variations of) connected information.

*Priming* as such is not a linguistic but a psychological concept. Though it appears as if most research focuses on *lexical priming* (where test under laboratory conditions are undertaken with words) the wider application of *priming* is widely acknowledged. (See, for example Habib:2001[45]). The term does, however, not appear until the later 20th century.

The early literature in which the term appears seems to be mostly concerned with the *priming* of language – words read and heard. According to Collins and Loftus (1975) it was Ross M. Quillian who first used the term: "Quillian's theory of **semantic memory** search and semantic preparation, or **priming**" (my highlighting). This refers to papers Quillian produced between 1961 and 1969. As can be seen, Quillian (1961, 1962, 1966, 1967, 1969 and Collins & Quillian 1969) laid

[45] "Priming" designates hypothetical processes that underlie the priming effect, the empirical finding that identification of objects is facilitated by the individual's previous encounter with the same or similar *objects*. (my italics). (Habib 2001:188)

the groundwork for all the research to come in the field of priming since the early 1960s. Papers written by Quillian and Collins (1969) and Collins (1969; 1970; 1972 (a/b); 1975) where these two looked at the process they name "retrieval from the semantic memory", and this book and these papers will be discussed in some detail below. All the seminal works that past and current research is based on go back, in one way or another, to this early research.

This led to investigations by Meyer & Schvanefeldt (1971) on whom Posner & Snyder (1975) in turn based their research. James H. Neely's (1976 & 1977) papers[46] [47] are entitled *Semantic priming and retrieval from lexical memory*. Neely very clearly refers to the work of these researchers as his main influence[48].

*Priming* together with *Lexical* appears, however, to be first brought into discussion by James H. Neely. Neely (1976) links the research in the 1960s and 1970s to Hoey (2005 *etc.)*.

Psychologists and psycholinguists approach language very differently from other linguists. More often than not, they base their results on carefully planned and executed experiments that other researchers must be able to re-stage. As psycholinguistic research into priming developed, a change in the investigators' methods becomes apparent. In early research, a key word is followed by another (single) word. *Priming* becomes apparent by the first term preparing for the comprehension of the next.

[46] Which Hoey refers to in *Lexical Priming* - 2005: 8

[47] Michael Hoey (2005) names the same paper as published in 1977. JHN published two parts of the same paper (with different subtitles) in two different publications in two consecutive years.

[48] Neely (1991) describes how Posner & Snyder's work was his main influence and how they were influenced by Meyer & Schvanefeld. All four appear in the bibliography of Neely (1977).

Later work became concerned with larger units within the text – syntactic or phonological priming.

Psycholinguistic methods appear to contrast with those used in corpus linguistics research, yet Gries notes that corpora have been used for psycholinguistic research since 1997 (cf. Gries 2009: 222).[49]

### 3.3.1 M Ross Quillian and the language learning machine

A researcher in Artificial Intelligence, M.R. Quillian (1962; 1969) describes, in theory, how to construct an Understanding *Machine* (1962), a *Teachable Language Comprehender* (1969). Talking about language translation, he states:

> … human translators do not translate "directly", and … really good mechanical ones cannot hope to either. (Quillian 1962: 17)

In providing the theoretical blueprint for a *mechanical translator*, he tries to simulate how the human mind learns language.[50] While the term *priming* is not yet introduced, Quillian deals with a number of issues that will resurface, over forty years later, in Hoey's Lexical Priming.

An initial concern of Quillian was how to deal with polysemy.

[49] Gries (2005) also turned to CL to look at priming. To his obvious amazement, all corpus-based results agreed at a rate of more than 90% with the experimental results. See chapter 3.4.3

[50] "The program's strategy is presented as a general theory of language comprehension." Quillian (1969)

> The resolution of a polysemantic ambiguity, by whatever method of translation, ultimately consists of exploiting clues in the words, sentences or paragraphs of text that surround the polysemantic word, clues which make certain of its alternate meanings impossible, and, generally, leave only one of its meanings appropriate for that particular context. The location and arrangement in which we find such clues is itself a clue, or rather a set of clues, which we may call syntactic clues. (Quillian 1962: 17)

His theoretical outline foreshadows Hoey's work. The problem of polysemy exists in an ambiguous sentence like "He reached the bank" but not in "He got a loan from the bank". In the latter, the clues are sufficient, as Quillian describes:

> Thus, in our example, a reference to money is one such semantic clue, and one which, should it appear in the sentence, could be exploited no matter what word it occurred in, whether one of those on our list or not. (…) Learning to understand a language would consist of learning which readings on which scales should be activated in response to each word of that language. (Quillian 1962: 18)

This is the part of Lexical Priming referred to by Hoey as *semantic association*[51].

Quillian actively spurns transformational linguistics.[52] In line with Brazil (1995), he seems to prefer the concept of linear grammar to the idea of sentence grammar when he says –

[51] See chapter 3.2.2

[52] The relation between TLC, a semantic performance model, and the syntactic "competence" models of transformational linguistics (Chomsky, 1965) is not clear. The efforts that have been made so far to attach "semantics" to transformational models seem, to this writer at least, to have achieved little success (Quillian 1969)

> This seems to me a crucial advantage over those other approaches to mechanical translation which, lacking any manageable representation of meaning, have to proceed as though the only clues that are useful in resolving polysemantic ambiguities are those in grammatical features and their locations, or else in established idiomatic phrases. *That human beings do not so limit themselves, but also utilize semantic clues extensively, would appear obvious from the fact that people are able to understand language that is full of grammatical and syntactical errors.*[53]
>
> (Quillian 1962: 18. My italics – MP-S)

In fact, by the time Quillian (1969) discusses his *Teachable Language Comprehender* (TLC), he speaks of a machine that still had not entered active service in 2010: a machine reader that has built up a semantic web in its memory:

> This memory is a "semantic network" representing factual assertions about the world. The program also creates copies of the parts of its memory which have been found to relate to the new text, adapting and combining these copies to represent the meaning of the new text. By this means, the meaning of all text the program successfully comprehends is encoded into the same format as that of the memory. In this form it can be added into the memory. (Quillian1969: 459)

Though the wording is different, it does not sound unlike Hoey's *everything heard or read, everything said or written* (see below) that primes a person to use words in one way and not another. In his paper on the TLC, Quillian gives the example of a text that is easily comprehended because it is natural. This is similar to the example Hoey uses where he

[53] Meyer & Schvanefeldt (1976) suggest that Quillian is right in an experiment where words are made harder to read.

refers to *a three hour car ride* and then brings in an example of words not usually found together - *rides between Oslo and Hammerfest use thirty hours up in a bus*[54] – which is harder to comprehend. He concludes:

> What the reader must have, then, as he reads the text above, is an extremely versatile ability to recognize the appropriate chunk of memory information from among literally thousands of others he may since have learned about "Presidents," about "fruit trees," and about "fathers". (…) we assume that there is a common core process that underlies the reading of all text- newspapers, children's fiction, or whatever--and it is this core process that TLC attempts to model. (Quillian 1969: 461)[55]

This, I would claim, is the first step Quillian takes towards identifying *lexical priming* as a psychological process. In fact, Quillian proposes to prime the machine in a way similar to how a young person would be primed to figure out words in contexts. He proposes to give

> twenty different children's books dealing with firemen and have TLC read all of these [and reckons that the machine] will require less and less input as it accumulates knowledge. (Quillian1969: 464)

In his references to natural language, he goes well beyond that:

> Natural language text communicates by causing a reader to recall mental concepts that he already has. It refers him to such already known concepts either with isolated words or with short phrases, and then specifies or implies particular relations between these. (Quillian 1969: 474)

[54] Both examples are from Hoey 2005: 5

[55] I here use Hoey's examples as Quillian refers to a story of (President) George Washington who felled his father's cherry tree. While this is apparently a commonly known story in the USA, I find it a less suitable example in this context.

This appears to be very close to Sinclair's Idiom Principle[56] and also to the idea that collocations are recalled. In other words, in natural language the mind is *primed* to connect concepts on hearing or reading words and short phrases.

It might be argued, however, that Quillian simply philosophises over the problem. He does not quote other research, and he makes only a few references to other works. Neither are his descriptions backed up by successful experiments at this stage. However, he makes clear that he is providing a theoretical basis for building an actual machine. Most importantly, his ideas have stood the test of time and provided a theory that is still quoted by Artificial Intelligence (AI) researchers in the 21st century.

In fact, the following:

> Essentially, it asserts that to read text a comprehender searches his (her, its) memory, looking for properties which can be considered related to that text.
> (Quillian 1969: 474)

sounds remarkably familiar to those who have read these lines from Hoey (2005):

> I have talked of the language user as having a mental concordance and of the possibility that they process this concordance in ways not unrelated to those used in CL. (Hoey, 2005: 14)

[56] cf. chapter 3.2.1.2

Quillian reckons that his TLC is fully teachable – not by working on big structures but by learning piece by piece. The structure would thereby develop through what is feasible and what is not. Once we substitute Speaker / Writer for the term machine, it becomes clear that Quillian gives a good grounding for the priming research to come:

> Overall, the most distinctive features of this theory, as compared with other models and theories of language of which we are aware, are its explicitness and detail and its reliance on "knowledge of the world". (Quillian 1969: 475)

### 3.3.2 Facilitating access to the semantic memory

Moving on from the theory, Quillian and Collins (1969; 1970 & 1972a), discussing retrieval from the semantic memory, publish the results of a series of experiments. The last of these makes use of the term *priming*. The research involved checking the reaction times of volunteers to find out that *true sentences* (tennis is a game) have a shorter reaction time than *false*[57] ones (football is a lottery). They linked these findings to what was termed *semantic memory*:

> Priming is understood to be a process by which concepts and their meanings in semantic memory are activated, regardless of the origin of that activation.
> (Collins & Quillian: 1972. Quoted in Ashcraft 1976: 490)

[57] "True" and "false" sentences – their terminology.

This work in turn started a whole flurry of experiments by psycholinguists like Loftus (1973), Posner & Snyder (1975a), Collins & Quillian (1975) themselves, and Ashcraft (1976) and, significantly, led to the seminal paper by Meyer & Schvanefeldt, entitled *Facilitation in recognizing pairs of words. Evidence of a dependence between retrieval operations*. (1971).

The importance in the context of these studies is the phrase *pairs of words*, which links to J. R. Firth's notion of collocation, the importance of which has been highlighted also by Halliday (1959 *etc.*) Sinclair (1991) and Hoey (2005 *etc.)*. Meyer and Schvanefeldt's paper links an insight derived from psycholinguistic experimental evidence with a theoretical concept that has acquired significance in corpus linguistics.

In Meyer and Schvanefeldt's experiment, candidates have to link English words to *unassociated words* or *related words*.

> We showed that such decisions are faster when one word (e.g., 'nurse') is preceded by another semantically related word (e.g., 'doctor'). [than linked with a unassociated word, e.g. *bread* – MP-S]
>
> [Positive] responses averaged 85 ± 19 msec. faster for pairs of associated words than for pairs of unassociated words. (Meyer & Schvanefeldt [1971] 1984: 20)

The response time for collocates, therefore, was shown to be decisively quicker than the one for unrelated terms. This indicated that the mind of the reader / listener has a mental, subconsciously made connection

between these two *nodes*. Meyer and Schvanefeldt point out that "the results of [their experiment] suggest that degree of association is a powerful factor affecting lexical decisions in the (…) task." (Meyer & Schvanefeldt 1971: 229)

Sinclair's (1991) view that collocations mainly occur within 5 steps on either side of a word is an observation of how words appear in texts. That there is a possible link to how words are linked in one's memory finds support in the following results described by Meyer and Schvanefeldt:

> (…) responses to pairs of associated words would be faster than those to pairs of unassociated words. This follows because the proximity of associated words in the memory structure permits faster accessing of information for the second decision. The argument holds even if the accessed information is (*a*) sufficient *only* to determine whether a string is a word and (*b*) does not include aspects of its meaning.
>
> (Meyer & Schvanefeldt 1971: 232)

The key here is the *proximity of associated words* – one word acts as prime and the mind is already set to expect a limited set of options to follow. Meyer and Schvanefeldt go on to claim that this is a mental process that does not only reside in the short-term memory:

> (…) any retrieval operation $R_2$ that is required sufficiently soon after another operation $R_1$ will generally depend on $R_1$. This would mean that human long-term memory, like many bulk-storage devices, lacks the property known in the computer literature as *random access* (cf. McCormick, 1959, p. 103). (Meyer & Schvanefeldt 1971: 232)

This would explain why computer users, understandably, feel that their machine cannot think or is illogical. The fact is, that the logic of a RAM (Random Access Memory) has little in common with the network that binds information together in the human memory.

Finally, Meyer and Schvanefeldt refine Quillian's concept of linking words as nested strings[58]. They note:

> We previously have argued that processing normally begins with a decision about the top string and then proceeds to a decision about the bottom one. Let us now assume that memory is organized by familiarity as well as by meaning, with frequently examined locations in one "sector" and infrequently examined locations in another sector.
> (Meyer & Schvanefeldt 1971: 232)

This means the *familiarity,* and hence the *priming* of a term, is mapped for its likely use and environment in the language-users' mind.

Meyer and Schvanefeldt claim, in their 1976 paper, unambiguously sub-titled *People's rapid reactions to words help reveal how stored semantic information is retrieved,* that their set-up differs from most other experiments in the field, in that they do not seek to measure speakers' mistakes but the reaction time people take making lexical

[58] See Quillian (1969: 472): *[this] does not output as a parsing a tree structure, but rather a set of nested strings. However, in building these strings it succeeds in "undoing" a number of syntactic transformations, replacing deleted elements and rearranging others.*

choices. Interestingly, the rate of error is remarkably low, indicating how sure-footed language users are in their native language:

> But the reaction times depended significantly on the set relations between the categories. When the meanings of the category names were closely related to each other, reaction times tended to be shorter.
>
> (…)
>
> People were about 55 ± 7 milliseconds faster on the average at recognizing a word like BUTTER if it followed the related word BREAD than if it followed the unrelated word NURSE (20). (Meyer & Schvanefeldt 1976: 30)

The difference in milliseconds, becomes significantly large when compared at this level. Meyer and Schvanefeldt do not use the term lexical priming, but it is clear to readers familiar with concordances that BREAD and BUTTER are likely to be in each others' company, while BREAD and NURSE are not. This, then, would experimentally confirm the foundations of the LP theory. Indeed, the notion of lexical priming, in all but name, is supported by another set of experiments described by the authors. Once words are made harder to decipher, the *semantic memory* assists recognition:

> Degrading the legibility with [a] pattern of dots increased reaction times by more than 100 Milliseconds. The harmful effect of degradation was significantly less, however, for related words than for unrelated words, suggesting that semantic relatedness helped to overcome the visual distortions produced by the degradation.
>
> (Meyer & Schvanefeldt 1976: 30)

Hoey (2005) notes that lexical priming does not simply mean connecting lexically / semantically related words. In fact, some primes (e.g. VERY) have little lexical content. That these still play an important part of the semantic memory is pointed out by Quillian (1969). Meyer and Schvanefeldt highlight that it is not necessarily the "meaning" of a word that makes it act as a prime and, consequently, ask for further investigation[59]:

> It is not true, however, that close relations of meaning always facilitate mental processing of words. Some processes are actually inhibited when they must deal with two words that have related meanings. (…) The apparent inhibition raises more questions about what semantic information is stored in human memory and how the information is used. (Meyer & Schvanefeldt 1976: 31)

This could be seen as an explanation why synonyms, though clearly related, are not fully interchangeable in all contexts. As spoken language production is not pre-planned and aims to be fluent with as little hesitation as possible, the words (chunks of words) that have least inhibition will tend to be the preferred choice. Hoey (2005) states that *words either prefer or avoid the company of others*. The *apparent inhibition* is assumed to be because these words, even if semantically related, have not been primed for the speaker to occur together.[60]

[59] I discuss the link of "meaning" and "priming" in section 3.4.1

[60] While we can speak of a *tall order* and a *tall boy*, there is a *high tower* (not *tall tower or *high boy). Talking of a *high order* (*highest order* is more common usage) actually means something very different compared to *tall order*.

### 3.3.3 Semantic Priming of the Lexical Memory

J.H. Neely's two papers (1976; 1977) are cited in Hoey (2005) and build on Meyer and Schvanefeldt's work. His *Semantic Priming of Lexical Memory*, for the first time, connects the words *priming* and *lexical*. In his 1976 experiment, volunteers see a Related (R), Unrelated (U) or Neutral (Nx)[61] semantic term as a prime before a target word. Exposure to these primes varies between extremely short (360 msec), medium (600 msec) and very long times (2,000 msec). Whatever the exposure, the R prime provoked a shorter reaction time. During short exposure, the difference between R and U is 40 msec (Nx lies in between). However the gap becomes marked for 600msec and longer exposure times. While a neutral prime runs in close parallel to the unrelated prime, the related prime has a response time of between 60 and 80 msec difference from the unrelated prime. (i.e. 540 msec instead of 600 msec). As with the Meyer and Schvanefeldt experiment, Neely's informants' error rate was remarkably low.

He relates in his discussion that –

> Activation spreads from the logogen[62] for the priming word to the logogens for semantically related words, and (2) the subject uses the priming word to direct his (…) attention for words that are semantically related to the priming word. (Neely 1976: 652)

[61] Described by Neely (p. 649: 1976) as follows: *a semantically neutral warning prime consisting of a series of Xs*.

[62] According to the **Logogen** model, the word frequency effect is explained by logogens having different thresholds, such that "logogens corresponding to words of high frequency in the language have lower thresholds" (Morton, 1969). Hence, high frequency (i.e. common) words require less perceptual information to raise their activation to threshold, hence are recognised more quickly than

Neely appears to say that the *threshold of perception* (see footnote below) of what is here referred to as a *logogen* is directed by the level of semantic relatedness. His final conclusions point in the direction of lexical priming:

> (…) In comparison to a noninformative and semantically neutral warning-signal prime, a word prime (1) facilitates lexical decisions about a subsequently presented semantically related word, (2) inhibits lexical decisions about a subsequently unrelated word, and (3) facilitates decisions about a subsequently presented nonword.
> (Neely 1976: 654)

With this, Neely underlined the importance of Meyer and Schvanefeldt's findings, while at the same time rebutting a theory of Posner & Snyder, who had postulated that priming was expectancy based and under the subject's control.

At this stage, experimental linguists had opened a gate to connect lexical decisions with concepts formed in the mind. That grammatical choices and lexical choices are entwined was under serious discussion. Zimmermann (1972), discussing automated text lemmatisation, comments:

> Die Konzeption eines Lexikons schließt die Konzeption einer Grammatik weitgehend ein: Lexikon und Regelsystem bilden eine Einheit. (…) eine Satzanalyse (oder weiter gefasst: eine Kontextanalyse)schafft erst die Voraussetzung dafür, Texte zu lemmatisieren. Die an der (Wort- oder Satz-) Oberfläche mehrdeutigen (Teil-)

---

low frequency words. Definition taken from Milton, N: *Word Recognition* @ http://www.epistemics.co.uk/staff/nmilton/papers/word-recognition.htm (accessed 07/09).

> Strukturen sind mittels der Informationen aus dem Kontext zu vereindeutigen und in den Rahmen der Strukturierung des Textes (oder bescheidener: der Sätze) entsprechend einzugliedern.[63] (Zimmermann1972: 3)

Nevertheless, despite the occasional paper linking into this type of research in the eighties – notably by Neely (1989) himself, the citation index of the papers published shows that the notion of priming, in the context of lexical memory and sentence grammar (semantic or syntactical) has not become prominent in linguistic discussion until quite recently[64].

## 3.4 Priming and Syntax

There seems to have been little significant research on the matter of semantic memory and priming for the next 20 years. Few of the subsequently published papers have been much cited (according to the citation indices) and most seem to simply confirm the results and conclusions of earlier researchers. Even the later Neely (1989) paper mainly re-iterates the findings of his 1976/77 papers.

While the comprehension of non-linear (i.e. *complex*) concepts in general was under discussion in the 1980s, in the 1990s and particularly

---

[63] The conception of a dictionary comprises almost totally the conception of a grammar: lexicon and rule system are one. (…) It is syntax analysis (or, in a wider sense, context analysis) that creates the basis for lemmatising texts. The structures that are ambiguous on the surface (of words or sentences) are to be disambiguated with the information gathered from its context and to be integrated into the framework of the text (or the sentences). *My translation. M P-S*

[64] To find this, I have made use of the University of Liverpool's *Summon, Scopus* and *Discover* systems. A further search was made on Google Scholar (last accessed 09/10): http://scholar.google.co.uk/scholar?start=10&q=priming+lexical+OR+memory+%22sentence+grammar%22&hl=en&as_sdt=2001&as_ylo=1962&as_yhi=2010&as_subj=bio+med+soc

in the 2000s, however, the notion of priming has become of renewed interest to psycholinguists. The main strands are two now: the priming observed when reading and the priming observed in the oral production / perception of language. The interest most share is the topic of syntactic and semantic priming.

### 3.4.1 The importance of compounds in research

The foundations for research on dependent clusters can be found in Gregory Murphy's *Comprehending Complex Concepts* (1988). Here, Murphy defines the *complex concept* as lying between the *simple* – that "can be represented as a single lexical item", and the "lexicalized (i.e., idiomatic) expression". In his paper, he quotes the example of "corporate lawyer" which is a fixed, complex adjective-noun expression. Murphy notes that the noun-noun expression "*corporation lawyer*" is not available for use and expressions like "corporate stationery" mean something very different from the term "corporate". Murphy hints at the fact that the listener would have to know which of the specific meanings a non-predicating term like "corporate" has and his paper can be seen as another stepping-stone towards acceptance of fixed collocations as a psycholinguistic notion.

Ratcliff and McKoon (1988) go much further in their research. The hypothesis they outline is that of *compound cue* priming. In terms of

retrieval from memory, they advance the theory that it is not concept trees (bird – animal – flight) but words that go together that make it possible to associate:

> The theory assumes that the prime and target form a compound cue and that this compound interacts with memory to produce a value of resonance, goodness of match, or familiarity that is determined by associations in long-term memory between the prime and target. If the prime and target are directly associated in memory, then the familiarity value will be larger than if they are not associated.
>
> (Ratcliff and McKoon 1988: 405)

This would cover a range of options. The "goodness of match" would determine in what sense "corporate" (see above) would be used if it compounds with "lawyer" rather than with "stationery". Likewise, the sense of "familiarity" would find few associations for "corporation lawyer" – "corporate lawyer" being the familiar combination. In fact, *compound cue* priming highlights that the human mind very seldom retains a single lexical item by itself in its memory. It usually is associated with another term. This notion of association goes beyond the confines of simple collocation. Referring to their earlier (1981) work, Ratcliff and McKoon (1988: 389) point out that "they have shown that priming can be obtained between concepts that are much more than four words apart." This raises issues, though, about collocation since it appears to contradict Sinclair's (1991) claim that there are no valid collocations beyond the five-word mark either side.

De Mornay Davies (1998), in his work on brain-damaged patients[65] finds that they lack the knowledge (in other words, the operating software) to use their semantic memory.

They do tend to *hyperprime*[66], seemingly retaining most of the semantic information associated with target words presented:

> It has often been reported for these patients that, whereas semantic representations, as assessed by off-line tasks, are degraded or inaccessible, their performance on semantic priming tasks suggests that much of the semantic information associated with these concepts is retained" (de Mornay Davies 1998: 390)

The importance of his work in this context is that he is able to demonstrate the long-term memory function of semantic association[67] and its automatic retrieval:

> Automatic semantic priming assumes that, on presentation of a word, the information about that word is retrieved as a result of lexical access, rather than being retrieved explicitly as a result of subjects' responses to task demands.
> (de Mornay Davies 1998: 391)

The concept of *lexical access* appears to be very close to *lexical priming*. De Mornay Davies is more explicit when he states:

---

65 described by DMD (1998: 390) as "*Patients with semantic memory breakdown"*.

66 Patients with semantic memory breakdown often show increased priming on semantic priming tasks compared to normals (``hyper-priming' ')

67 See cpt. 3.2.1.4

> Even if two words are not ``semantically related' ' in the strictest sense (i.e. they do not come from the same superordinate category), their frequent association produces a relationship at the ``meaning'' level. (de Mornay Davies 1998: 394)

This foreshadows Hoey, saying that each term is primed to mean something as a result of frequent association.

De Mornay Davies finds that there is still a strong drive by researchers to try and find a meaning-driven correlation of words. However, this would neither explain idiomatic use, nor his findings with brain-damaged patients. There is, however, a lexical and semantic automatism:

> .. activation in the lexical network could be controlled by co-occurrence frequency, such that words that often co-occur in speech or text (`collocates' ) would be more strongly linked in a phonological or orthographic lexical network. Lexical co-occurrence, therefore, has no connection with meaning-level representations, and many researchers argue that associative priming results from lexical-level co-occurrence.
>
> (de Mornay Davies 1998: 402)

Regrettably, he does not specify who these "many researchers" are; the bases of his claims are the findings of his own experiments. Being more specific than Ratcliff and McKoon, he anticipates Hoey's later claim that it is the property of each word to be primed to either prefer or avoid the company of other specific words, noting that this is the case because the mind co-associates these words, rather than because it links each individual word to concepts or meanings. This approach to *meaning* is also noted by the pragmaticist Siobhan Chapman:

> Many [linguists] would argue that it does not even make sense to try to discuss 'meaning' as a feature independent of context. The meaning of a word is entirely defined by how speakers use it in context; (…) these linguists reject the distinction between semantics and pragmatics as an unnecessary imposition on human communication. (Chapman 2006: 116)

By 2000, researchers had gathered enough evidence to conclude that priming is an automatic process, a single process not split into stages. Hernandez *et al.* (2001) confirm that -

> … No evidence was found for a stage in which lexical priming is present but sentential priming is absent – a finding that is difficult to reconcile with two-stage models of lexical versus sentential priming. We conclude that sentential context operates very early in the process of word recognition, and that it can interact with lexical priming at the earliest time window.
> (Hernandez et al. 2001: 191)

There has been, too, a body of work indicating how compounds or collocates play an important role for the human mind in the association of lexical items. On the basis of this, a host of new experiments and research has been undertaken in the past since the late 1990s.

### 3.4.2 Is priming verb or noun-driven?

In recent years, experimental psychologists and psycholinguists have sought to schematise what types of words are more likely as effective

primes for what follows. It has been as a question what is likely to act as triggers to prime what follows. This shall be considered in this section. In order to look at the relevant work here, it is helpful to introduce the notion of colligation, a term which owes its origin to Firth (1957). Hoey's (1996) definition of colligation, which is the one used in this thesis is inspired[68] by Michael Halliday's use of the term. Sinclair (1991), Hunston (2001) and Partington (1998) have all adhered to the concept in a very similar sense.

It is the category and function with which a word occurs that constitute colligation:

> 1. the grammatical company a word or word sequence keeps (or avoids keeping) (…);
> 2. the grammatical functions preferred or avoided by the group in which the word or word sequence participates;
> 3. the place in a sequence that a word or a word sequence prefers (or avoids).
>
> (Hoey 2003b: 389 also in Hoey 2005: 43)

This concept is particularly relevant to the issue whether certain grammatical functions are more likely to be associated with effective primes. A number of psycholinguists have made a case for either verbs or nouns being more important primes for words to follow. While *collocation* simply looks at how words co-occur, colligation looks at "the grammatical function preferred or avoided by the group in which the word or word sequence participate" (see 2. above). Consequently, if either verbs or

[68] personal communication

nouns act as key prime, it may be their colligational rather than their collocational role that is of importance.

Hoey also introduces another term to describe priming of semantic functions. This is *Semantic Association* (Hoey: 2005) This term is inextricably linked with the concept of colligation as it defines how we associate a word in its grammatical context. Concepts similar to semantic association are described in the research experiments undertaken by psycholinguists. In this context, the terms *semantic preference* or *syntactic preference* are being used. For example, Novick *et al* (2003) say that

> It is also worth noting that properties of the primes used in this experiment may also speak to the relative contribution of verb-specific syntactic and semantic preferences to parsing decisions. They also suggest that thematic role and syntactic preferences are activated during word recognition. (Novick *et al.* 2003: 71)

and note that both influence combinatory processing. Novick *et al* (2003) and Salamoura & Williams (2006) both cite Trueswell and Kim (1998) as having found that during sentence reading in L1 the **syntactic preferences** activated by a briefly displayed single verb were enough to bias the readers' resolution of temporary syntactic ambiguities.

Salamoura & Williams (2006) introduce us to yet another term:

> the processing of a (…) Dutch verb prime should be sufficient to bias speakers' **structural preferences** in a subsequent English target sentence according to the feature-based account of cross language syntactic priming.
>
> (Salamoura & Williams 2006: 301)

Yet, apart from work with Dutch speakers (see also de Goede 2006), the issue of verb priming appears to be of little relevance to lexical priming in English.

#### 3.4.2.1 Noun-driven priming

Gagné (2000) makes a case for head-noun driven priming in preference to modifier-driven priming:

> (…) although the modifier is more influential in the selection of a relation used to interpret the combination, the head noun is more influential in integrating the combination with existing knowledge. As a result, the head noun might receive more activation than the modifier. (…) A second possibility for why less priming was observed after a modifier prime than after a head noun prime concerns the interplay between the relation activated by the modifier prime and the relations activated by the modifier's relational distribution. When there is a discrepancy between the relation used in the modifier prime and the dominant relations activated on the basis of the modifier's relational distribution, the interpretation of the target combination will be slowed. When there is no discrepancy, the interpretation of the target combination will be facilitated.
>
> (Gagné 2000: 251)

While the first sentence describes a commonsensical process, it does not explain how these two words are set in relation in the first place. However, Gagné claims that the head noun gains its prominence through being meaningful: the mind can connect it with world knowledge. This approach stands in contradiction to de Mornay Davies (cf. chapter 3.4.1).

At all events Gagné's experimental results reaffirm the view that it is lexical co-occurrence that facilitates the "interpretation of the target combination" in that the language-user has been primed to accept a certain modifier-head noun combination, while rejecting, or avoiding, another.

Cleland and Pickering (2003) describe three experiments to confirm the importance of nouns in driving priming. They look solely at short-term memory–responses by second parties in dialogue and their re-use of head nouns:

> Experiment 1 found that repetition of the head noun between prime and target increased the tendency to repeat syntactic structure. Experiment 2 found an increased tendency toward syntactic repetition when the head nouns in prime and target were semantically related versus when they were unrelated (but less of a tendency than when they were the same). Experiment 3, however, found no tendency toward an increased effect when the head nouns were phonologically related versus when they were unrelated.
>
> (Cleland and Pickering 2003: 225)

The results of Experiment (1) might be explained in terms of the way respondents try to home in on the genre / tone of the previous speakers to fulfil the co-operation principle. If this short-term priming is used on repeated basis, these words might then move to long-term memory. The enhanced effect described in Experiment (2) is what semantic priming seems to be about. As for the results from Experiment (3), listeners and speakers have no reason to connect ship and sheep (their examples), so the dispreference is to be expected in the light of lexical priming theory.

Homonyms might have resulted in more interesting results. Most importantly, however, these experiments reinforce Hoey's theory. They indicate how, in the narrow confines of a controlled experiment, users are primed to employ specific words in both their lexical use (collocations) and semantic positions (colligations).

### 3.4.3 The value of context

In a continuum from collocation to colligation is the propensity already discussed by Quillian in 1962 – for word meaning to be disambiguated by the context it is found in. A considerable number of words have little concrete meaning by themselves, either because of the level of de-lexicalisation they have undergone or because of their role as function words. Also, as has been suggested above, even the role of synonyms is suspect – they are hardly ever fully interchangeable when presented in context.

Novick *et al.* (2003) provide evidence that word meaning is disambiguated by the context in which it is found:

> In this regard, it is interesting to note that priming effects appeared to be restricted to the argument preferences of the primes, and not to other aspects of the prime verb meaning, such as the verb's "core meaning." (Novick *et al.* 2003: 71)

This would appear to undermine any theories that lexical words (in this case, *verbs*), have a *core meaning* that remains stable whatever the

context. On the contrary, it appears that Novick *et al.* are suggesting that the context selects the meaning of the word.

Novick *et al.* (Ibid.) set up an experiment to investigate the way participants disambiguate verb meanings in sentences. Participants had to decide, from the wider context, what the most likely meaning conveyed by an ambiguous term was. This linked in with "the probability of each option, given a word and its local context."

Novick *et al.*'s 2003 paper on spoken word recognition reads like a blueprint for the theory that Michael Hoey started to outline during conferences from the same year on:

> Several conclusions about the nature of sentence comprehension arise from these results:
>
> 1. Lexical knowledge encodes detailed information about the syntactic possibilities for words, directly influencing the manner in which words are combined to form sentence-level representations. This is true of verbs and also of other word classes, such as nouns.
>
> 2. Those lexical-combinatory representations are encoded in a distributed manner and shared between words in a way that crosses grammatical class boundaries.
>
> 3. The lexical representations that guide sentence processing include combinatory information of a sort that may go beyond classical syntactic notions. This information may include event-structural information, including information about which specific classes of arguments a particular word tends to associate with.
>
> 4. The findings in general align well with constraint-based lexicalist theories of parsing. Word recognition appears to play an important role in the grammatical analyses of sentences. (Novick *et al.* 2003: 72)

I have quoted the conclusions in full to highlight the parallel conclusions drawn between Novick *et al.* and Hoey. Though lexical priming is not mentioned as such, *constraint-based lexicalist theories of parsing* would certainly include it. Clearly, points one and two mirror the concept of colligation, while points two and three also encompass semantic association. Point four, to conclude, highlights that grammar is lexically-driven and lexical occurrence and position determine the grammatical structure, not vice versa.

### 3.4.4 Priming in spoken usage – mirroring preceding word use

As section 3.4.3 has shown, the same notions of priming hold true for both the listener and the reader. As my thesis works with spoken corpora, it is important to highlight the work undertaken in experimental linguistics over the past decade. Though a majority of this research is found in applied linguistics, mainly in connection with comparison of the use of two languages, the results can still be seen as valid and important in the wider context of this thesis.

An experiment confirming the importance of collocates in producing primings in spoken communication is described by de Mornay Davies (1998), referring to Williams (1996), who

> compared the effects of four types of prime - target pairs: semantically similar, category coordinates, collocates (lexical co-occurrences) and associates (from word association norms). Only collocates produced significant priming in a pronunciation task when both prime and target were intact. (de Mornay Davies 1998: 395)

Initial work on spoken priming focussed on the short-term memory effect. That is to say, this research looked at how far a listener would reuse words, phrases or constructions when it was his or her turn to speak. Melinger & Dobel, introducing their work on German and Dutch, state:

> research on sentence production has revealed a tendency for speakers to reuse structures they have previously encountered. **This pattern of speaker behavior** (sic) **is known as syntactic or structural priming.**
> (my highlighting) (Melinger & Dobel 2005: B11)

Melinger & Dobel used for their experiment verb-prime constructions that are rare in spontaneous (unplanned) speech. The constructs that speakers produced after having listened to their prior speakers mirrored the (less common) constructs used. This could be explained by the co-operation principle, where speakers and listeners try to take account of each other in communication. This process would happen without the speakers being necessarily conscious of it. This is as an important finding in the context of my thesis. If the language characteristics occurring in a small (geographically limited) speech community are reinforced by daily use, a new set of primings could be assumed to have

been coined. The implications of this were highlighted by Pavel Trofimovich as early as 1992:

> In contrast to the facilitative effects of a repeated phonological context or of a semantically related word which rarely last more than a second, auditory word-priming effects are long lasting. For example, reliable processing benefits for repeated spoken words are maintained over delays of 8 s (Cole, Coltheart & Allard, 1974), minutes (Church & Schacter, 1994), days, and even weeks (Goldinger, 1996). These findings suggest that auditory word-priming effects have a long-term memory component. (Trofimovich 1992: 481)

Trofimovich looks at word priming in (spoken) context, comparing learners both in L1 and L2 contexts. Like Darnton (2001) he finds there is intrinsic value in repeated exposure and use of words in their contexts for the learners. He quotes Church and Fisher (1998) who say that

> (we have) recently identified auditory word priming as a likely mechanism supporting spoken-word processing and learning. (…) **because auditory word priming does not require access to word meaning**, it may reflect the process whereby listeners build and use presemantic *auditory* representations. (my highlights) (Trofimovich 1992: 482)

This is a departure from the concept of priming in context. The lack of knowledge of the *word meaning* presupposes that the hearer simply gains priming by hearing the same word in similar constructions and surroundings on a repeat-basis. Trofimovich's experiments show that priming, indeed, can be achieved this way:

> … results of this experiment revealed that, in both English and Spanish, the participants were faster at initiating word production in response to a repeated than an unrepeated word. That is an auditory word-priming effect (a temporal benefit in the processing of repeated vs. unrepeated words) was obtained in both languages.
>
> (Trofimovich 1992: 489)

However, it is under discussion whether this effect described above is lexically driven.

> Some psycholinguists have argued that the persistence effects that have been called syntactic or structural priming are in reality lexically driven. One of the most important arguments has been the observation that syntactic priming is increased dramatically when the lexical items in the prime and target are repeated.
>
> (Desmet and Declercq 2006: 621)

This would appear to back Trofimovich's findings. The *lexical boost effect* (Bock) appears time and again. Priming clearly is reinforced by repeated use.

Influenced by Melinger and Dobel, Salamoura and Williams (Cf. chapter 3.4.4), looked at translations by Dutch L1 speakers from English (their L2). Their research indicates that (fluent) L2 speakers are unlike beginners (and simple translation software) in that they do not seek translation word-by-word but by trying to locate an exact equivalent in the target language. This equivalent can be primed by a single lexical item, while the priming activates at the same time the use of the respective construct.

## 3.5 Priming and the Corpus

Up to this point, all the evidence for the existence of *Lexical Priming* and its workings have been based on experimental evidence by researchers into artificial intelligence (AI); cognitive linguists and psycholinguists, only very few people have tried to find proof for this notion in the real-occurring texts produced by writers and speakers – the corpus. Leaving the work of John Sinclair and Michael Hoey aside, let us turn to an account of the latest corpus-based psychoanalytical work by both European and US-American researchers.

Looking at the work by S.T. Gries and Nick Ellis *et al*, it becomes apparent that the two strands of empirical research – experiment-based and corpus based – are finally brought together. Ellis *et al* quote Meyer and Schvaneveldt (1971), while Gries highlights the fact that

> … although it has sometimes been argued that only experimental data can contribute to studies of priming, the analysis shows that ... the corpus based results for datives are very similar to the experimental ones." (Gries 2005: 365)

Gries introduces his study with a brief overview, stating that -

> … syntactic priming: (…)Levelt and Kelter (1982) and Branigan *et al.* (1999) report that priming (in spoken and written production respectively) is fairly short-lived.
> (Gries 2005: 368)

That priming is a short-lived and short-term memory issue, however, is only discussed in earlier *syntactic priming* discussions. Later research

has accommodated the notion that there is also the long-term, more fixed priming. Still, Gries notes that his colleagues appear to be locked into their traditional methods, as he does through quoting Branigan:

> Corpora have proved useful as a means of hypothesis generation, but unequivocal demonstrations of syntactic priming effects can only come from controlled experiments (Branigan *et al.*, 1995: 492; cf. also Pickering & Branigan, 1999: 136).
> (Quoted in Gries 2005: 369)

It appears from this that neither Branigan nor Pickering did any work with corpora at all but all know about some investigations based on corpus research. Branigan and Pickering seem unwilling to consider to look beyond the scope of "controlled experiments" and appear to be set against the use of corpora-based research argument without giving any further reasons why. This, however, this has not stopped Gries (as well as Ellis 2006a & 2006b) from conducting corpus-based experiments. While using data from the ICE-GB corpus, Gries analyzes two different pairs of syntactic patterns, the so-called "dative alternation" and "particle placement of transitive phrasal verbs": In order to investigate syntactic priming corpus-linguistically, Gries identified all ditransitive constructions and all prepositional datives with **to** and **for** in the British component of the International Corpus of English (ICE-GB) (cf. Gries 2005: 370). Gries himself seems to be taken aback by how well the data from his corpus match experimental results:

> In the present data, the ratios of the primed structure vs. the non-primed structure are 1.5 and 1.9 for prepositional datives and ditransitives respectively. By comparison, in her classic study, Bock (1986: 364) reports percentages instead of raw frequencies where the corresponding ratios of the percentages are 1.5 and 2.1 for prepositional datives and ditransitives respectively; the differences between her ratios and mine are obviously negligible. This also indicates that ditransitives prime more strongly than prepositional datives.
>
> (…)
>
> In sum, not only has the corpus-based analysis of syntactic priming revealed significant priming effects for ditransitives and prepositional datives, the results are also strikingly similar to those of previous experimental studies in terms of strength of effects, the influence of morphological characteristics of the verbs, construction-specificity, directionality and distance effects (i.e. the time course of priming). (Gries 2005: 373f.)

Gries' results are remarkable. All hypotheses were matched, with a very small reported rate of error. It is remarkable how well theory and results match. Throughout a great number of experiments discussed, Gries is able to find significant priming effects.

Gries (2005) echoes Hoey's (2005) definition of colligation (see above). The results presented by Gries make a good case for corpus linguistics working in tune with psycholinguistic methods:

> While I do not rule out discourse-motivated factors of priming at all, it is hard to explain all the similarities between the different kinds of results and still simply uphold the claim that all this is epiphenomenal. Without doubt, further experimental evidence is necessary, but it seems as if the utility of corpus-based, explorative results should not be underestimated prematurely.

> (…) **the fact that lexical activation decays too fast makes it unlikely that the long duration of priming effects observed here and in other (experimental studies) is just a lexical memory effect.** (my highlights) (Gries 2005: 387)

The latter part of the quote appears to move the discussion away from where Gries started his paper: priming effects go beyond syntactic priming found in exchanges. It works on a far deeper and more profound level.

In experiments that, similarly to Gries', compared volunteers' reaction times with (BNC) corpus evidence, Ellis *et al* (2006a; 2006b) came to similar conclusions. Ellis *et al* (2006b) seems to mirror and expand the experiments undertaken in 3.4.4 – where native speakers are compared with non-native ESL speakers[69]. They confirm Gries' results. Having Sinclair's (1991) idiom principle in mind, however, Ellis *et al.* outline that primings work in different ways for the two groups:

> Fluent Native speakers much more affected by MI (**M**utual **I**nformation)
>
> Non Native ESL speakers more affected by Frequency (Ellis *et al:* 2006b[70])

This is based on the following definitions:

- Frequency - need to have come upon the string before (strong effects of frequency in vocabulary acquisition and processing).

[69] The research is based on the most frequent phrases found in spoken and written academic texts in the BNC.

[70] Ellis *et al.* : 2006a and Ellis *et al.* : 2006 are PowerPoint presentations, hence no page numbers are given.

- MI - the bindings of words within a formula which make the formula distinctive and functional as a whole.

(Ellis *et al:* 2006b)

There is logic to this. All listeners / readers can be sure that "high frequency patterns are processed more fluently" (Ellis *et al:* 2006b). For all that, a learner of a new language will merely recognise strings he or she has been exposed to frequently before. A native speaker, however, is not just more likely to have heard /read the formula before: they will also be more open to a more loose form of repetition – as long as the *bindings of the words* remain consistent.

In other work, Ellis *et al.* (2006a) look at collocations and semantic prosody[71]. Ellis describes the set up of their tests as straightforward:

> We investigated the frequency and strength of these collocations in the BNC then looked for processing effects using the lexical decision paradigm.
>
> (Ellis *et al.*: 2006a)

This means that the researchers extracted frequently occurring collocates (clusters) from the BNC (for example: lose weight – frequent; receive virginity – infrequent (*sic*)) and then measured the reaction time (RT) it took to make a lexical decision. As a result, the team found that "Language processing (as indexed by this lexical decision task) is intimately sensitive to patterns of collocations in usage." (Ellis *et al.*: 2006a). The graphs of the corpus-occurrence patterns and the reaction

[71] Semantic prosody, based on definitions of Louw and Sinclair, is here described as the consistent aura of meaning with which a form is imbued by its collocates & the general tendency of certain words to co-occur with either negative or positive expressions.

times run in close correlation to each other for all the above-mentioned patterns. This is not that clear-cut, however, when it comes to semantic prosody. This may be due to the fact that semantic prosody is a vulnerable concept, as it is not easily replicable[72], and has been disputed. Still, the results of Ellis *et al.* (2006a) can be summarized in the schema given:

| | Usage Corpora | Lexical access | Semantic access | Selection for production |
|---|---|---|---|---|
| Collocation | yes | yes | yes | Not studied |
| Semantic Prosody | yes | no | yes | Not studied |

Table 1: results of Ellis *et al.* (2006a) summarized

The last column, *selection for production* is probably left open for further research. In a way, the *selection for production* is already made – by the choice of corpora.

The researchers conclude that –

- Written language processing is intimately tuned to frequencies of actual usage
- We process frequent collocations faster than infrequent ones
- we do *not* see ready evidence of *semantic* generalization here
- It appears the fluent processing associated with spread of activation in 'semantic priming effects' are due to memory for particular word associations.
- There is little by way of semantic generalization at this level of processing at least.

(Ellis *et al.* : 2006a)

[72] This fact has been highlighted by John Sinclair.

The first point is in total agreement with what Gries found in his experiments and what Hoey (2005) claims. The second and the last point also confirm what de Mornay Davies and others have claimed – that priming is not down to something that is based on *semantic generalisations* but more due to automatic decisions made because of *word associations in the memory*.

All in all, this should determine that corpus studies are as valid for psychological sciences as carefully structured experiments are. Likewise, the experiments undertaken to date confirm conclusion drawn by corpus linguists about the nature of language comprehension and language production.

## 3.6 Sociolinguistics, Psycholinguistics, Priming - and how they relate to each other

One feature that links sociolinguistics, psycholinguistics and corpus linguistics together is that their findings are based on real occurring (written or spoken) text[73]. All three appear to have started in around the 1960s-1970s, too. The major difference between the 1970s and now is that a) far more data are now available and b) a more objective, more powerful means of investigation is at easy disposal of researchers – the computer. This opens up whole new avenues of research.

[73] A case can be made that this chapter also needs to make reference to (neuro-) cognitive linguistics. Though work by Wallace Chaffe (1982) and Sidney Lamb (2000) has been consulted by me, I found it difficult to integrate this into framework of this thesis. Conversely, however, Sydney Lamb had never heard of *Lexical Priming* (personal communication, Cardiff LINC September 2010).

### 3.6.1 Pattern and Corpus Linguistics

Like the psycholinguists (cf. chapter 3.1 and 3.2), *Labov* and Wolfram found in their data the same evidence that Hoey (2005) would later use to develop his theory of Lexical Priming:

> Every lexical choice starts off a series of options and predilections that result in an amazing fluency in any situation in which the speaker has been primed to perform. (Hoey 2005: 163)

To trace the usage of words, their primings, Hoey uses corpus linguistics. Biber *et al.* (1998) describe the uses of corpus linguistics to investigate register variation, language acquisition & development as well as stylistic investigations. Institutional talk, in particular politicians' talk is widely investigated, most notably by Partington (2003). Corpora are now widely used in *Discourse Analysis* (Baker: 2006).[74] [75] From there, it is only a small step to the concept of *Colligation* as developed by Sinclair and Hoey, where the language structure is driven by the lexis.

[74] To date, corpus linguists have made a notable impact in many disciplines of linguistics. Since the publication of the first Collins Cobuild Dictionary in 1987, a new dictionary needs to take into consideration recourse to a corpus. There are a variety of researchers in mainland Europe doing contrastive studies based on corpora (*i.e.* de Groot 1989; Carreiras & Perea 2002; Salamoura & Williams 2004: Desmet & Declercq, 2005; Melinger & Dobel, 2005 & 2006; Trofimovich, 2005; de Beaugrande, 2007).

[75] Patrick Hanks, lexicographer and corpus linguist, has pointed out that lexicography is well aware that issues like peer-pressure, pressure through ridicule etc. bring about small differences in language use (personal communication). Though I am not aware of any investigation of such changes, any work in this direction would most like make use of corpora.

> Every word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word. If one of the effects of the initial priming is that regular word sequences are constructed, these are in turn primed. (Hoey 2005: 9)

A single use may not even register, repeat usage, however, primes the listener/speaker to appropriate the term or term sequences for their own use. In this context, work undertaken since the late 1960s has an intrinsic value and importance in the context of corpus-based analysis, and, furthermore, to the notion of lexical priming. Wolfram describes an important area of distinction between language varieties – frequency of use:

> But studies of sociolects which were done during the 1960s - particularly those which followed the Labovian quantitative orientation, indicated that sociolects were often not differentiated by discrete sets of features alone, but also by variations in the frequency with which certain features or rules occurred. (Wolfram 1978: 2)

Wolfram highlights here that the "variations in the frequency of sets of features" rather than a complete collection of variations are the ones that distinguish one variation from another. While Labov, Trudgill and others initially focussed on phonological differences, Wolfram casts the net wider – and opens the door to expand the tools and approaches to dialectology amongst other things –

> Further, it is necessary to identify relevant linguistic environments (phonological, grammatical, and semantic) which may affect the variation of items. (Wolfram 1978: 8)

This is a crucial point in this research. Wolfram makes clear that an expansion of dialectology and sociolinguistics beyond its traditional brief and stretching out to the *phonological, grammatical, and semantic* is possible. In short, all people who are native speakers have access to about the same sets of features. The point of distinction appears to be, however, how these sets of features vary in their frequency.

On the surface of it, corpus linguists, psycholinguists and socio-linguists are all alike that all of them look at real (natural occurring) data. They also have in common the focus on frequency of occurrence. The difference is usually the different sections of similar material that all three groups focus upon. What has been expressed by Biber about assembling data for a corpus would be seen as equally relevant for the other two groups of researchers:

> Finding patterns of use and analysing contextual factors can present difficult methodological challenges. Because we are looking for typical patterns, analysis cannot rely on intuitions or anecdotal evidence. (…) Furthermore, we need to analyse a large amount of language from many speakers, to make sure that we are not basing conclusions on a few speakers' idiosyncrasies. (Biber *et al.* 1998: 3)

We have already shown the link between psychological research and Hoey's *Lexical Priming*. As early as 1978, Wolfram (a socio-linguist) describes the link between linguistics and psychology:

> Linguistic theory, if studied seriously, has as its goal accounting for exactly the capabilities people have in using their language-no more and no less. Linguistic theory, then, can be viewed as a special kind of study in **psychology.** Taken seriously, every capability built into a linguistic theory constitutes a claim that the same capability is built into the language control parts of the human brain and speech mechanism.
>
> (My highlights) (Wolfram 1978: 12)

That "linguistic theory can be seen as a special kind of study in psychology" is expanded by Prucha (1972), who looks at communication and context. The psychological processes in acquiring language should be seen in the context of its social and cultural background, as Prucha points out:

> The theory of language behaviour and language acquisition cannot be established without the study of the communicating man, and the study of the communicating man cannot be isolated from the communication context in a broad sense, i.e. also involving the social and cultural background. (Prucha 1972: 9)

Prucha brings together the different strands under discussion. Psycholinguistics is, according to him, concerned with the individual's own language processes. Sociolinguistics, however, looks at the context in which each utterance is made. Corpus linguistics provides material to study *the communicating man* as an adult speaker. Prucha (1972) pointed out that-

> … little is known of the communicative competence of adult speakers. Undoubtedly, however, the concept of communicative competence is very useful, as it unites the psycholinguistic and sociolinguistic aspects. (Prucha 1972: 10)

This was over 35 years ago. Though the term *communicative competence* was coined by Hymes (71), Hudson's (1980 [1996]) discussion makes clear that this concept goes way beyond the confines of grammar and into the area of cultural conventions:

> Some parts of communicative competence may be due to universal pragmatic principles of human interaction (…), but there are certainly other parts that vary from community to community and which have to be learned. (Hudson 1980 [1996]: 224f.)

A connection can be drawn between Hudson and the corpus linguist Hoey's claim that "every word is primed for use in discourse as a result of the cumulative effects of and individual's encounters with the word" (see above). This seems to confirm Hudson's view that a competent speaker has to be competent in the nuances of word use in order to be seen as a competent speaker within his or her community. This is the one point where the highly competent L2 speaker may still fail. They can be able to construct a sentence that is accepted as "grammatically correct" – still this sentence would not be uttered by an L1 speaker. On a micro-scale, the same is true for an L1* speaker who comes from a community that uses one English variant and moves to another community that uses another English variant – where he would be for some purposes an L2* speaker.

As Wolfram, apparently unaware of the work of Meyer / Schvandefeldt, pointed out six years after their seminal work on the human mental capacity of priming was published:

> Ultimately, then, linguistic theory will only be shown correct or incorrect when much more is understood about the operation of human brain neurology. (Wolfram 1978: 12)

Looking at the sum of sociological, economic and cultural differences hinted upon in this section, I believe that a case can be made that the inhabitants of Liverpool stand apart from the average English speaker. As a community apart, its own forms of expression show how language reflects the social position of Scousers as a group.

# Chapter 4 The use of 1st person singular *I* in SCO and MAC

## 4.1 Statistical testing in the research chapters

The pairwise comparisons that will be undertaken in chapters 4 to 10 will be subjected to statistical testing, in order to establish which are statistically significant results. To do so, I will use Paul Rayson's *Log-likelihood Calculator*[76] to undertake tests for all pairwise comparisons in the thesis where there are at least a minimum of five occurrences in both cases. No such tests are undertaken for comparisons where the smaller corpus has fewer than 5 occurrences, as they are likely to be unreliable. Where there are, however, noticeable proportional differences of use, though total numbers found are below 5 occurrences, these will be discussed with the given caveat that low numbers prevent one from drawing any fully conclusive results.

Where statistical testing is undertaken, the comparison will be between the SCO corpus and the MAC corpus[77]. Here, the focus will be on those pairs which indicate that they are significant above the 99.9% level. According to Rayson[76] (see also: Rayson *et al.* (2004)) the level of significance in *Log-likelihood tests* is defined as follows:

[76] http://ucrel.lancs.ac.uk/llwizard.html (last accessed 1/10/10)
[77] bar one exception in chapter 10.2 where the comparison will be SCO:BNC/C.

95th percentile; 5% level; $p < 0.05$; critical value = 3.84
99th percentile; 1% level; $p < 0.01$; critical value = 6.63
99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13

As the majority of the total numbers recorded in SCO in pairwise comparisons with MAC are between 5 and 100, the highest level of probability ($p < 0.0001$) will typically be focused on. This is the equivalent of the critical value reading in a *log-likelihood* calculation (LL) of 15.13 or above.

## 4.2 Introduction to *I*

> Deictic reference is a communicative practise based on a figure-ground structure joining a socially defined indexical ground, emergent in the process of interaction, and a referential focus articulated through culturally constituted schematic knowledge. The horizon of schematic knowledge (…) that practise presupposes, is also produced in the practise. (Hanks 1990: 515)[78]

The use of personal pronouns (*I, you* etc.) is for Hanks (who writes with reference to the language of the Maya) necessarily entwined with cultural practice. The interesting point here is that the "schematic knowledge (…) that practise presupposes, is also produced in the practise". This can be read as knowledge gained through practice. In the context of language use, this seems to link to the propositions of lexical priming. In spoken

[78] American English spelling used in the original.

corpora, the most frequently occurring word of reference is *I*. *I* is one of the many so-called *stance-markers* and found, in particular, in spoken English. Fasulo and Zucchermaglio (2002) say that *I* can be seen as the most direct deictic pointer:

> The first person singular pronoun, 'I', is in principle the least ambiguous among pronouns from a grammatical point of view: indeed, it refers only to one person (unlike 'we', whose members could be vague, and include or not include listeners) and does not risk misidentification (like 'you', who in the presence of many could lead to uncertain attribution). (Fasulo & Zucchermaglio 2002: 1122)

*I* being "in principle the least ambiguous" does not mean, however, that *I* occurs only in a very restricted set of contexts. It simply indicates that other personal pronouns can be more vague when employed.

There appears to be not as much research on the first person singular pronoun available as might be expected. There is widespread reference to the *academic I* (or the lack of it). More literature on the first person singular use appears to occur in psychological and cultural research than in language studies:

> A conception of a person is also coded in the use of person-indexing pronouns, or *deixis*, such as "I" and "you" in English. Deixis are used to indicate extralinguistic entities in discourse: Personal deictic pronouns index the speaker and the addressee within the specific social context. (…)

> Hanks[79] argued that deictic systems evolve, to a large extent, through culturally specific, situated practices. Specific uses of personal deixis in everyday discourse may require users to pay close attention to … personal relationships.
>
> (Kashima & Kashima, p.464: 1998)

Words like *I* or *You* therefore do not exist outside the social context, meaning they tend to be found in less abstract texts such as casual conversation. The reference to Hanks is of particular interest in the context of this thesis: "culturally specific, situated practices" are, after all, what human beings, in the course of their socialization, are primed to follow. As this thesis looks at priming in spoken language, the highest occurring deictic, *I*, is expected to reveal culturally specific usage.

Indeed, Fasulo and Zucchermaglio (2002) claim, based on a sample taken from 10 informants, that utterances with *I* have four discursive functions:

> Four basic classes were identified on the basis of their semantic and pragmatic meaning: *Epistemics, Decisionals, Operatives,* and *Impersonals*. (…) *Epistemic IMU* [I-marked utterance] refers to the speaker's state of knowledge. The range of Epistemics found in the corpus include parentheticals … probability such as *I think*, parentheticals of necessity (mostly of the negative form, such as *I am not convinced*), verbs of perception used in a metaphorical fashion such as *I see*, references to cognitive states such as *I remember*, and expressions of one's inclination for a certain possible line of action, such as *I am in favor* or *I agree*.
>
> (…)

[79] See Hanks, W. F. (1990: 514)

> *Decisional* utterances are those in which the speaker defines his stance toward a given line of action by proposing it to the interlocutors or committing himself to it. … These are modals such as *I shall, I can, I want, I say, I go* (sic)(…)
>
> *Operatives* … are utterances directly concerned with practical operations; they can be reports of things done, in the past tense, of simple announcements of next actions, in the present tense. E.g. *I came here, I begin to* (...)
>
> *Impersonal IMUs* are those where the agent is not the speaker, but a generic person doing the action in question. E.g. *If I click, When I'm doing …*
>
> (Fasulo & Zucchermaglio pp.1125ff.)[80]

Fasulo and Zucchermaglio also note that there is also strong use of *I* as the first word when interrupting a speaker (they refer to them as "cutoffs"). This might be an area worthy of further investigation.

## 4.3 I *in the spoken corpora*

*I* in virtually all sets of spoken utterances plays an important role and can be found in almost every corpus of spoken English as one of the three highest-occurring words:

> Conversation is interactive as a form of personal communication. It is not surprising, then, that conversation shows a frequent use of the first-person *I* and *we* and the second-person pronoun *you*. (Biber *et al.*2002: 5)

[80] For their Italian speakers, the authors found that *operative I* is the most commonly used (over 1/3 of all occurrences). Italian is a pro-drop language, which means one can drop the subject ("I" included). This happens especially in spoken Italian. –Thanks to Pierfranca Forchini for clarifying this point.

As such, the pronoun is a potentially valuable pointer to differences of use between speech communities. If *I* is highly frequent, it does not automatically follow that its nearest collocates and clusters are similar in their frequency in two corpora.

This chapter looks at how this high frequency, freely collocating, word is used in both MAC and SCO and whether this indicates important differences of use.

However, the number of instances of *I* occurring in a single cluster depends very much on how far *I* occurs in speeches or interviews or in casual conversation. For example, the BNC/C subcorpus of the BNC[81] has *I* as the highest occurring word at 3.28% of all words. This compares to a figure of 2.26% for the use of "I" in the spoken BoE (209,583 out of a total of 9.2 million words). This corpus also includes speeches and radio-interviews. MAC records relatively few instances of *I* in its spoken corpus: out of 3.3 million words, *I* occurs 37.076 times – 1.12%. In SCO *I* appears in 2.26% of its 120.000 word corpus.

Table 1 below shows the distribution of *I* in the various corpora:

| Word | Relation | Total "I" | Total Corpus (Tokens) |
|---|---|---|---|
| I (MAC) | 1.13% | 37,127 | 3,300,000 |
| I (SCO) | 2.26% | 2,693 | 119,079 |
| I (BNC/C) | 3.28% | 132,397 | 4,022,428 |
| I(BoE)[82] | 2.26% | 209,583 | 9,272,579 |

Table 1: *I* use in three spoken corpora

[81] The BNC subcorpus (BNC/C) has a total of 4.022.428 tokens. It consists of the *Conversation* BNC SPOKEN files.
[82] BoE – Bank of English(Collins) – this refers to the UKSpoken subcorpus

Table 1 also highlights the importance of *I* in spoken English. The *Relation* column shows what the relative frequency of use is within the whole corpus.

## 4.4. "I" collocates

*I* has the tendency to collocate widely and only short (2-word: 2w) clusters are found with relatively high frequencies, whereas longer clusters (3w and longer) are comparatively rare. In Table 2, the 15 most frequent collocates of *I* are listed. SCO is the point of comparison with both MAC and BNC/C.

Table 2 must be read in two ways. Firstly, within each corpus, the ranking of the collocates (the relative use of the collocates in relation to each other) must be taken into account. Secondly, the relative percentages of the usage of *I* collocates across the corpora needs to be discussed.

Table 2: 15 most frequent collocates to SCO “I” compared to MAC and BNC/C occurrences

| SCO Ranking | Word | % SCO | Total | % MAC | Total | MAC Rank | Word | % BNC/C | Total | BNC/C Rank |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | KNOW | 14.8 | 480 | 13.5 | 4,987 | 5th | KNOW | 15.4 | 20,386 | 6th |
| 2 | AND | 14.8 | 478 | 15.8 | 5,860 | 3rd | AND | 17.3 | 22,895 | 4th |
| 3 | TO | 14 | 453 | 14.1 | 5,224 | 4th | TO | 18.1 | 23,984 | 1st |
| 4 | THE | 14 | 451 | 13.2 | 4,882 | 6th | THE | 17.4 | 23,099 | 3rd |
| 5 | IT | 13.9 | 444 | 21.0 | 7,627 | 1st | IT | 17.5 | 23,231 | 2nd |
| 6 | YOU | 12.5 | 405 | 18.0 | 6,684 | 2nd | YOU | 17.1 | 22,660 | 5th |
| 7 | A | 11.7 | 381 | 10.9 | 4,042 | 10th | A | 13.9 | 18,389 | 8th |
| 8 | THAT | 10.6 | 344 | 12.3 | 4,578 | 7th | THAT | 12.5 | 16,488 | 10th |
| 9 | PAUSE | 10.2 | 331 | n/a | n/a | n/a | PAUSE | >0.1 | 10 | n/a |
| 10 | WAS | 9.7 | 312 | 7.5 | 2,793 | 17th | WAS | 9.7 | 12,777 | 15th |
| 11 | DON'T | 9.6 | 308 | 11.9 | 4,399 | 8th | DON'T | 12.9 | 17,130 | 9th |
| 12 | LIKE | 8.5 | 277 | 5.0 | 1,847 | 36th | LIKE | 5.9 | 7,862 | 27th |
| 13 | THINK | 8.1 | 264 | 11.4 | 4,215 | 9th | THINK | 12.3 | 16,262 | 11th |
| 14 | YEAH | 8.1 | 262 | 7.0 | 2,545 | 27th | YEAH | 5.0 | 6,685 | 35th |
| 15 | IN | 7.9 | 256 | 7.0 | 2,589 | 26th | IN | 7.2 | 9,575 | 21st |

Table 2: 15 most frequent collocates to SCO "I" compared to MAC and BNC/C occurrences

### 4.2.1 Differences in ranking

It is not obvious when the – rather similar – frequencies are compared that there are, amongst the *I* collocates, striking differences in ranking when SCO is compared to MAC and BNC/C. We find that the most frequent collocate for *I* in SCO is *know*. By contrast, this is the 5th most frequent collocate in MAC and 6th most frequent in BNC/C. This shows us, however, that this collocate of *I* with *know* is a vital element of SCO use but is less important as a collocate in either MAC or BNC/C. Looked at in another way, however, the use of *I* with *know* is equally important across the corpora, but other collocates, which are important in MAC and BNC/C, are less so in SCO.

While the percentages are broadly similar for *was* (9.7% in SCO and 7.5% in MAC) and for *yeah* (8.1% in SCO and 7% in MAC[83]) the ranking of these words as collocates is recognisably different. We see that in SCO, *was* is the 10th most frequent collocate of *I* while it is ranked only 17th in MAC and 15th in BNC/C. More striking still, *yeah* is the 14th most frequent collocate with *I* in SCO, but it is only the 27th most frequent in MAC and is even less important in BNC/C where it is ranked 35th most frequent *I* collocate. Similarly, *like* appears as a collocate with *I* in SCO, ranking 12th (accounting for 8.5% of *I* collocates) but it is only ranked 36th (5%) of all uses in MAC. In the BNC/C, *like* is ranked as the 27th collocate (5.9%) with *I*.

[83] see also Table 3 below.

Conversely, the highest-ranking collocate in MAC is *it* and this is also the second most frequently occurring collocate in BNC/C, whereas *it* ranks only as the fifth-highest occurring collocate in SCO. Their proportional frequencies differ notably too, as will be shown below. This means that in SCO, *it* is not avoided as a collocate, but it is not as strong a collocate of *I* as it is in MAC or BNC/C. Similarly, *you* is the second-ranking collocate in MAC, co-occurring in 18.0% of all uses of *I* , but it only ranks sixth in SCO with 12.5% of all co-occurrences.

### 4.2.2 Collocates with different proportional use

In 4.2.1 we looked at which collocates with *I* were the most likely ones used in each corpus. 4.2.2 looks how far collocates appear with divergent frequencies in the different corpora. Focussing on the SCO – MAC comparison, Table 2 shows that the majority of the *I* collocates are occurring with similar frequencies. It also shows a number of collocates where there is a visible divergence between the corpora. The degree of divergence is shown in Table 3 below:

| Item | IT | YOU | THINK | DON'T | YEAH | WAS | LIKE |
|---|---|---|---|---|---|---|---|
| Ratio | 1: 1.51 | 1: 1.44 | 1: 1.41 | 1: 1.23 | 1: 0.86 | 1: 0.77 | 1: 0.59 |
| LL* | 21.68 | 13.04 | 5.59 | 0.36 | 26.51 | 46.69 | 106.28 |

Table 3: Collocates with highest difference in SCO: MAC comparison.
Ratio with sum of entries normalised to SCO=1. *LL stands for *Log-Likelihood.*

There are three words where the ratio of use between SCO and MAC is close to 1:1.5. *Think* co-occurs with *I* in 11.4% of all uses of *I* in MAC (12.3% in BNC/C) but accounts for only 8.1% of instances in SCO. YOU is

strongly used as a collocate with *I* in MAC, where it is ranked the second most frequent collocate, whereas in SCO, it ranks 6th most frequent. It co-occurs with *I* in 18% of all uses of *I* in MAC (17.1% in BNC/C), but only accounts for 12.5% of cases in SCO. A stronger divergence still can be found in the use of *I* with *it*. In SCO, *it* co-occurs with *I* in 13.9% of all uses while it is more prominent in MAC, where it co-occurs with *I* in 21.0% (17.5% in BNC/C) of all cases. Only *think* and *don't* show no divergence of statistical significance, while the difference between SCO and MAC in the use of *like* as an *I* collocate is statistically highly significant.

Three other collocates listed in Table 3 are more prominently used in SCO. *Was* as a collocate of *I* is recognisably more prominent in SCO as it accounts for 9.7% of all cases while *was* as collocate of *I* in MAC appears in only 7.5% of all uses. *Like* appears as a collocate with *I* in SCO ranking 12th and accounts for 8.5% of all uses but it is only ranked 36th (accounting for 5% of all uses) in MAC. In the BNC/C, *like* is ranked as the 27th collocate of *I*, co-occurring in 5.9% of instances of *I*.

### 4.3.1 "I" 2-word clusters

As has been pointed out before, *I* is fairly free-associating. As such, there are relatively few long clusters. *I* can be found in significant numbers mostly in 2-word (2w) clusters. This demonstrates the dividing line between *collocation* and *clustering*: *collocation* refers to the

relationship between two words that do not stand in a fixed position to each other (throughout this thesis, a *collocate* is a word either five words to the left (L) or the right (R) of the target (or node) word). A *cluster,* by contrast, refers to a word that stands in a fixed position to the target word[84].

O'Keefe *et al.* (2007) give a valuable overview of the top 20 two-word chunks of their 5-million-word CANCODE spoken corpus, and I give here an excerpt:

| rank | item | frequency |
|---|---|---|
| 2 | ***I*** mean | 17,158 |
| 3 | ***I*** think | 14,048 |
| 6 | ***I*** don't | 11,975 |
| 8 | and ***I*** | 9,722 |
| 11 | ***I*** was | 8,174 |

Table 4: chunks with *"I"* amongst CANCODE top 20 2w chunks (top 5 "I" 2w clusters).

This gives a good indication what to look for in SCO, MAC and BNC/C:

| rank | item SCO | freq. | item MAC | freq. | item BNC/C | freq. |
|---|---|---|---|---|---|---|
| 1 | I DON'T | 282 | I DON'T | 3,811 | I DON'T | 15,982 |
| 2 | I MEAN | 249 | I MEAN | 3,663 | I MEAN | 15,258 |
| 3 | AND I | 225 | I THINK | 3,326 | I THINK | 14,228 |
| 4 | I WAS | 205 | I , I | 3,302 | AND I | 10,704 |
| 5 | I THINK | 197 | I KNOW | 2,728 | I I* | 9,846 |
| 6 | I KNOW | 148 | | | | |
| 6 | I-I | 148 | | | (* no comma in BNC/C concordance) | |

Table 5: Most frequent 2w clusters (chunks) with *I* in SCO, MAC and BNC/C.

Tables 4 and 5 show that the most frequent 2w clusters involving *I* are found both in SCO and CANCODE, and that the degree of convergence

[84] Consequently *I* with *the* is a collocate: **the I* would be rare if at all found in a concordance, while *the movie I went to see* is likely. On the other hand, both *I was* and *was I* are possible (2w) clusters.

with MAC and BNC/C is also very high. Consequently, this section will necessarily focus on the medium-high frequent *I* clusters where differences of use is apparent.

#### 4.3.1.1 "I" 2w clusters: areas of divergent use

In this section, I want to look at some 2w clusters that were highlighted in section 4.2 as recognisably different in their collocational frequency in the two corpora. Above, I have shown that while *I* with *got* is proportionally more frequent in MAC, *I* with *was* as collocate is proportionally more frequent in SCO. This is one of the kind of clusters Fasulo and Zucchermaglio (2002) describe as *operatives*. However, both *I* with *like*[85] and *know* (which Fasulo and Zucchermaglio would class as *epistemic*) are also far more frequent in SCO than MAC.

#### 4.3.1.2 "I" 2w clusters: SCO more frequent

Table 6 present five clusters, all of which appearing with a statistically significant higher proportional frequency in SCO. *I like* and *I just* will be discussed in detail in the *Discourse Particle* section (Chapter 9), I will focus here on the two other clusters.

[85] This refers to only one possible use of *LIKE* (preference). See chapter 8.2.6 for more details.

| Cluster | occ. SCO | % SCO | occ. MAC | % MAC | log likelihood |
|---|---|---|---|---|---|
| I WAS | 205 | 6.3 | 1683 | 4.5 | 42.93 |
| I WASN'T | 13 | 0.48 | 62 | 0.17 | 9.55 |
| WHAT I | 157 | 4.8 | 1187 | 3.2 | 42.96 |
| WHAT I MEAN | 61 | 2.3 | 237 | 0.6 | 59.74 |
| I JUST | 119 | 3.7 | 504 | 1.4 | 104.03 |
| I LIKE | 81 | 2.5 | 488 | 1.4 | 39.04 |

Table 6: SCO 2w "I" clusters divergent.

*I was* has been found, in a detailed review of its usage, to be dissimilar in SCO in marginal uses only: In the 1683 occurrences of *I was* in MAC, only 62 are instances of *I wasn't*. In the 205 occurrences of SCO, however, *I wasn't* is proportionally used far more often, appearing 13 times However, this is significant only at a 99.0% level while all other comparison are significant above the 99.99% level.

The 2w cluster *what I* is proportionally used more in SCO than in MAC (or, for that matter, the BNC/C). We also see that *what I* appears in the majority of 3w clusters (38.9%) as being part of *what I mean* in SCO. In MAC, *what I mean* occurs proportionally less often (19.97%). *What I mean* appears to be the reason why *what I* clusters are overall found with a higher proportional percentage in SCO when compared to MAC. This will be discussed in detail in 4.3.4 below. Table 6 also shows higher overall proportional use of forms of *what I want* in SCO compared to MAC but all of these clusters appear with far lower frequencies than *what I mean*.

#### 4.3.1.3 "I" 2w clusters: MAC more frequent

By a noticeable margin, the majority of the most frequently used 2w clusters with *I* occur with higher proportional frequencies in MAC than in SCO. Most of the clusters in question are the highest-frequency 2w *I* clusters.

| Cluster | occ. SCO | % SCO | occ. MAC | % MAC | MAC=1:SCO | log likelihood |
|---|---|---|---|---|---|---|
| I DON'T | 282 | 8.7 | 3811 | 10.26 | 1. \| : \| 0.847953 | 0.10 |
| I MEAN | 249 | 7.7 | 3663 | 9.9 | 1. \| : \| 0.777778 | 1.00 |
| I'VE | 230 | 7.1 | 3272 | 8.8 | 1. \| : \| 0.806818 | 0.21 |
| I THINK | 197 | 6.1 | 3326 | 9.0 | 1. \| : \| 0.67777 | 8.11 |
| I KNOW | 148 | 4.6 | 2728 | 7.4 | 1. \| : \| 0.621622 | 12.93 |
| I - I | 148 | 4.6 | 3302 | 8.9 | 1. \| : \| 0.516854 | 42.93 |
| I SAID | 102 | 3.1 | 2014 | 5.4 | 1. \| : \| 0.574074 | 42.96 |
| I CAN | 96 | 3.0 | 1821 | 4.9 | 1. \| : \| 0.612245 | 59.74 |
| BUT I | 92 | 2.8 | 1809 | 4.8 | 1. \| : \| 0.6250 | 103.18 |

Table 7: "I" 2w clusters MAC more frequent. Normalisation to MAC=1 in the right-hand column

Table 7 shows that I *don't*, *I mean* and *I've* are found with the same ranking of occurrence in SCO and MAC, they are proportionally about 2% more frequent in MAC. Neither of them are at statistically significant divergent levels. A higher frequency of use is also the only difference to be found between SCO and MAC in the medium-high frequency clusters – *I think* (which is found at a level of difference that is only 99.0%

significant) as well as *I said* and *I can*. Both *I said* and *I can* appear mostly in the same nesting (*I said I, and I said; I can , I can't*)[86].

When we look at the 1809 occurrences of *but I* in MAC, 246 occurrences are *but I mean* - equalling 0.7% of all uses of *I*. That is nearly three times as frequent, proportionally, as in SCO, where there are 7 occurrences of *But I mean* – these equal 0.25% of all uses of *I*. Likewise, *but I think* occurs in 0.3% of *I* uses in MAC, but only in a marginal 0.15% (4 occ.) of uses in SCO.

The occurrence pattern of *I know* will be discussed in-depth in chapter 9.3.2, as it is an instrumental part of longer clusters.

*I think* appears in 9% of all uses of *I* in MAC (3326 occ.) but is about one-third lower in SCO: 6.1 % (197 occ.). This does include the negation *I don't think*, which is more dominant in MAC (1.6% compared to 1.1%).

> … in modals of probability like *I think* or *I believe*, a certain state of affairs is by the laws of rationality true in many but not all possible worlds. These expressions can then be considered mitigation devices and, in the taxonomy proposed by Caffi (1999), would be classified as ''hedges'', i.e., affecting ''the illocutionary force of the utterance'' and modulating the relationship between the speaker and the saying.
>
> (Fasulo and Zucchermaglio 2002: 1127)

This points to strong use of *epistemic I* in MAC and suggests that conventionally accepted hedging appears to be slightly stronger in *I* clusters in MAC than in SCO.

[86] MAC also has a sizeable number of the cluster *I says* – with 612 occurrences this makes up 1.2% of all uses of "I". It clearly reflects a regional accent captured by the MAC corpus and there is not one occurrence of it in SCO (though, anecdotally, *I says* is a characteristic figure of speech in Scouse narrative).

While *I think it ('s)* is the most frequent cluster with *I think* in both corpora, it is found in 14.2% of all uses of *I think* in SCO, but only in 10.73% of all uses in MAC. Another difference is the occurrence of *I think er(m)*. This accounts for only 0.87 % (29 occ.) of all uses of *I think* in MAC, yet is recorded in 2.53% (5 occ.) of all *I think* uses in SCO. The strongest divergence is found when we look at the second-most common 3w *I* cluster containing *I think* in MAC: *I think I*. This accounts for 9.26% (308 occ.) of all uses of *I think* in MAC. Similarly, *I think I* is found 1,589 times out of a total 14,228 times of *I think* in BNC/C – 11.17%. In stark contrast, *I think I* is marginal in SCO, where we find only two occurrences (1%).[87]

The comparatively robust stance taken through the use of *I think* is augmented by the use of *but I*. These clusters appear significantly more frequently in MAC than they do in SCO.

With regards to *I – I*, as will be shown in a later chapter, in particular with the term *really*, MAC shows a tendency for single word repetition and multiple repetition that is not replicated in SCO. The cluster *I-I* appears nearly twice as often in MAC than SCO. The repetition of *I* itself in SCO fits into a pattern. Indeed, multiple repetition is rare in SCO but quite a common feature in MAC. Consequently, *I – I – I* accounts for only 0.8% (42 occurrences) of all of *I* in SCO – and that is the maximum repetition found in relevant quantities. In MAC, it appears in

[87] In SCO, these two occurrences are *I think I'm* to be precise. The use of *I think he* and *I think that('s)* is roughly the same in both MAC and SCO, though.

1.8% (946 occ.) of all occurrences of *I* and even *I - I - I - I* is attested, accounting for 0.47 % (174 occ.) of all occurrences[88].

This could be seen as a reflections of SCO speakers as more confidently fluent speakers.

#### 4.3.2.1 Long clusters with the negations I'M NOT and I CAN'T

In this section we explore how far the usage of **I'm not** and **I can't** differs in the two corpora.

| item | SCO tot. | SCO % | MAC tot. | MAC % | Log-Likelihood | BNC/C tot. | BNC/C % |
|---|---|---|---|---|---|---|---|
| I CAN'T | 51 | 1.6 | 902 | 2.4 | 3.23 | 2899 | 1.7 |
| I'M NOT | 78 | 2.4 | 867 | 2.2 | 3.13 | 2756 | 1.6 |

Table 8: Occurrence distribution of *I can't* and *I'm not* amongst *I* use in SCO, MAC and BNC/C

Table 8 above shows the proportional frequencies of the 2w clusters *I can't* and *I'm not* are similar both in relation to each other and in the three corpora. Figure 1 (next page) shows the highest occurring clusters, with SCO as the point of comparison. Deeper analysis shows, however, that only one 3w cluster incorporating the 2w cluster *I'm not* is employed with about the same proportional frequency: *I'm not gonna*.

By contrast, the 3w cluster *no I'm not* (a very finite statement) is the only one of the clusters incorporating the 2w *I'm not* that is used markedly more often in MAC than SCO. Table 9 gives the respective proportional

[88] The number is similar in BNC/C, where *I - I - I – I* occurs 136 times.

figures as 6.2% in MAC (8.2% in BNC/C) compared to 3.8% in SCO. Further *I'm not* clusters found in MAC but rarely in SCO are *well I'm not* (45 occ.) and *I'm not going* (44 occ.), where the former is not recorded in SCO and the latter appears only twice in SCO.

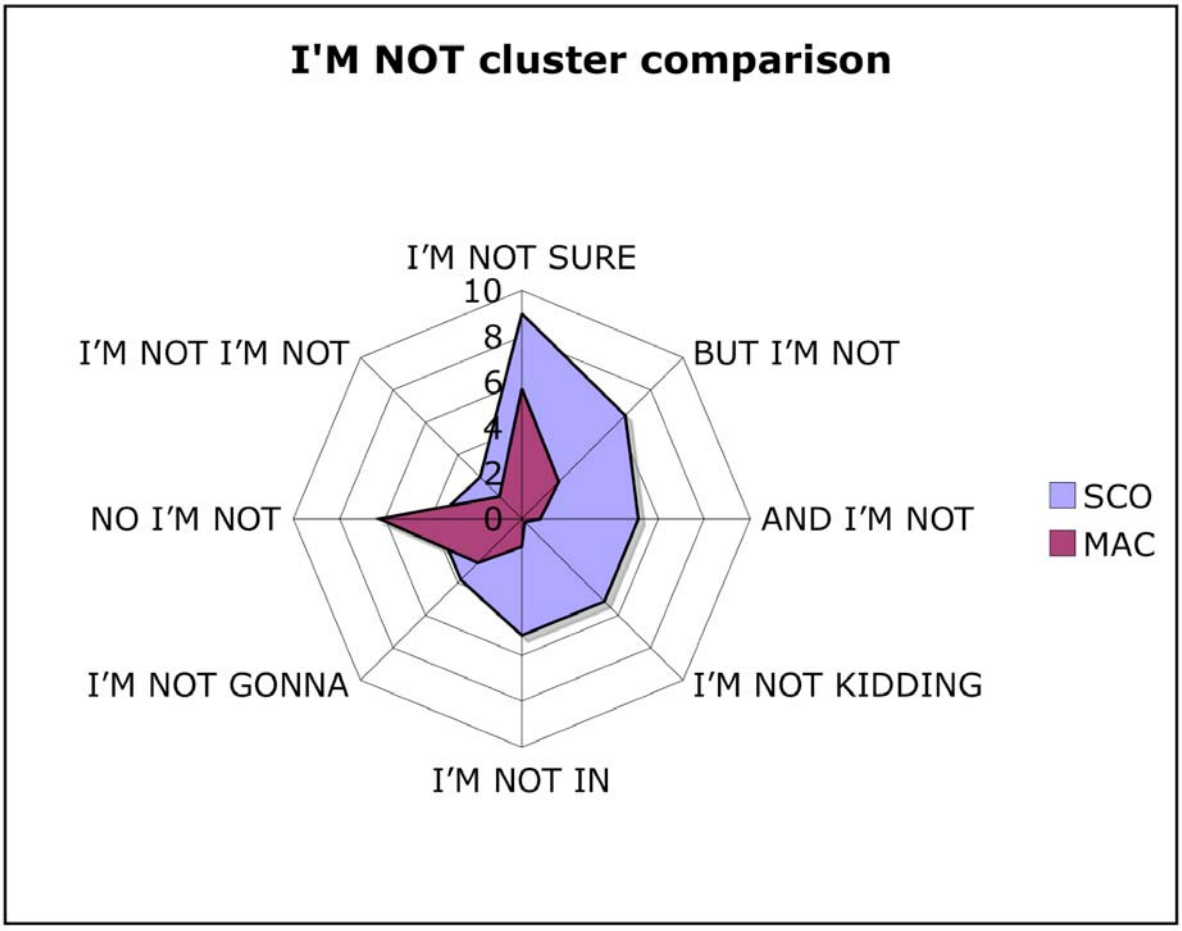



Figure 1: SCO vs. MAC differences of use in I'M NOT clusters made visible

Table 9: Comparison of I'M NOT clusters (% in relation to I'M NOT) in SCO, MAC and BNC./C. (Percentages as of *I'm not* occ.).

| *I'M NOT cluster* | SCO Freq. | SCO % | MAC Freq. | MAC % | BNC/C Frq. | BNC/C % | log likelihood |
|---|---|---|---|---|---|---|---|
| I'M NOT SURE | 7 | 9.0 | 49 | 5.7 | 228 | 8.3 | 1.17 |
| BUT I'M NOT | 5 | 6.4 | 20 | 2.3 | 118 | 4.3 | 3.37 |
| AND I'M NOT | 4 | 5.1 | 7 | 0.80 | 53 | 1.9 | |
| I'M NOT KIDDING | 4 | 5.1 | 2 | 0.2 | 11 | 0.4 | |
| I'M NOT IN | 4 | 5.1 | 10 | 1.2 | 29 | 1.1 | |
| I'M NOT GONNA | 3 | 3.8 | 23 | 2.7 | 190 | 6.9 | |
| NO I'M NOT | 3 | 3.8 | 54 | 6.2 | 225 | 8.2 | |
| I'M NOT I'M NOT | 2 | 2.6 | 12 | 1.4 | 42 | 1.5 | |

Table 9: Comparison of I'M NOT clusters (% in relation to I'M NOT) in SCO, MAC and BNC./C. (Percentages as of *I'm not* occ.).

Conversely, the hedge *I'm not sure* is noticeably more widely employed in SCO (9.0%) than in MAC (5.7%). However, *I'm not sure* is nearly as frequent in BNC/C (8.3%) as in SCO. This is one of the rare occasions where MAC is the outlier[89]. As the statistical test shows, the difference is of no significance.

More striking is the occurrence of the cluster *I'm not kidding*, which appears proportionally 25 times more often in SCO than in MAC (over 12 times more frequently than in BNC/C). *I'm not kidding* appears to be likely to be a SCO-specific phrase. To a lesser degree, this is also true for *and I'm not*. Because of the low numbers, no statically secure conclusions can be made, yet were we to project the proportional occurrences on to larger corpora, the difference would be significant[90]. This indicates how a particular form of negation with *I* may have a different field of semantic association for SCO speakers when compared to MAC speakers.

Turning now to **I can't**, we can find beyond the similarities that the main difference across the corpora is the unequal distribution of verbs following this cluster. This is shown in Table 10. As far as the low numbers allow a judgement here, it has to be the following:

[89] This is also underlined by the fact that the long cluster BUT I'M NOT SURE is recorded only once in MAC, but twice in the smaller SCO

[90] Were we to double the corpora and therefore the occurrence numbers, the LL figure for *And I'm not* would be 10.18 (above the 99.0% significance level) and for *I'm not kidding* it would be 16.73 (99.99%).

Table 10: SCO highest-occurring terms to the right of I CAN'T (percentages per I CAN'T occurrences) and MAC / BNC/C equivalents.

| Term/s (R.) | SCO total | SCO % | MAC total | MAC % | BNC/C tot. | BNC/C % |
|---|---|---|---|---|---|---|
| - see | 3 | 5.9 | 70 | 7.8 | 227 | 7.8 |
| - remember | 4 | 7.8 | 66 | 7.3 | 373 | 12.8 |
| - I | 1 | 1.6 | 42 | 4.2 | 101 | 3.5 |
| - I CAN'T | - | - | 8 | 0.8 | 45 | 1.6 |
| - understand | - | - | 37 | 4.1 | 71 | 2.4 |
| - get | 3 | 5.9 | 19 | 2.1 | 150 | 5.2 |
| - imagine | 3 | 5.9 | 6 | 0.7 | 16 | 0.6 |
| - do | 3 | 5.9 | 10 | 1.1 | 155 | 5.4 |
| - think | 2 | 3.7 | 16 | 1.8 | 102 | 3.5 |

Table 10: SCO highest-occurring terms to the right of I CAN'T (percentages per I CAN'T occurrences) and MAC / BNC/C equivalents.

*I can't*, in MAC, has a strong preference to be followed on the right (R.) by these verbs: *see; remember; understand*; as well as a repetition of *I*. All other verbs are clearly occurring with a lower frequency after *I can't*. *See, remember* and *understand* are all verbs that describe internal states or forms of perception.

By contrast, the verb cluster of perception *I can't imagine* is used with similar frequency to many other clusters (i.e. *I can't see*) in SCO, yet the use of *I can't imagine* is rather low in use in both MAC and BNC/C.

The clusters *I can't do* and *I can't get* are relatively frequent in their use in SCO but marginal in their use in MAC, *do* and *get* are verbs that may reflect external states. (In BNC/C, however, the proportional figures are close to SCO).

This seems to highlight – as far as the low figures for both corpora allow – that the semantic associations of *I can't* are usually bound to verbs of perception in MAC, while SCO users employ *I can't* equally with verbs of internal and external states. Where all three corpora are compared, the phrase *I can't imagine* stands out as being in relatively strong use in SCO and marginal in use in MAC and BNC/C. One may draw the conclusion that *I can't imagine* is being used by SCO speakers instead of the phrase *I can't understand,* especially as the latter is not recorded at all in SCO.

### 4.3.3 Longest available clusters

As described before, *I* easily collocates with a large variety of words and there are many 2w clusters while there are few really long clusters with *I* that appear with any relevant frequency in any corpus. Most of the long *I* clusters incorporate *I don't*:

| Cluster | SCO tot. | MAC tot. | Log-Likelihood |
|---|---|---|---|
| YOU KNOW WHAT I MEAN | 46 | 153 | 54.06 |
| I DON'T KNOW I | 15 | 135 | 2.19 |
| THAT'S WHAT I | 13 | 161 | 0.13 |
| I DON'T KNOW WHAT | 16 | 113 | 5.31 |
| I DON'T KNOW WHETHER | 7 | 53 | 1.91 |
| I DON'T THINK IT | 6 | 55 | 0.81 |
| I DON'T KNOW I DON'T | 6 | 52 | 1.03 |
| I DON'T KNOW WHY | 4 | 40 | n/a |

Table 11: Longest *I* clusters in SCO compared.

Table 11 makes clear that only one of the long *I* clusters appears with a statistically significant difference in proportional frequency where SCO is compared with MAC. This is the most frequent 5w *I* clusters in SCO and its occurrence pattern will be discussed in more detail in 4.3.4.

### 4.3.4 /You know /what I /I mean / – *2w clusters form a longer, meaningful, cluster*

The 2w clusters *you and I mean know* appear proportionally less in SCO than in MAC, while the opposite is true for *what I*. As Table 12 shows, neither the 2w cluster *you know* nor *I mean* appears with statistically significant differences of proportional frequencies either:

| 2w cluster | SCO total | SCO % | MAC total | MAC % | Ratio with entries normalized to MAC=±1 | log likelihood |
|---|---|---|---|---|---|---|
| YOU KNOW | 143 | 5.31 | 2,613 | 7.04 | 1. \| : \| 0.754261 | **11.72** |
| WHAT I | 157 | 4.8 | 1,187 | 3.2 | 1. \| : \| 1.5 | **42.96** |
| I MEAN | 249 | 7.7 | 3,663 | 9.9 | 1. \| : \| 0.777778 | **1.00** |
| I MEAN I* | 20 | 0.742 | 1238 | 3.34 | 1. \| : \| 0.222156 | **75.79** |

Table 12: *you know*, *I mean* and *what I* occurrence percentages in MAC compared to SCO. **I Mean I* and *I mean, I* combined figures.

Despite these differences, investigation into the respective environments of these three 2w clusters reveals that they tend to form constitute parts of clusters with you *know* and *I mean*. While the most frequent 3w cluster *incorporating you know*, *you know I,* appears with similar proportional frequencies, the 2w cluster *I mean* and in particular the 3w *I* cluster *I mean I*[91] are found with far lower proportional frequencies in SCO than in MAC. Consequently, further investigation into the uses of *I mean* groups of clusters is needed.

Schourup says that *I mean* has been investigated widely and concisely describes its function in discourse thus:

> … *I mean* indicates that what is said and what is meant may well be substantially non-equivalent and, unless repair is undertaken, could lead to misunderstanding. It is thus important that *I mean* but not *like* prefaces corrections. (Schourup, p.148: 1985)

[91] This mirrors the use of *I think I* as discussed in section 4.3.1.3.

Furthermore, Brinton (2003)[92] describes *I mean* as follows:

> As such, *I mean* has procedural meaning and is best analyzed as a discourse, or pragmatic, marker. (…)*I mean* also expresses of range of speaker attitudes.
>
> (Brinton, p.1: 2003)

As *I mean* is meant to indicate a repair or clarification in what has just been said, it is unsurprising to find the cluster *I mean* as part of the hesitation / repetition feature *I mean I* in MAC. This cluster appears 1.8% of all times *I* is used. Another, similar, cluster is *I mean (pause) I* which occurs 1.5% of all *I* uses. Taken together they are one of the highest–occurring 3w clusters with *I* in MAC. By contrast, *I mean I* appears only in 0.74% in SCO. While a speaker would expect *I mean* to be followed often by *I* in MAC, this is not the case in SCO.

If we look at how 2w clusters contribute to the formation of longer clusters we find that while there are a number of clusters which are used with proportionally the same frequency in both SCO and MAC, other clusters show divergent frequencies. Staying with the 2w cluster *I mean* as part of *what I mean*, we find it occurs in 1.8% (61 times) of all uses of *I* in SCO. In MAC, *what I mean* occurs 237 times (0.64%).

Amongst these clusters, we find two SCO specific uses: There is the 4w cluster *See what I mean* which occurs 8 times in SCO (13.1% of all uses of *what I mean* and 0.3% of all *I* uses). In the far larger MAC

[92] Brinton also provides an excellent overview on the literature regarding *I mean* usage.

corpus, however, *see what I mean* is only recorded 6 times – 2.53% of all occurrences of *what I mean*[93]. This is about the same percentage found in BNC/C, where *see what I mean* occurs 87 times (2.73% of the 3,185 occurrences of *what I*). Furthermore, these 2w clusters form part of the 5w cluster *you know what I mean* which, with 46 occurrences and 1.4% of all uses of *I* in SCO appears to be a fixed phrase. In marked contrast, *what I mean* accounts for only 0.7 % of all *I* occurrences in MAC. The long cluster *you know what I mean* appears in MAC in only 0.42% (153 occ.) of all *I* uses. It is even less frequent in the BNC/C – 326 occurrences (0.25%)[94]. This means the phrase would appear 171 times in every 10, 000 words in SCO but only 41.3 times in MAC and only 24.8 times in BNC/C. This not withstanding, *you know what I mean* is the longest *I* cluster of high frequency in SCO, MAC and BNC/C.

Looking at the nesting of *you know what I mean*, we find that it is followed in nearly a third of all cases by a pause in MAC. In SCO, pauses occur in 8 out of 43 cases (18.6%). Another four instances (8.7%) of *You know what I mean* in SCO are followed by a short pause and twice (4.3%) they are followed by a laugh or a hesitation marker (*erm*). In all other cases, they are either directly followed by somebody else's turn or part of a longer utterance. This use of pause may indicate that the speaker has ended her / his turn but there is uncertainty whether what was said is agreed with or understood. The 240 concordances lines in MAC are similar: Twice, the phrase is followed by *erm*, 17 times (0.05 %) it is part

[93] LL value of 24.82 - highly significant therefore.
[94] See Cpt. 9.3.3 on *clusters* for a detailed discussion of BNC/C use.

of a longer phrase. 42 times (0.12%) it is followed by a short pause and 24 times (0.06%) it is at the end of a turn. This indicates that, while the phrase is employed far more widely in SCO, the usage itself does not differ in the majority of cases from the way it nests in MAC.

There is, however, evidence of divergence of the phrase in a minority of cases, where *You know what I mean* finds itself in the neighbourhood of different words: *You know what I mean* is preceded by the utterance *yeah* a number of times in both corpora (9 = 3.75% in MAC; 4 times = 8.7% in SCO). *Yeah* has been said by the previous speaker, so the phrase is employed to check understanding.

Similarly, while *you know what I mean* is followed by *so* three times (6.5%) in SCO (always still part of the same utterance) *so* follows *you know what I mean* only twice (0.8%) in MAC.

This means that SCO speakers would be more likely to assume *you know what* to have preceded *I mean*. This can be seen as different primings (collocationally and colligationally) when it comes to the phrase *I mean.* It also tentatively suggests that the *nesting* of the phrase is dissimilar in the two corpora.

## 4.4 Conclusions of *"I"* usage in the corpora

*I* is one of the most frequently occurring words in casual spoken English. As such, its usage gives sufficient evidence of occurrence patterns in corpora. We see that, amongst long *I* clusters, the differences

are marginal - with the one exception of the phrase *You know what I mean*. *You know what I mean* appears to be a set, fixed phrase in English, being the most frequent long *I* cluster in SCO, MAC and BNC/C. It appears particularly favoured by SCO informants, in whose speech it appears over three times more frequently than amongst MAC sources. When SCO and MAC are compared, *you know what I mean* appears to be set (*nesting*) amongst different words in its use as well.

When we look at *I* collocates, we notice that, while the proportional frequencies between SCO, MAC and BNC/C may be similar, the *order* (or *rank*) of their occurrences are different. A SCO speaker would use *know* as the most likely collocate with *I*, while the collocate preferred by MAC or BNC/C speakers appears to be *it*. Therefore, the collocate *know* has higher attraction to the target word *I* in SCO than in MAC. When we turn our attention to the proportional *frequency of use* on the other hand, we see that *know* is, proportionally, about as frequent in SCO as it is in MAC. *I* with *yeah*, *like* and *just* are clearly more frequently used in SCO compared to MAC. These collocates are all used in a more prominent way in Scouse than UK-wide English. Had we seen widely differing use of *I* between the corpora at all times that could have been seen as evidence that one corpus reflects either a different language[95].

[95] With the search chunk *have a* Glenn Hadikin also used the SCO corpus in a comparison to BNC/C and his own corpora of Korean L2 English speakers. He, too found that SCO and BNC have a high degree of proportional similar use. (Hadikin, G.: Lexical Priming in L2 English: a comparison of two Korean communities: presentation given at UoL School of English PG seminar, 28/05/09)

We found that *I* tends to cluster widely and therefore that 2w clusters are most appropriate for comparison. Here, we have seen that all the most frequent SCO 2w *I* clusters diverge in their proportional frequencies significantly from their use in MAC. On the other hand, the most frequent MAC 2w *I* clusters are used in a similar way in SCO. It is only amongst the medium-high frequency clusters that occur more often in MAC, in cluster like *I said* or *I can* and, in particular, *but I*, that significant divergence is apparent.

Overall, the available data does support the notion of *lexical priming*. There is some difference and these indicate that is variety. It does not, however, show the kind of difference that support the view that there is such a strong difference between SCO and MAC that the former could be classed as a dialect. There are some noticeable variations in use, but, on the whole, both corpora use the term in very similar ways.

At the same time, MAC has a higher proportional frequency of clusters like *I know, I mean, I mean I* and *I think* that hint at a stronger assertiveness in the tone of the speakers. In MAC we also see a much higher occurrence of frequent I repetition (e.g. *I - I - I*).

It is, to sum up, in these noticeable differences that evidence of priming can be traced: SCO speakers tend to have a set of words and phrases they are more likely to both use and expect in the vicinity of *I* in a number of cases. This can be seen as a social strategy, and, one may conclude, the choices of language that SCO speakers appear to be primed

to make reflect at times what can be called *strategic hedging.* Speakers tend to avoid definitive statements or assertiveness (i.e. strong single-word repetition). This is a way of being cautious with utterances in order to protect the speaker from being countered. Given the historical development of the city of Liverpool – immigrants from all over the UK and Europe; casual labour where different people would work together from one day to the next; the conflict of Catholics versus Protestants – it can be seen why *strategic hedging* might have become internalised by speakers in Liverpool.

| 2w cluster | MAC total | MAC % | SCO total | SCO % | Ratio with entries normalized to MAC=±1 | log likelihood |
|---|---|---|---|---|---|---|
| I DON'T | 3,811 | 10.26 | 282 | 8.7 | 1. \| : \| 0.847953 | 0.10 |
| I MEAN | 3,663 | 9.9 | 249 | 7.7 | 1. \| : \| 0.777778 | 1.00 |
| I'VE | 3,272 | 8.8 | 230 | 7.1 | 1. \| : \| 0.806818 | 0.21 |
| AND I | 2,717 | 7.4 | 225 | 6.9 | 1. \| : \| 0.932432 | 3.51 |
| I WAS | 1,683 | 4.5 | 205 | 6.3 | 1. \| : \| 1.4 | 42.93 |
| I THINK | 3,326 | 9 | 197 | 6.1 | 1 \| : \| 0.677778 | 8.11 |
| WHAT I | 1,187 | 3.2 | 157 | 4.8 | 1. \| : \| 1.5 | 42.96 |
| I KNOW | 2,728 | 7.4 | 148 | 4.6 | 1. \| : \| 0.621622 | 12.93 |
| I - I | 3,302 | 8.9 | 148 | 4.6 | 1. \| : \| 0.516854 | 38.15 |
| YOU KNOW | 2,613 | 7.04 | 143 | 5.31 | 1. \| : \| 0.754261 | 11.72 |
| I JUST | 504 | 1.44 | 119 | 3.7 | 1. \| : \| 2.56944 | 104.03 |
| I SAID | 2,014 | 5.4 | 102 | 3.1 | 1. \| : \| 0.574074 | 13.98 |
| WHEN I | 1,265 | 3.4 | 97 | 3.0 | 1. \| : \| 0.882353 | 0.27 |
| I CAN | 1,821 | 4.9 | 96 | 3.0 | 1. \| : \| 0.612245 | 10.23 |
| BUT I | 1,809 | 4.8 | 92 | 2.8 | 1. \| : \| 0.583333 | 12.30 |
| SO I | 999 | 2.7 | 83 | 2.5 | 1. \| : \| 0.925926 | 1.36 |
| I LIKE | 487 | 1.41 | 81 | 2.5 | 1. \| : \| 1.77305 | 39.21 |
| I GOT | 625 | 1.7 | 76 | 2.2 | 1. \| : \| 1.29412 | 15.82 |
| YEAH I | 842 | 2.1 | 47 | 1.50 | 1. \| : \| 0.714286 | 3.31 |
| KNOW WHAT I | 269 | 0.73 | 51 | 1.89 | 1. \| : \| 2.58904 | 31.71 |
| I I I | 946 | 2.55 | 42 | 1.56 | 1. \| : \| 0.611765 | 11.30 |
| YEAH I KNOW | 259 | 0.699 | 15 | 0.56 | 1. \| : \| 0.800114 | 0.77 |
| YOU KNOW I | 390 | 1.05 | 26 | 0.966 | 1. \| : \| 0.92 | 0.18 |
| I DON'T KNOW | 1312 | 3.53 | 116 | 4.3 | 1. \| : \| 1.21813 | 3.95 |
| I DON'T THINK | 588 | 1.58 | 39 | 1.45 | 1. \| : \| 0.917722 | 0.30 |
| I SAID I | 530 | 1.43 | 20 | 0.67 | 1. \| : \| 0.468531 | 10.14 |
| I MEAN I | 1238 | 3.34 | 20 | 0.74 | 1. \| : \| 0.222156 | 75.79 |
| I WAS LIKE | 0 | 0 | 19 | 0.6 | n/a | n/a |

Table 13: *"I"* 2w and 3w clusters,
MAC : SCO proportional frequency comparison with MAC normalized to ±1.

## 4.5 Corpora conductive to comparison

That we get different results in proportional frequency of use when MAC and SCO can have a number of causes:

1. SCO presents different results because it is recording different socio-geographic usage.
2. MAC could be an outlier and therefore make SCO evidence look different.
3. SCO and MAC cover transcripts of informants that are participating in different genres.
4. The small size of SCO magnifies differences into apparent significance.

The genre issue (3) is irresolvable. However, Table 13 shows those clusters where there is a clear difference in proportional use between SCO and MAC. To find out, however, how far MAC can be seen as an outlier (point 2)), it needs to be compared to another general corpus. Throughout this chapter, to be sure that MAC is a comparator that reflects UK spoken English well, at times BNC/C figures are shown to make the difference in use between SCO and the comparators clearer. To demonstrate that MAC mostly (though not at all times) records proportional frequencies of use that are comparable to the ones found in the BNC/C, Table 14 shows *I* use 2w and 3w clusters ratios with entries normalised to MAC=±1.

| 2w cluster | MAC total | MAC % | BNC/C total | BNC/C % | Ratio with entries normalized to MAC= ±1 |
|---|---|---|---|---|---|
| I DON'T | 3,811 | 10.26 | 15982 | 9.45 | 1. \| : \| 0.9211 |
| I MEAN | 3,663 | 9.9 | 15258 | 9.00 | 1. \| : \| 0.9091 |
| I'VE | 3,272 | 8.8 | 10,611 | 6.28 | 1. \| : \| 0.7136 |
| AND I | 2,717 | 7.4 | 10704 | 6.33 | 1. \| : \| 0.8555 |
| I WAS | 1,683 | 4.5 | 9255 | 5.48 | 1. \| : \| 1.2178 |
| I THINK | 3,326 | 9 | 14228 | 8.42 | 1 \| : \| 0.9356 |
| WHAT I | 1,187 | 3.2 | 3185 | 1.90 | 1. \| : \| 0.594 |
| I KNOW | 2,728 | 7.4 | 8655 | 5.12 | 1. \| : \| 0.6919 |
| I - I | 3,302 | 8.9 | 9846 | 5.82 | 1. \| : \| 0.6652 |
| I JUST | 504 | 1.44 | 2704 | 1.60 | 1. \| : \| 1.111 |
| I SAID | 2,014 | 5.4 | 9490 | 5.60 | 1. \| : \| 1.037 |
| WHEN I | 1,265 | 3.4 | 4105 | 2.40 | 1. \| : \| 0.7059 |
| I CAN | 1,821 | 4.9 | 3289 | 1.95 | 1. \| : \| 0.398 |
| BUT I | 1,809 | 4.8 | 6309 | 3.70 | 1. \| : \| 0.7708 |
| SO I | 999 | 2.7 | 4299 | 2.54 | 1. \| : \| 0.9407 |
| I LIKE | 487 | 1.41 | 1720 | 1.30 | 1. \| : \| 0.922 |
| KNOW WHAT I | 269 | 0.73 | 678 | 0.51 | 1. \| : \| 0.699 |
| I I I | 946 | 2.55 | 1334 | 1.00 | 1. \| : \| 0.3922 |
| YEAH I KNOW | 259 | 0.699 | 681 | 0.52 | 1. \| : \| 0.744 |
| YOU KNOW I | 390 | 1.05 | 1299 | 0.98 | 1. \| : \| 0.9333 |
| I DON'T KNOW | 1312 | 3.53 | 5901 | 4.46 | 1. \| : \| 1.263 |
| I DON'T THINK | 588 | 1.58 | 2585 | 1.95 | 1. \| : \| 1.234 |
| I SAID I | 530 | 1.43 | 1276 | 0.96 | 1. \| : \| 0.6713 |
| I MEAN I | 1238 | 3.34 | 2959 | 2.23 | 1. \| : \| 0.6677 |

Table 14: *"I"* 2W AND 3W CLUSTERS, MAC : BNC/C proportional frequency comparison with MAC normalized to ±1.

This can be used to directly compare in how far MAC and BNC/C show similar proportional percentages of use for these *I* clusters. While there are clear outliers where MAC is far more frequent in its proportional use – for example, *I can* or repetition of *I* as in *I-I* or *I-I-I* - overall the differences are found to be less strong than those shown in Table 12 where MAC *I* cluster use is compared to *SCO* occurrences.

This leaves point (1) and (4). Dealing with the latter first, there is no doubt that, where the difference between 3 or 4 occurrences results in, say 10 full percentage points, statistical significance would be based very much on chance events. Though the reader has to be aware of this, the variations often found occur at a subtle level. Consequently, both the consistency of variation and, in particular, the fact that a clear number are found to have total occurrence figures that are close to those found in the much larger comparators support the validity of claims that are made for SCO. This, tentatively, leads me to conclude that while the points (2) - (4) made above may play some role in explaining the difference between SCO and its comparators, the main reasons for differences found appear to be based on the fact that SCO records different socio-geographic usage.

# Chapter 5 Uses of Indefinite Pronouns with SOME* & ANY* and *ONE & *BODY

While chapter 4 looked at the most frequently used personal pronoun, *I,* a decision was made not to look at *you* as many of the SCO informants (and, it can be expected, MAC or BNC/C informants) would use either *you* or the appropriate personal name (e.g. Michael). This would still give valid data for the use of *you* but does restrict the situations where casual conversation is concerned. In compiling SCO, however, I noticed the prominent use of *indefinite pronouns*.

Starting point are the words some and *any*. These describe quantities in casual spoken English. Their vagueness enables simple, informal reference to almost everything that requires a plural: goods as well as people. While the uses of *anything* and *something* will be looked at, the focus is on the reference to other people. Consequently, a second part of the discussion will look at the words *one* and *body*. This discussion aims to be an introduction to the use of *indefinite pronouns* – *someone, anybody, everyone* etc. – discussed in chapter 6.

### 5.1.1 Definitions

There appears to be little or no corpus linguistic research specifically concerned with the usage of *indefinite pronouns.*[96] For this reason, the grammatical compendia provide an initial definition:

> ... a set of words which we may call NONASSERTIVE FORMS: ***anybody, anywhere, yet,*** etc. These in turn contrast with corresponding **ASSERTIVE** FORMS ***(some, somebody, somewhere, already,*** etc) which are associated with positive statements:
>
> [l] Have you found ***any*** mistakes ***yet?***
>
> [2] Yes, I have found ***some already.***
>
> [3] No, ***I*** haven't found ***any yet.***
>
> The contrast between assertiveness and nonassertiveness is basically a logical one: whereas a sentence like [2] asserts the truth of some proposition, the question [l] and the negative statement [3] do not claim the truth of the corresponding positive statement.
>
> (Quirk *et al*. 1985: 83)

Quirk *et al.* also point out that there are situations when *some* appears in a negative use: "Conversely, ***some*** is often used in negative, interrogative, and conditional sentences, when the basic meaning is assertive" (1985: 390).

The issue of *scope* will not be discussed here, as the main focus is on divergent use between SCO and MAC. Indeed, the MACMILLAN dictionary does not mention *scope* while it makes a connection between *any* and *some* use, as its definition of *any* shows:

[96] They seem to be, however, key terms for certain communications / politics / sociology research paper titles.

(usually in negatives and questions) used instead of SOME for saying or asking whether there is a small amount of something or a small number of people and things.

used when it is not important to say which person or thing you are referring to, because what you are saying applies to everyone and everything. (2002: 51)

The dictionary links *any* to negatives and questions. It implies that there is a degree of inter-changeability with *some* and infers that *any* can be found in a context of either persons or things.

### 5.1.2 Corpus-based usage

Looking at the use of *any* and *some* in SCO, BNC/C and MAC, the totals point to the fact that the speakers tend to use *some** rather than *any** utterances:

| MAC word | Freq. | % | SCO word | Freq. | % |
|---|---|---|---|---|---|
| ANY* | 15403 | 0.47 | ANY* | 243 | 0.22 |
| SOME* | 27026 | 0.82 | SOME* | 386 | 0.35 |
| **BNC/C** | **Freq.** | **%** | | | |
| ANY* | 13617 | 0.34 | | | |
| SOME* | 18362 | 0.47 | | | |

Table 1: SOME and ANY frequencies & proportional % in MAC, SCO and BNC/C.

*Some** is the more-frequently used form in all three corpora. In MAC, however, either form is more frequently used then in either SCO or BNC/C, and *some** occurs considerably more often than *any*. Though total occurrence within the entire corpus are lower for BNC/C than MAC, both are higher proportionally than in SCO.

#### 5.1.3.1 ANY clusters comparison

The most frequent two-word clusters are the following:

| MAC cluster | Freq. | % | SCO cluster | Freq. | % |
|---|---|---|---|---|---|
| ANY OF | 1329 | 8.6 | ANYTHING LIKE | 10 | 4.1 |
| HAVE ANY | 1224 | 7.9 | OR ANY | 10 | 4.1 |
| ANY MORE | 917 | 6.0 | DO ANY | 9 | 3.7 |
| ANY OTHER | 847 | 5.7 | GOT ANY | 9 | 3.7 |
| **BNC/C** | **Freq** | **%** | | | |
| ANY MORE | 824 | 6.1 | | | |
| GOT ANY | 675 | 5.0 | | | |
| OR ANYTHING | 395 | 2.9 | | | |
| ANY OF | 380 | 2.7 | | | |

Table 2: Most frequent 2w ANY* clusters in MAC, BNC/C and SCO.

Table 2 above shows that *any* is used in markedly different ways in the three corpora. All the most-frequent clusters in MAC are more frequent proportionally than the top clusters in SCO. MAC and SCO have not a single *any* 2w cluster in common amongst their most frequent clusters. While BNC/C is also different, two of its four top 2w *any* clusters appear also in MAC. While *any of* is proportionally more frequent in MAC, proportional use of *any more* is the same. This underlines that MAC and BNC/C data is more similar while SCO differs more strongly. Indeed, SCO is unique in using *anything like* as its most frequent usage. More on that below.

Most frequent three-word clusters:

| MAC cluster | Freq. | % | SCO cluster | Freq. | % |
|---|---|---|---|---|---|
| ANY OF THE | 449 | 2.9 | ANYTHING LIKE THAT | 9 | 3.7 |
| IS THERE ANY | 389 | 2.5 | OR ANYTHING LIKE | 5 | 2.0 |
| ANY KIND OF | 262 | 1.7 | YOU GOT ANY | 4 | 1.6 |
| YOU HAVE ANY | 260 | 1.7 | ANYTHING YOU WANT | 4 | 1.6 |

| BNC/C cluster | Freq. | % |
|---|---|---|
| HAVEN'T GOT ANY[97] | 245 | 1.8 |
| YOU GOT ANY | 116 | 0.9 |
| YOU WANT ANY | 99 | 0.7 |
| ANYTHING LIKE THAT | 89 | 0.7 |

Table 3: Most frequent 3w ANY* clusters in MAC, SCO and BNC/C

The three-word clusters show, in contrast to the 2w clusters, one phrase that can be seen as equivalent across the corpora. The question *is there any* (MAC) seems to have its equivalent in the more personal *you got any* (BNC/C and SCO). This appears a clear example of pragmatic association where one speech community uses a different wording for the same speech act as the other.

It is remarkable that *anything* is by far the prevalent usage of ANY* in SCO.

| cluster | SCO Freq. | SCO % | MAC Freq. | MAC % | Log-Likelihood | BNC/C Freq. | BNC/C % |
|---|---|---|---|---|---|---|---|
| ANYTHING LIKE THAT | 9 | 3.7 | 12 | >0.1 | 46.66 | 89 | 1.8 |
| OR ANYTHING LIKE | 5 | 2.0 | 10 | >0.1 | 22.87 | 50 | 0.3 |
| ANYTHING YOU WANT | 4 | 1.6 | 3 | >0.1 | n/a | 12 | >0.1 |

Table 4: Most divergent *anything* 3w SCO clusters compared to MAC occurrences.

As Table 4 shows, the most frequent SCO 3w cluster, *anything like that* appears proportionally about twice as frequent in SCO as in BNC/C. All

[97] Appears 177 times (1.2%) in MAC.

three frequent *anything* 3w clusters in SCO are rare in MAC and the Log-Likelihood test shows that the difference is, indeed, significant. This makes the use of *anything* in 3w clusters marginal in MAC and BNC/C compared to the prominent use found in SCO. These findings hint at the fact that *any* * appears to be used in very different contexts in SCO when compared to MAC.

#### 5.1.3.2 SOME clusters comparison

We turn now to *some* and begin by focussing on three-word (3w) *some* clusters:

| MAC cluster | Freq. | % | SCO cluster | Freq. | % |
|---|---|---|---|---|---|
| SOME OF THE | 2774 | 10.3 | SOMETHING LIKE THAT | 11 | 2.8 |
| SOMETHING LIKE THAT | 791 | 2.9 | SOME OF THE | 6 | 1.6 |
| SOME OF THEM | 788 | 2.9 | OR SOMETHING LIKE[98] | 6 | 1.6 |
| SOME OF THESE | 628 | 2.3 | YOU SAY SOMETHING | 3 | >1.0 |
| **BNC/C cluster** | **Freq.** | **%** | | | |
| SOMETHING LIKE THAT | 478 | 2.6 | | | |
| SOME OF THEM | 281 | 1.5 | | | |
| SOME OF THE | 278 | 1.5 | | | |
| YOU WANT SOME | 205 | 1.1 | | | |

Table 5: Most frequent three-word clusters of *some* in MAC, SCO and BNC/C

While a comparison between Tables 3 and 5 shows that *any* and *some* appear in some uses that are similar in their respective corpora (i.e. *any of the* and *some of the*) we also find that, unlike *any*, *some* appears to be

[98] Appears 195 times (1.0%) in BNC/C; 378 times (1.4%) in MAC

predominantly used for statements. None of the top clusters in any of the corpora indicates use within a question. Table 5 shows clearly that we find *something like that* as the one 3w *some* cluster that is used with proportionally the same frequency across all three corpora.

| Cluster | SCO Freq. | % | MAC Freq. | % | LL* |
|---|---|---|---|---|---|
| SOMETHING LIKE THAT | 11 | 2.8 | 791 | 2.9 | 0.01 |
| SOME OF THE | 6 | 1.6 | 2774 | 10.3 | 44.19 |
| OR SOMETHING LIKE | 6 | 1.6 | 378 | 1.4 | 0.06 |

Table 6: Most frequent SCO *some** 3w clusters and MAC equivalents compared.
(*LL based on occurrences: total number of **some* in the corpora)

Looking at the most frequent **some* 3w clusters in SCO, we find that *or something like* is used with about the same proportional frequency in both SCO and MAC. Conducting the statistical test we find that *some of the* would be expected to be more frequent to be not divergent from the occurrence pattern in MAC. The 2w cluster *some of* in MAC usually appears as part of a larger cluster, describing a part (or portion) of a larger group of objects or people. This is beyond doubt the predominant use of *some* in MAC. This pattern is mirrored in BNC/C.

### 5.1.4 SOME and ANY conclusions

Though *some* is more frequent than *any* in both corpora, the two terms appear to be used for very similar utterances. We find *some of the* is the top 3w cluster in MAC and so is *any of the*. Other high frequency clusters are similar to these. However, when the proportional frequencies

of the respective clusters are directly compared, inter-changeability of *some* with *any* is called into question.

Given the clusters *have you got any* / *you got any* the preference of *any* for negative statements / questions and *some* for positive statements that MACMILLAN describes is confirmed.

While SCO, with one exception (the underuse of *some of the* in SCO compared to MAC), is in agreement with both MAC and BNC/C when it comes to *some** clusters, *anything like that* and *or anything like* can be found to be significantly more (proportionally) often in SCO.

## 5.2 Uses of *ONE & *BODY

### *5.2.1 Corpus-based usage*

5.1 found strongest use for *anything / something* in SCO. As my main interest lies in how a third party, another person, is being referred to, in 5.2 the focus will be on **one* and **body*. The WordSmith software provides a wildcard option. This means that parts of words can be requested and all endings (or beginnings) of these words will appear as well. It gives an insight into how frequent words and their combinations are – and if these combinations are at all common.

| item | SCO Freq. | SCO % | MAC Freq. | MAC % | Log-Likelihood | BNC/C Freq. | BNC/C % |
|---|---|---|---|---|---|---|---|
| *BODY | 91 | 0.09 | 6890 | 0.21 | 128.81[99] | 4622 | 0.11 |
| Some- | 41 | 45.00 | 2366 | 34.30 | 2.73 | 1592 | 34.40 |
| Every- | 20 | 22.00 | 1704 | 24.70 | 0.29 | 870 | 18.82 |
| Any- | 9 | 10.00 | 1418 | 20.60 | 6.20 | 806 | 17.44 |
| No- | 18 | 20.00 | 909 | 13.20 | 2.55 | 552 | 11.94 |

Table 7: Proportional freq. of *BODY as part of the total corpus and the relative freq. of compounds with –body

As Table 7 above shows, **body* is not particular frequent in either corpus. Significantly, **body* is more than twice as frequent proportionally in MAC than in SCO or BNC/C however. Relative BNC/C figures are, apart from *everyone*, closer to MAC than SCO. The one remarkable figure is the inverted use of anybody and *nobody*. *Nobody*, on the other hand, is markedly more frequent in its proportional use in SCO than in the comparators. Not one of these differences is, however of statistical significance as the test shows.

The picture is different when **one* is focussed on:

| item | SCO Freq. | SCO % | MAC Freq. | MAC % | Log-Likelihood | BNC/C Freq. | BNC/C % |
|---|---|---|---|---|---|---|---|
| *ONE | 622 | 0.59 | 44,836 | 1.36 | 781.08[95] | 38,086 | 0.95 |
| Some- | 34 | 5.50 | 1068 | 2.40 | 17.78 | 727 | 1.93 |
| Every- | 35 | 5.60 | 745 | 1.66 | 35.26 | 438 | 1.16 |
| Any- | 28 | 4.50 | 643 | 1.40 | 25.34 | 392 | 1.04 |
| No- | 28 | 4.50 | 455 | 1.00 | 39.05 | n/a | n/a |

Table 8 Proportional freq. of *ONE as part of the total corpus and the relative freq. of compounds with –one

Table 8 shows that **one* is far more frequently used than **body* throughout the whole of all corpora. **One* is proportionally nearly twice as frequent in BNC/C than SCO and even more frequent in MAC. While

[99] Log-Likelihood figure for **body* is based on frequency in relation to total size of the respective corpus

*one itself is found to be statistically occurring far less often in SCO as would be expected in comparison to MAC, the points of comparison here – *someone, everyone, anyone* and *no-one*[100], appear statistically far more frequently in SCO than could be expected. We also find that total numbers for **one* (apart from *everyone*) are higher than for **body* equivalents in SCO.

While in the MAC (and BNC/C) we find again the definitive preference for *someone* ahead of all other **one* alternatives, in SCO corpus, *someone* is less frequent than *everyone* and not much more frequent than *anyone*.

#### 5.2.3.1 Clusters with ONE

##### 5.2.3.1.1 ONE most frequent clusters

Looking at cluster–usage of **one*, it appears in both corpora quite frequently as a reference to other things (or people).

As Table 9 below shows that, overall, SCO **one* 2w clusters appear in frequencies similar to those found in the comparators. The log-likelihood test shows that of these 2w clusters *that one* is significant at $p < 0.01$ (the value is higher than 6.63) - in other words, *that one* appears proportionally less often than expected in SCO, while it meets the expected number of occurrences in MAC.

[100] Log-likelihood with reference to the total number of occurrences of **one* in SCO and MAC.

| *ONE 2w cluster | SCO Freq. | SCO % | MAC Freq. | MAC % | Log-Likelihood | BNC/C Freq. | % |
|---|---|---|---|---|---|---|---|
| ONE OF | 60 | 10.8 | 4,317 | 19.8 | 0.00 | 2,787 | 7.3 |
| THAT ONE | 26 | 8.0 | 3,041 | 6.9 | 7.12 | 3,028 | 8.0 |
| THE ONE | 20 | 6.5 | 2,188 | 5.0 | 3.97 | 1,808 | 4.7 |

Table 9: Proportional freq. of *ONE as part of the total corpus and the relative freq. of compounds with –one

| Cluster | SCO Frq. | % | MAC Frq. | % | Log-Likelihood | Frq. BNC/C | % |
|---|---|---|---|---|---|---|---|
| ONE OF THEM | 16 | 2.6 | 430 | 0.96 | 11.27 | 604 | 1.6 |
| ONE OF THE | 15 | 2.4 | 2146 | 4.9 | 8.88 | 671 | 1.8 |
| ONE OF THOSE | 8 | 1.3 | 436 | 0.94 | 0.56 | 468 | 1.2 |
| ONE OF THESE | 8 | 1.3 | 355 | 0.8 | 1.59 | 329 | 0.9 |

Table 10: Proportional use of the most frequent ONE – clusters in SCO, MAC and BNC/C.

Table 10 shows that, amongst 3w *one* clusters *one of them* is significant at $p < 0.001$ (the value is higher than 10.83). Again, this cluster occurs far less often in SCO as is expected, while the expected value is matched in MAC.

### 5.2.4 Clusters with *BODY

As we have seen, *BODY mostly appears in the use of *somebody* or *anybody*. Table10 shows the most frequent clusters with the target terms. These proportional percentages must be interpreted with caution however as the total numbers of these clusters are very low in SCO.

| Cluster | Frq. SCO | % | Frq. MAC | % | Log-Likelihood | Frq. BNC/C | % |
|---|---|---|---|---|---|---|---|
| SOMEBODY ELSE | 7 | 6.8 | 342 | 5.0 | 1.15 | 191 | 4.1 |
| TO SOMEBODY | 5 | 4.5 | 107 | 1.6 | 5.34 | 74 | 1.6 |
| AND SOMEBODY | 3 | 3.4 | 129 | 1.9 | N/A | 65 | 1.4 |
| SOMEBODY WAS | 3 | 3.4 | 38 | 0.6 | N/A | 25 | 0.5 |
| EVERYBODY ELSE | 0 | 0 | 144 | 2.1 | N/A | 100 | 2.2 |
| ANYBODY ELSE | 2 | 2.3 | 198 | 2.9 | N/A | 95 | 2.0 |

Table 11: 2w clusters with **body* compared in three corpora (LL for SCO:MAC).

First of all, **body* 2w clusters (Table 11) show a high degree of similarity in the proportional use of MAC and BNC/C. Second, all three corpora have *somebody else* as their most frequent 2w cluster, with roughly similar proportional frequencies within **body* use. Third, *everybody else* and *anybody else* are used with similar frequencies in MAC and BNC/C – and these two clusters are half as frequent as *somebody else*. Looking at the *log-likelihood test*, we find that *somebody 2w* clusters cannot safely be assumed to occur with different proportional frequencies of a significant kind either. *To somebody* is only significant around the 95% level - not enough, given that there is only the minimum of 5 occurrences of this cluster. There is only a pointer that *somebody was* is used proportionally more often in SCO (3.4% of all uses of *some**) than in either MAC or BNC/C but the occurrence total is too low to make any safe claims.

## 5.3 Conclusions:

## The uses of SOME- & ANY- and -ONE & -BODY

There are a number of important conclusions to be drawn from the above findings.

As for the use of the target terms, the above research indicates divergent use between the comparators and SCO. **one* and **body* are proportionally more frequent terms in MAC (and BNC/C) than SCO. However, when looking at *some-, every- , any-* and *no -one* use, these are all more frequently used proportionally in SCO corpus. This has to be seen, however, within the framework of low total numbers in SCO.

*Some-, every- , any-* and *no –body* usage appears to be more complex. Overall, the corpora use the word-combination at similar proportional frequencies. Yet the highest occurring combination, *somebody*, is proportionally more frequent in SCO and the use of *anybody* is proportionally twice as frequent in SCO.

As such, this serves as a qualified introduction to the following sections on the uses of *some-, every- , any-* and *no -one* and *–body*. Trends and divergences have already become apparent indicating that valuable information can be gained from further investigation.

# Chapter 6 Talking about other people in Casual English

## 6.1 Introduction: core words used

This chapter is concerned with how non-present third parties are named and referred to by Scouse speakers. In many conversations, we can identify three parties: the Speaker, the Listener(s) and a non-present third party. While the first two take turns, the third is referred to, but in most cases, is non-present[101]. Dickerson (2000:393) indicates that a level of self-definition is given when contrasting the *"I"* to the *"other"*. Our data are largely concerned with such conversations.

Transcribing SCO, it became apparent that few persons were addressed by name, while it was noticeable that third-party referents were being used by almost all informants in different circumstances. The reason for looking at these referrers is, therefore, less a matter of raw frequency than one of wide dispersal of use.

A third party can be somebody very close – a partner or member of the family. As such, he / she or they can be expected to be referred to by their name. If, however, the speaker refers to somebody that he or she is emotionally more distant to, or a group or class of people, the point of

[101] A clear exception is the question *Anybody here from …?* .

reference may constitute a rather more vague description. The speaker may, for example, make use of *indefinite pronouns.* According to Biber *et al.* -

> .. indefinite pronouns refer to entities which the speaker or writer cannot or does not want to specify more exactly. (Biber *et al.* 2000: 351)

There is also a link to the discourse particles and intensifiers discussed in later chapters. Both these third-party referrers and intensifiers are vague descriptors as Duguid (2009) notes:

> Hyperbole is to some extent vague. (…) it avoids precision. We can see in the next set from the keywords [i.e. *really, bit, stuff, something*] how lack of precision can also be combined with understated vagueness, which has been identified as a strong indication of an assumed shared knowledge and can mark in-group membership (Carter and McCarthy 2006: 202). (Duguid 2009: 16)

It is interesting to see that Duguid brings use of intensifiers and referrers together. The view that this "shared knowledge marks group membership" highlights a characteristic of casual speech and gives further reason to research these items in the context of this thesis.

As there are clear proportional differences of use with regard to how the core terms occur, a closer look at how these terms are defined seems necessary. For this, all the terms have been re-categorised according to the definitions given in *Macmillan English Dictionary.*

Items typically employed for these occasions are:

| item | Macmillan English Dictionary for Advanced Learners definition |
|---|---|
| anyone | (usually in negatives & questions) used in stead of 'someone'. Used when it is not important which person you are referring to. |
| anybody | anyone. |
| everybody | everyone. |
| everyone | every person in a group. Used for talking about people in general. |
| somebody | someone. |
| someone | Used for referring to a person when you do not know or do not say who the person is. |
| nobody | no one. |
| no-one | not any person, nobody. |
| folk(s) | people in general. People of a particular type / place. Folks: (spoken) used for talking to a group of people. |
| people | the plural of person. Used for ref. to humans in general. Men and women who work in the same organisation. |

Table 1: Core terms discussed and their dictionary (Macmillan) definitions

Though everybody is aware that all these terms are being used in appropriate contexts and are different in their meaning and nuance, little research appears to have been published on the specific use of them.

Table 1 shows that the Macmillan dictionary equates *anyone* with *anybody, everyone* with *everybody,* and *someone* with *somebody*. More detailled research will show in how far this is justified.

The aim of this chapter is to explore whether there is any marked difference between the use of these terms in UK Casual Spoken English in MAC[102] when compared to Liverpool Casual Spoken English.

[102] As pointed out in Chapter 2, BNC/C data will be taken into account where there are substantial differences in use between SCO and MAC.

| Word: | Freq. SCO | % | Ratio of use | Freq. MAC | % | Ratio of use | LL |
|---|---|---|---|---|---|---|---|
| **a. ANYBODY** | 8 | 0.007 | a:b~1:2 | 1,919 | 0.058 | a:b ~ 2:1 | 86.05 |
| **b. ANYONE** | 17 | 0.015 | | 1,045 | 0.032 | | 13.93 |
| ***c.* EVERYBODY** | 19 | 0.017 | c:d~1:2 | 2656 | 0.080 | c:d ~ 2:1 | 90.02 |
| **d. EVERYONE** | 35 | 0.032 | | 1391 | 0.042 | | 4.99 |
| **FOLKS** | 3 | 0.003 | | 602 | 0.018 | | n/a |
| **NOBODY** | 18 | 0.016 | | 1050 | 0.032 | | 23.56 |
| **NO-ONE** | 12 | 0.01 | | 250 (MAC:MED) | 0.010 | | 28.57 |
| **PEOPLE** | 226 | 0.205 | | 25816 | 0.782 | | 752.16 |
| **e. SOMEBODY** | 41 | 0.037 | e:f~1.2:1 | 3200 | 0.096 | e:f ~1.5:1 | 62.35 |
| **f. SOMEONE** | 34 | 0.031 | | 2133 | 0.065 | | 29.54 |
| ***TOTAL*** | | 0.364 | | | 1.18 | SCO:MAC ~1:3 | |

Table 2(a): Comparison of the proportional frequencies of occurrence of 3rd party referral core terms in SCO and MAC. The *ratio of use* compares the ratio of core terms.

### 6.1.1 Proportional distribution of usage

Looking at Table 1 in conjunction with Table 2, we find how the pronouns and descriptions of a third party are clearly divided. Both SCO and MAC use the "pronoun with negatives / question; person not important" less than the "pronoun used to refer to person not known or not named". The most frequent term, *people*, is general and all-inclusive. At the same time *folk*, though technically describing a group of people, has a very specific (and rare) use.

Despite the differences in proportional frequencies found for *everybody* and *everybody*, neither term is used in SCO in a way notably different from that in MAC. For this reason, *everyone/ everybody* is not discussed in further detail.

Used least of all are references to *nobody* and *no-one*.

On the whole, items listed in Table 1 appear proportionally more frequently in MAC than in SCO. The combined occurrences in SCO of the items listed in Table 1 constitute 0.37% of all the words of the corpus. In MAC, the figure is 1.12% of all the words. The ratio of use is 1:3 when SCO is compared to MAC. This mirrors results shown in chapter 5.

Table 2 highlights that the differences of use between SCO corpus and MAC corpus are three-dimensional. Apart from the higher proportional frequencies in MAC compared to SCO, we also find that *anyone* appears nearly twice as often as *anybody* in SCO, while the

reverse is found in MAC. Likewise, *everyone* is nearly twice as frequent as *everybody* in SCO; but the reverse proportions are found in MAC. However, in both corpora *somebody* is used slightly more often than *someone*.

*People* is by a fair margin the most frequently used term to refer to third parties in both corpora. There are other terms, however, where frequency of use in both the Liverpool and the Macmillan corpora is negligible: *folks* and *no-one*. Relatively rare in SCO only are *anyone, anybody,* and *nobody*. Clusters occuring with these items are too few to provide any insights that can be validated. Consequently, only the most marked divergencies found in comparison will be discussed.

The most-used terms are *people*, *somebody* and *someone*. *People* (SCO:0.205%¦ MAC: 0.782%) is used nearly four times as often in MAC as it is in SCO; and *somebody* (SCO:0.037%¦ MAC: 0.096%) is used nearly three times as often in MAC as it is in SCO.

Quirk *et al.* (1985) discuss compound pronouns and provide details from the LOB & Brown corpora:

[a] The frequencies of compound pronouns with *any-*, *every-*. and *some-* that have personal reference are as follows in the LOB and Brown corpora of printed BrE and AmE. respectively:

*Table* 6.46b Frequencies of compound pronouns with *any-*, *every-*, and *some-*

| item | BrE | AmE |
|---|---|---|
| anybody | 32 | 42 |
| anyone | 141 | 140 |
| everybody | 33 | 72 |
| everyone | 106 | 94 |
| somebody | 27 | 57 |
| someone | 117 | 94 |

The table shows that in both corpora, the compounds in *-one* are consistently more frequent than the corresponding compounds in *-body;* but also that compounds in *-body* are more frequent, and compounds in *-one* are less frequent, in AmE than in BrE.

(Quirk *et al.* 1985: 378)

Comparing the frequency-list of Quirk *et al.* with Table 1 it must be noted that the tendencies in LOB are closer to SCO than to MAC for *anybody /anyone* and *everybody /everyone* because the *–one form* is more used in both LOB and SCO. However, the proportional relations of *–one* to *–body* forms in LOB seem to be very different from those of SCO and MAC. Given that LOB (and Brown) are dated corpora by now[103], this can be seen as evidence of diachronic change. Given that *–one* forms appear to be more old-fashioned, stronger use of *–one* forms in SCO can be seen as an indication of a more conservative use of these words. This is confirmed when comparison is made with Biber *et al.* (2000: 352) where the relative use of *–one* to *–body* forms (by total frequency) are much closer[104].

[103] Both LOB and Brown corpora go back to 1961.
[104] I do not provide their table here, as they do not give numbers, just bar-charts. They show that *somebody* is used more often than *someone* (in line with SCO and MAC results) and no-one is rare and less used than nobody (ditto).

This chapter tries to find whether there is a marked difference between SCO use and MAC use as regards all these terms.

## 6.2 NOBODY

### *6.2.1 NOBODY in concordance*

The item *nobody* appears only 15 times in SCO which throws up considerable problems for any serious comparison, so I will here just focus on the most striking differences between the corpora. Not one 2w cluster in SCO has the sufficient number of occurences to undertake reliable statistical tests. All that is being described has to be seen as a tendency found in SCO in comparison to the comparators.

The most frequent 2w cluster in SCO – *nobody has* – appears 3 out 15 times (20.0%). In MAC, it is proportionally far less used (37 times – 3.5%). In the 658 occurrences of *nobody* in BNC/C, it is even rarer. It appears only 10 times (1.5% of all uses of *nobody).* The most frequent 2w and 3w clusters found in MAC are, with the exception of *nobody nobody,* also amongst the most frequent clusters in BNC/C. This finding is compounded by the fact that not one of the top 10 2w MAC *nobody* clusters appears in SCO, neither does any of the top 15 3w MAC clusters appear in SCO. Therefore, despite the low numbers, it has to be noted that the most-used two-word clusters in SCO corpus are rather infrequently used in MAC corpus.

While this is an instance of collocational difference, the colligational differences between SCO and MAC are also striking. The total of 15 concordance lines of *nobody* in SCO above can be analysed as follows to represent their colligational structure:

| | |
|---|---|
| NOBODY followed by a verb: 13 | (87%) |
| NOBODY followed by an adverb: 2 (hardly;ever) | (13%) |
| NOBODY followed by a present tense verb: 6 | (33%) |
| NOBODY followed by a past modal verb: 4 | (26%) |
| NOBODY followed by a future modal verb: 2 (will) | (13%) |
| NOBODY appears as *subject* of a sentence uttered in 15/15 times | (100%) |

It must be noted that *nobody* in SCO always appears as the subject of an utterance. The verbs either directly follow nobody or appear after *nobody* + auxiliary verb. In SCO, *nobody* is followed by an equal number of present tense and past-tense verbs. In MAC, however, present-tense verbs follow *nobody* in the majority of cases. Furthermore, every single SCO concordance line has *nobody* as clause subject.

There are indications, however, that although MAC uses *nobody* as a subject in a clause, it also uses it as an object, for example, *you know nobody,* and both as subject and object in *nobody knows nobody*. This is also shown utterances containing in the most frequent MAC *nobody* 2w cluster, *and nobody*. At the beginning of an utterance, this tends to be the subject: *And nobody ever discussed him*. It can also be found, in the middle of an utterance, as an object: *Sarah and nobody else!*

## 6.3 ANYBODY and ANYONE

Though the difference of meaning between anybody and *anyone* appears to be fractional, the proportional frequency of use is different. Consequently, one word can be seen to be primed to appear more in one environment (context) than the other. The following discussion will highlight different uses of the two core terms and explore whether the core terms are also employed differently in different corpora.

| | ANYBODY | ANYONE | Ratio of use |
|---|---|---|---|
| SCO Freq. | 8 | 16 | |
| SCO use per 100k words | 7.5 | 13 | ~1:2 |
| MAC Freq. | 1,919 | 1,045 | |
| MAC use per 100k words | 58.2 | 32.1 | ~2:1 |
| Log-Likelihood | 86.05 | 13.93 | |

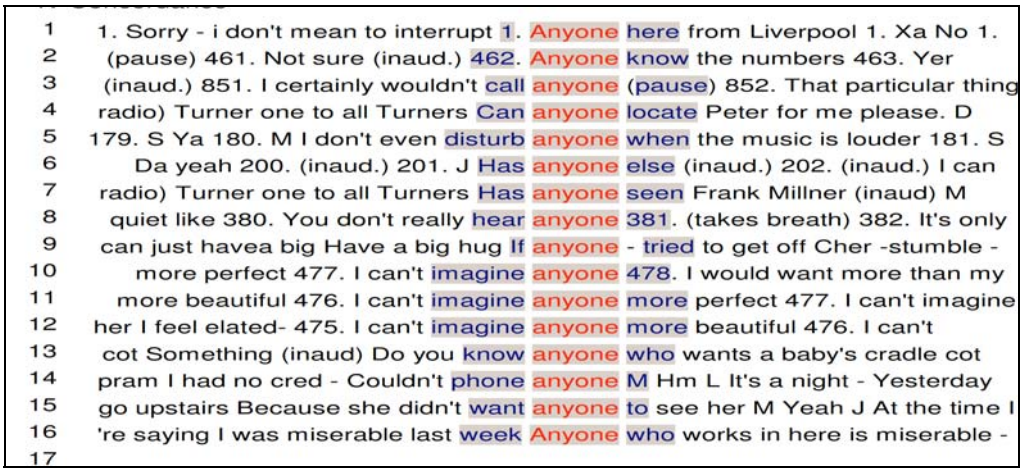
Table 2(b): *Anybody* and *Anyone* occurrence in SCO and MAC.

Table 2(b) shows that *anyone* occurs proportionally twice as often than *anybody* in SCO. This is inverse to the pattern of use found in MAC, where *anybody* is proportionally more frequent[105]. This already indicates a divergent use of the term. Furthermore, *anybody* is the more significant item, as it is proportionally nearly 8 times more frequent in MAC than in SCO; *anyone* is proportionally about 2.5 times more frequent in MAC than SCO. However, having only 8 occurrences for *anybody* means that there is not enough material for a conclusive comparison available for this word.

[105] This is also true for BNC/C, where *anybody* occurs 910 while *anyone* occurs only 441 times.

### 6.3.1 ANYONE

Given that there are only 16 concordance lines in SCO, I can compare these line by line with the total usage in MAC.



| | |
|---|---|
| 1 | 1. Sorry - i don't mean to interrupt 1. Anyone here from Liverpool 1. Xa No 1. |
| 2 | (pause) 461. Not sure (inaud.) 462. Anyone know the numbers 463. Yer |
| 3 | (inaud.) 851. I certainly wouldn't call anyone (pause) 852. That particular thing |
| 4 | radio) Turner one to all Turners Can anyone locate Peter for me please. D |
| 5 | 179. S Ya 180. M I don't even disturb anyone when the music is louder 181. S |
| 6 | Da yeah 200. (inaud.) 201. J Has anyone else (inaud.) 202. (inaud.) I can |
| 7 | radio) Turner one to all Turners Has anyone seen Frank Millner (inaud) M |
| 8 | quiet like 380. You don't really hear anyone 381. (takes breath) 382. It's only |
| 9 | can just havea big Have a big hug If anyone - tried to get off Cher -stumble - |
| 10 | more perfect 477. I can't imagine anyone 478. I would want more than my |
| 11 | more beautiful 476. I can't imagine anyone more perfect 477. I can't imagine |
| 12 | her I feel elated- 475. I can't imagine anyone more beautiful 476. I can't |
| 13 | cot Something (inaud) Do you know anyone who wants a baby's cradle cot |
| 14 | pram I had no cred - Couldn't phone anyone M Hm L It's a night - Yesterday |
| 15 | go upstairs Because she didn't want anyone to see her M Yeah J At the time I |
| 16 | 're saying I was miserable last week Anyone who works in here is miserable - |
| 17 | |

Concordance 1 : All ANYONE concordance lines in SCO.

Concordance 1 demonstrates four striking features:

(1) There is the repeat cluster I *can't imagine anyone* which appears 3 times (18%) [pragmatic association / collocation]. Given that this is uttered by a single person, it has to be discounted.

(2) The structure *negative (+ Aux.) + Verb + anyone* appears eight out of sixteen times. (50%) [colligation]. The difference found in comparison to the use in MAC is marginal.

(3) Six uses of *anyone* are a question (37%) [colligation].

(4) Only in three cases (16%) does *anyone* start a turn (lines 1, 2 & 6). All three are identified to be questions [colligation / semantic association].

However, as the total figures for SCO are extremely low (which means that a single extra or a single few occurrence could change

percentages considerably) I will just focus on the *anyone* clusters that are found to be extremely marginal in MAC:

| | SCO clause | Occ. | MAC clause | Occ. |
|---|---|---|---|---|
| 1 | **Anyone here** from Liverpool? | 1 | **Anyone here** support Arsenal? | 1 (0.1%) |
| 2 | **Anyone know** the numbers? | 1 | *Does **anyone know** what?* | 3 (0.3%) |
| 3 | **Anyone who work**s here ... ? | 1 | **Anyone who**<br>**Anyone who work**ed at the UN | 13 (1.2%)<br>1 (0.1%) |
| 4 | ***Has anyone** else* | 2 | ***Has anyone** been on telly* | 44 (4.1%) |

Table 3: *Anyone* and *anyone questions* at the start of a turn in SCO and MAC.

*Anyone* occurrences shows the use of ellipsis among Scouse speakers. While both SCO and MAC also use the question format *Auxiliary Verb +anyone + Lexical Verb.* (*Has anyone seen…* ), clusters like *anyone here* and *anyone who* with *work** appear in both corpora with the same total occurrence numbers -when MAC is 27.5 times as large as SCO and higher occurrences can be expected.

## 6.4 *SOMEBODY* and *SOMEONE*

*Somebody* and *someone* are far more frequent in both SCO and MAC than *anybody* and *anyone*. The proportional differential in which these two terms are used is, however, much smaller.

As Table 2(c) below shows, in SCO and MAC the proportional frequency of use of *somebody*, when compared to *someone*, stand in roughly the same relation to each other[106].

| | SOMEBODY | SOMEONE | Ratio of use |
|---|---|---|---|
| **SCO Freq.** | 41 | 34 | |
| **SCO use per 100k words** | 37 | 31 | ~1.2:1 |
| **MAC Freq.** | 3200 | 2133 | |
| **MAC use per 100k words** | 96 | 65 | ~1.5:1 |
| **Log-Likelihood** | 62.35 | 21.54 | |

Table 2(c) : SOMEBODY and SOMEONE occurrence frequencies and relation in SCO and MAC.

The statistical test shows that both *somebody* and *someone* are used significantly less frequently in SCO than in MAC. In both corpora, however, the ratio of use between *somebody* and *someone* is fairly close. There can be found, however, a difference in collocates between *somebody* and *someone*. This difference is more pronounced in SCO than MAC.

[106] In BNC/C the difference is stronger than in MAC. In BNC/C, there are 1717 occurrences of *somebody* compared to 778 occurrences of *someone*. The re44lation being 2.2:1.

Table 4: SOMEBODY and SOMEONE collocates distribution in SCO and MAC.

| SCO S'body | Freq. | % | SCO S'one | Freq. | % | MAC S'body | Freq. | % | MAC S'one | Freq. | % |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TO | 8 | 20 | TO | 11 | 34 | TO | 961 | 30 | TO | 764 | 36 |
| ELSE | 6 | 15 | THAT | 8 | 24 | AND | 786 | 25 | THE | 495 | 23 |
| I | 6 | 15 | YOU | 7 | 21 | YOU | 702 | 22 | WHO | 477 | 23 |
| IT | 6 | 15 | I | 6 | 18 | THE | 697 | 21 | A | 470 | 22 |
| OR | 6 | 15 | OR | 6 | 18 | A | 587 | 18 | YOU | 461 | 22 |
| THE | 6 | 15 | IF | 5 | 15 | THAT | 563 | 18 | AND | 442 | 21 |
| WAS | 6 | 15 | WHO | 5 | 15 | I | 502 | 15.7 | THAT | 433 | 21 |
| AND | 5 | 12 | | | | IT | 484 | 15 | I | 314 | 15 |
| ON | 5 | 12 | | | | IN | 408 | 12.75 | IS | 292 | 13.7 |
| | | | | | | WHO | 388 | 12.5 | IN | 276 | 13 |
| | | | | | | ELSE | 387 | 12.5 | IT | 271 | 12.9 |
| | | | | | | OF | 330 | 10 | OF | 256 | 12 |
| | | | | | | IF | 307 | 10 | IF | 249 | 12 |
| | | | | | | IS | 289 | 9.8 | FOR | 197 | 9 |
| | | | | | | ETH | 254 | | WITH | 184 | 8.6 |
| | | | | | | THEY | 245 | 7.7 | ELSE | 171 | 8 |
| | | | | | | OR | 240 | 7.5 | OR | 160 | 7.5 |
| | | | | | | WAS | 237 | 7.4 | BE | 156 | 7.4 |
| | | | | | | ON | 229 | 7.2 | WAS | 146 | 6.8 |
| | | | | | | FOR | 228 | 7.2 | ON | 145 | 6.8 |

Table 4: *Somebody (S'body)* and *Someone (S'one)* most frequent collocates in SCO and MAC.

### 6.4.1 SOMEONE

When looking at the ways both *somebody* and *someone* cluster, it becomes apparent that their uses are highly restricted. Neither in MAC nor in SCO can we find any significant amount of three-word clusters. Both terms tend to collocate mostly with a fixed set of other words. This is especially true of the less frequent of the two, *someone*.

| SOMEONE collocate (1-5ws) | SCO Frq. | % | MAC Frq. | % |
|---|---|---|---|---|
| TO | 11 | 34 | 764 | 36 |
| THAT | 8 | 24 | 433 | 21 |
| YOU | 7 | 21 | 461 | 22 |
| I | 6 | 18 | 314 | 15 |

Table 4(b): Top 4 collocates of SOMEONE in SCO and MAC.

Table 4(b) shows how close even the proportional usage in terms of collocates is. *Someone* tends to prefer the company of the same most frequent terms both MAC and SCO.

Likewise, the proportional frequencies of all the most-used clusters for the term *someone* are near-identical. This is demonstrated in Table 5:

| cluster | SCO Frq. | % | MAC Frq. | % | LL |
|---|---|---|---|---|---|
| SOMEONE WHO | 11 | 15 | 764 | 18.5 | 0.12 |
| SOMEONE ELSE | 8 | 9 | 433 | 8.6 | 0.16 |
| SOMEONE'S | 7 | 9 | 461 | 8 | 0.02 |
| IF SOMEONE | 6 | 9 | 314 | 7.4 | 0.18 |

Table 5: *Someone* 2w cluster frequency (as part of the total of *someone* occ.) in SCO and MAC.

This leads to the conclusion that *someone* is one of those words where the usage is pretty much identical between SCO speakers and speakers from across the UK represented by MAC.

### 6.4.2 SOMEBODY

Proportionally, the use of *somebody* is more prominent in MAC compared to SCO. It is even more prominent than the use of someone. While the latter is twice as frequent in MAC corpus, *somebody* is proportionally 2.6 times as frequent (0.037% compared to 0.096%).

As with *someone*, we find that *somebody* is highly restricted in its use. In neither corpus can any meaningful number of three-word clusters be found. The only clusters with a relevant amount of repetition (frequency) are the two-word clusters. Amongst these, by far the most frequent for both corpora is the cluster *somebody else.*

*Someone* usage serves as an example of how a word might be employed in almost identical ways in the two speech communities under comparison. A good example for this is the repeat-use of *somebody* in a clause. Accordingly, we find in SCO –

> But it's always L8[107] like (...) if someone's mugs somebody or robs somebody.

Which has got exactly one equivalent in MAC:

[107] Liverpool L8 – Toxteth, which has a certain notoriety.

> You can murder somebody or rape somebody, you're still eligible, for me in that sense.

This particular example has to be appraised with care. A single occurrence has no relevance. On the other hand, the obvious parallels of use – the conditional clause used to describe a criminal act committed on *somebody* - may indicate that this is a very specific, though rare, employment of *somebody with or*.

| cluster | SCO Frq. | % | MAC Frq. | % | LOG-LIKELIHOOD |
|---|---|---|---|---|---|
| SOMEBODY ELSE | 7 | 17.0 | 431 | 13.5 | 0.36 |
| TO SOMEBODY | 4 | 9.8 | 112 | 4.7 | |
| AND SOMEBODY | 3 | 7.3 | 133 | 5.2 | |
| SOMEBODY WAS | 3 | 7.3 | 41 | 1.4 | |
| SOMEBODY WHO | 3 | 7.3 | 147 | 5.8 | |
| IF SOMEBODY | 3 | 7.3 | 135 | 5.3 | |

Table 6: *Somebody* 2w cluster proportional frequency (relative to total no. of *somebody* occurrences) for SCO and MAC

When we look at the most frequently occurring *somebody* 2w clusters, as in Table 6, it is clearly shown that *somebody* occurrence pattern in short clusters, similar to the use of *someone*, is almost identical in the two corpora.

### 6.4.3 Conclusions & Comparison: SOMEONE and SOMEBODY

On the whole it can be said that that the prefix *some-* , when used to refer to a third party, tends to bring out broadly similar use between SCO and MAC speakers. It is not simply the most common clusters that occur

with broadly the same proportionally frequency. All the other clusters of significant use also occur proportionally as often.

That there is such a level of concurrence is actually a good sign. If I had found difference of use for every single term the research focuses upon, it could put my own corpus into question. Finding only difference, after all, would indicate we are looking at a completely different language, or that the SCO corpus is too restricted for adequate comparisons to be drawn.

As it is, the difference between *somebody* and *someone* in MAC and SCO is restricted to the level of usage. Though *somebody* is used more frequently than *someone* in both corpora, it is even more frequent in MAC. More importantly, *someone* appears about twice as often and *somebody* nearly three times as often in MAC as in the SCO corpus. As the data from the LOB corpus indicate, stronger use of –ONE may indicate a more conservative pattern of usage.

## 6.5 PEOPLE usage

### *6.5.1 Introduction and numbers of occurrence*

*People* is a term that, by itself' can be seen to be very broad in its meaning. It fits in with the category of naming a third party – the same as *someone* or *anybody* It needs to be pointed out, though, that *people* is found to be far more frequent than all other referral terms. This section sets to find out how specific a meaning *people* can have given the

immediate context it is found in. Leading on from this, the issue is again to see whether there are uses which mark a clear difference between SCO and MAC speakers.

In this section I will look at how the term *people* collocates with a number of key items. I will examine to what extent *people* is used with different frequency in the two corpora and in what contexts – i.e. in what clusters. This leads on to question of prosody and social preference – this section will try to find out to what extent a socially sensitive term like *people* reflects levels of speakers' attitude through its use.

When looking at the spoken use of *people*, one point needs to be made first and foremost. Compared to all other third-party referrers, *people* is relatively frequent in both corpora. This makes any statement about their comparative uses more relevant, as small differences of use do not affect the overall percentages disproportionally.

| Item | Freq. SCO | % | Freq. MAC | % | Freq. BNC/C | % |
|---|---|---|---|---|---|---|
| PEOPLE | 226 | 0.205 | 25,816 | 0.782 | 4,692 | 0.12 |

Table 7: *People* frequencies of occurrence in SCO, MAC & BNC/C with % of total corpus

The raw figures of Table 7 reveal some contradictory findings. It can be seen that the term *people* is used over 3.7 times more often in MAC than in SCO. *People* is, however, relatively low in its occurrence in BNC/C: it occurs 4,692 times, which is 0.12% of the corpus total – proportionally only half as often as SCO. When the use of *people* clusters is discussed, BNC/C will be employed as a further comparator.

| SCO total | Word/colloc. | % SCO | % MAC | MAC total | Log Likelihood |
|---|---|---|---|---|---|
| 226 | PEOPLE (relative use in corpus) | 0.21 | 0.78 | 25,300 | 752.16 |
| 47 | THE | 20.8 | 41.0 | 10,574 | 27.23 |
| 41 | TO | 18.1 | 30.5 | 7,873 | 13.15 |
| 40 | OF | 17.7 | 27.4 | 7,082 | 8.87 |
| 40 | AND | 17.7 | 25.3 | 6,527 | 5.70 |
| 37 | YOU | 16.4 | 11.0 | 2,848 | 5.03 |
| 33 | A | 14.6 | 15.6 | 4,028 | 0.15 |
| 33 | IN | 14.6 | 15.4 | 3,980 | 0.10 |
| 31 | THAT | 14.2 | 24.4 | 6,290 | 12.42 |
| 28 | I | 12.4 | 6.9 | 1,789 | 7.77 |
| 24 | KNOW | 10.6 | 6.2 | 1,592 | 5.90 |
| 23 | DON'T | 10.3 | 3.2 | 836 | 21.02 |
| 22 | LIKE | 10.0 | 4.1 | 1,046 | 12.73 |
| 21 | YEAH | 9.3 | 1.4 | 374 | 41.78 |
| 19 | IT | 8.4 | 8.2 | 2,128 | 0.01 |
| 18 | JUST | 8.0 | 2.5 | 658 | 16.31 |
| 18 | THEY | 8.0 | 6.3 | 1,619 | 0.94 |
| 16 | THESE | 7.1 | 4.1 | 1053 | 4.04 |
| 15 | WERE | 6.6 | 4.8 | 1,244 | 1.37 |
| 15 | NOT | 6.6 | 4.3 | 1,121 | 2.33 |
| 14 | WHO | 6.2 | 15.1 | 3,890 | 15.12 |

Table 8: SCO and MAC most frequent collocates of PEOPLE (percentages relative to total of core term frequencies.)

### 6.5.2 PEOPLE and its collocates

Table 8 gives the occurrence (in %) of *people* in relation to the total corpus. On the whole, the statistical test reveals that *people* appears far less often than could be expected in SCO. All collocates give percentages in relation to the overall figure for *people*.[108] Table 8 shows that, despite differing percentages of occurrence, both MAC and SCO mostly share the same most-frequently occurring collocates. Indeed, most of the top ten highest occurring collocates are the same.

In Table 8, MAC lists as collocate of *people* only *have* (a collocate in 10.4% of all occurrences of *people*) but not *'ve*. SCO records *have* and *'ve* and the combined use of *have* and *'ve* in SCO is 8.4%.

#### 6.5.2.1 Frequency of collocates

It has already been mentioned that MAC makes greater use of the term *people* than SCO. The most frequent collocate to *people* - *the* - occurs twice as often in MAC. More to the point, with 41% of instances of *people* occurring with this collocation (compared to 20.8% in SCO) *the* occurs nearly every other time as a collocate when *people* is used in MAC, this is statistically significantly fewer occurrences of *people* with *the* in SCO. Similarly, almost all of the top-ten most occurring collocates in

[108] Because of the way MAC has been transcribed, pauses do not show up in this table.

MAC occur about 10 % more often than the same collocates in SCO. This seems to indicate that *people* in MAC is employed in far more fixed expressions and usages – its nesting can be seen to be more restricted. *People* in SCO, however, appears to have the ability to collocate more freely.

#### 6.5.2.2 Where collocates' frequencies differ

Table 9 highlights all those *people* collocates that are found to be statistically significantly different in their frequency of occurrence when SCO is compared with MAC:

| ***People*** **collocate** | **Rank SCO** | **Occ. SCO** | **Occ. MAC** | **Rank MAC** | **LL** |
|---|---|---|---|---|---|
| **THE** | 1 | 47 | 10,574 | 1 | 27.23 |
| **TO** | 3 | 41 | 7,873 | 2 | 13.15 |
| **THAT** | 8 | 31 | 6,290 | 6 | 12.42 |
| ***DON'T*** | 11 | 23 | 836 | 51 | 21.02 |
| ***LIKE*** | 13 | 22 | 1,046 | 41 | 12.73 |
| ***YEAH*** | 14 | 21 | 374 | 93 | 41.78 |
| ***JUST*** | 16 | 18 | 658 | 61 | 16.31 |

Table 9: *People* collocates most divergent between SCO and MAC

According to the log-likelihood test, *the, to* and *that* are found to occur less often than expected, the other collocates listed in Table 9 occur more often than expected in SCO. This mirrors both their proportional frequency of occurrence and ranking. When we look at clusters incorporating *people,* we will see whether these differences in collocations are reflected in differences in cluster use.

### 6.5.3 PEOPLE clusters

The Macmillan Dictionary hints in its entry for *people* at the prominent use of the 2w/3w clusters *people like* and *people like that*.

Based on the research undertaken with SCO and MAC, *people* appears in quite a number of longer clusters. As a result, many 2w and 3w part mostly form part of 4w or even 5w clusters. Consequently, we find that the top three-word cluster is part of the most-occurring 4-word cluster: *lot of people* and *a lot of people* – the highest occurring 3w/4w clusters in both SCO and MAC.

O'Donnell (2009) points out that -

> Adjusted frequency list is a simple index-based method of producing frequency lists where status of clusters/n-grams as 'single choice items' is reflected in frequency of all smaller items. (Summary slide).

This means that we are looking at not just chunks that are found to be highly frequent but also need to focus on in what context these chunks appear, as they might be a constituent part of a larger cluster. O'Donnell points out that, in those cases, focus should be concentrated on the longer clusters. For example, there are 15 occurrences of *of people* in SCO – 8 of which are constituent part of *lot of people*. This exemplifies the extent to which *people* is a term where smaller chunks are often found to be constituent part of longer clusters, making it logical to compare shorter and longer clusters in adjusted frequency tables.

| 2 - 4 word cluster | SCO total | SCO % | MAC % | MAC total | Log Likelihood |
|---|---|---|---|---|---|
| OF PEOPLE | 15 | **6.6** | **6.1** | 2364 | 4.14 |
| PEOPLE IN | 12 | **5.3** | **4.0** | 1194 | 0.01 |
| SOME PEOPLE | 9 | **4.0** | **3.6** | 1021 | 0.23 |
| OTHER PEOPLE | 8 | **3.5** | **3.2** | 881 | 0.13 |
| LOT OF PEOPLE | 8 | **3.5** | **3.2** | 845 | 0.06 |
| A LOT OF PEOPLE | 6 | **2.7** | **3.1** | 791 | 0.62 |
| PEOPLE HAVE | 5 | **2.2** | **3.1** | 821 | 0.74 |
| PEOPLE THAT | 5 | **2.2** | **3.4** | 927 | 1.38 |
| PEOPLE IN THE | 3 | **1.3** | **1.57** | 406 | n/a |

Table 10: Comparison of SCO and MAC 2-4 word PEOPLE clusters with similar proportional frequencies of use

Table 10 shows those 2 – 4w *people* clusters where there is no strong degree of difference between their proportional frequencies. These also are amongst the highest occurring *people* clusters: *of people* ranking as the highest frequency 2w cluster in SCO, the third-highest in MAC. The longer *a lot of people* cluster is the highest frequency cluster in SCO, MAC (and BNC/C).

| *2 - 4 word cluster* | *SCO total* | *SCO %* | *MAC %* | *MAC total* | Ratio with entries normalized to SCO=±1 | LL |
|---|---|---|---|---|---|---|
| WHEN PEOPLE COME IN | 3 | **1.3** | **>0.025** | 1 | (n/a) | |
| WHEN PEOPLE COME | 3 | **1.3** | **0.028** | 7 | 1. \| : \| 0.0215 | |
| NOT MANY PEOPLE | 3 | **1.3** | **0.10** | 25 | 1. \| : \| 0.0769 | |
| PEOPLE JUST | 5 | **2.2** | **0.41** | 106 | 1. \| : \| 0.1863 | 8.54 |
| PEOPLE SAY | 9 | **4.0** | **0.72** | 187 | 1 \| : \| 0.1975 | 15.67 |
| PEOPLE COME | 5 | **2.2** | **0.466** | 118 | 1. \| : \| 0.2118 | 7.70 |
| PEOPLE DON'T | 9 | **4.0** | **1.24** | 304 | 1 \| : \| 0.310 | 9.12 |
| PEOPLE FROM | 7 | **3.1** | **1.10** | 273 | 1. \| : \| 0.3548 | 5.75 |
| HOW MANY PEOPLE | 4 | **1.8** | **0.64** | 164 | 1. \| : \| 0.3555 | |
| PEOPLE LIKE | 7 | **3.1** | **1.4** | 364 | 1. \| : \| 0.4516 | 3.35 |
| PEOPLE WERE | 7 | **3.1** | **1.5** | 418 | 1. \| : \| 0.4838 | 2.37 |
| WHEN PEOPLE | 4 | **1.8** | **0.95** | 240 | 1. \| : \| 0.5277 | |
| THESE PEOPLE | 9 | **4** | **2.15** | 555 | 1 \| : \| 0.5375 | 2.78 |
| MANY PEOPLE | 10 | **4.4** | **2.98** | 769 | 1. \| : \| 0.6772 | 1.36 |
| PEOPLE WHO | 13 | **5.8** | **9.9** | 3378 | 1. \| : \| 1.707 | 11.69 |
| THE PEOPLE | 8 | **3.5** | **9.2** | 2,556 | 1. \| : \| 2.628 | 12.21 |
| PEOPLE ARE | 4 | **1.8** | **4.9** | 1573 | 1. \| : \| 2.722 | |
| PEOPLE WHO ARE | 0 | 0 | **3.2** | 826 | | |
| THE PEOPLE WHO | 0 | 0 | **2.2** | 697 | | |
| OF THE PEOPLE | 0 | 0 | **2.0** | 502 | | |
| OF PEOPLE WHO | 0 | 0 | **1.5** | 395 | | |

Table 11(a): PEOPLE 2-4 w clusters divergent where SCO is compared to MAC.

### 6.5.3.1 PEOPLE divergent use of 2-4-word clusters

While the highest-frequency *people* clusters appear with similar proportional frequencies, it is amongst the medium-high proportional frequencies of *people* clusters that we find differences. Table 11(a) shows those 2w clusters and their proportional frequencies in SCO and MAC that can be found to be constituent part of longer, relatively frequent clusters in both corpora.

### 6.5.3.2 PEOPLE: MAC-dominant use of clusters

Table 11(b) is an extract of Table 11(a). It shows the three 2w *people* clusters that are noticeably more frequent in their proportional use in MAC:

| 2 - 5 word cluster | SCO total | SCO % | MAC % | MAC total | LL |
|---|---|---|---|---|---|
| PEOPLE WHO | 13 | 5.8 | 9.9 | 3378 | 11.69 |
| THE PEOPLE | 8 | 3.5 | 9.2 | 2,556 | 12.21 |
| PEOPLE ARE | 4 | 1.8 | 4.9 | 1573 | |
| PEOPLE WHO ARE | 0 | 0 | 3.2 | 826 | |
| THE PEOPLE WHO | 0 | 0 | 2.2 | 697 | |
| OF THE PEOPLE | 0 | 0 | 2.0 | 502 | |
| THE PEOPLE WHO ARE | 0 | 0 | 0.6 | 167 | |
| OF PEOPLE WHO | 0 | 0 | 1.5 | 395 | |
| OF THE PEOPLE WHO ARE | 0 | 0 | >0.1 | 15 | |

Table 11(b): *People* 2-5w clusters more prominent in MAC

Table 11(b) shows that *people* collocates like *of, the* and *who* which are all statistically more frequently found in MAC than SCO, play a role in a number of 2-3w *people* clusters that show divergence between SCO and MAC. Table 11(b) demonstrates that 2w clusters like *people who* and *the people* can be found of the longer clusters *people who are, the people who, of the people* and *of people who* all of which are medium-low frequency *people* clusters in MAC but are not found at all in SCO.

Table 12 below demonstrates, furthermore, that there is a high degree of convergence between MAC and BNC/C *people clusters.* While there is little difference in proportional occurrence between MAC and BNC/C, this indicates that there is a clear difference between the comparators and SCO however.

| *2 - 4 word cluster* | *MAC total* | *MAC %* | *BNC/C %* | *BNC/C total* |
|---|---|---|---|---|
| WHEN PEOPLE COME IN | 1 | **>0.025** | **n/a** | n/a |
| WHEN PEOPLE COME | 7 | **0.028** | **n/a** | 2 |
| PEOPLE COME | 118 | **0.466** | **0.55** | 26 |
| NOT MANY PEOPLE | 25 | **0.1** | **0.34** | 16 |
| PEOPLE JUST | 106 | **0.41** | **0.5** | 25 |
| PEOPLE SAY | 187 | **0.72** | **0.92** | 43 |
| PEOPLE DON'T | 304 | **1.24** | **1.00** | 47 |
| PEOPLE FROM | 273 | **1.1** | **0.4** | 19 |
| HOW MANY PEOPLE | 164 | **0.64** | **0.90** | 42 |
| PEOPLE LIKE | 364 | **1.4** | **2.95** | 135 |
| PEOPLE WERE | 418 | **1.5** | **1.30** | 60 |
| WHEN PEOPLE | 240 | **0.95** | **0.82** | 38 |
| THESE PEOPLE | 555 | **2.15** | **3.00** | 144 |
| MANY PEOPLE | 769 | **2.98** | **2.96** | 139 |
| PEOPLE WHO | 3378 | **9.9** | **6.80** | 317 |
| THE PEOPLE | 2,556 | **9.2** | **8.20** | 384 |
| PEOPLE ARE | 1573 | **4.9** | **4.90** | 231 |
| PEOPLE WHO ARE | 826 | **3.2** | **1.80** | 81 |
| THE PEOPLE WHO | 697 | **2.2** | **1.77** | 78 |

Table 12: MAC frequency occurrence pattern compared to BNC/C for the *people* clusters that are most divergent between SCO and MAC in proportional frequencies.

### 6.5.3.3 PEOPLE in SCO – dominant use of 2 word clusters

A fair number of 2w clusters are marginal in MAC or BNC/C while they appear to be a preferred choice in SCO use. This section will concentrate on all those 2w *people* clusters that do not form part of a longer, frequent *people* cluster which are clearly divergent in their use when SCO and MAC are compared.

| 2 word cluster | SCO total | SCO % | MAC % | MAC total | LL |
|---|---|---|---|---|---|
| PEOPLE SAY | 9 | **4.0** | **0.72** | 187 | 15.67 |
| PEOPLE DON'T | 9 | **4.0** | **1.24** | 304 | 9.12 |
| PEOPLE JUST | 5 | **2.2** | **0.41** | 106 | 8.54 |
| PEOPLE LIKE | 7 | **3.1** | **1.4** | 364 | 3.35 |
| THESE PEOPLE | 9 | **4** | **2.15** | 555 | 2.78 |
| PEOPLE WERE | 7 | **3.1** | **1.5** | 418 | 2.37 |

Table 11(c): *People* 2w clusters that are proportionally more frequent in SCO.

Table 11 (c) shows that people collocates like *just, say, don't* and *like* which have been seen to be significantly more frequent in SCO compared to MAC appear again to be proportionally more frequent when found in the combinations of 2w clusters. However, only one, *people say* appears with 99.99% level of statistical significance. Although, in its total usage, SCO speakers are proportionally over 4 times more inclined to use *people say* than MAC speakers, the nesting of this cluster seems to be the same in both corpora – as a preference for being part of the 3w cluster *some people say* shows.

*People don't* (divergent at a 99.0 % level of significance) gives little clue in the concordance lines why it should be different in its proportional use where SCO is compared to MAC. In MAC, it is used a number of times as part of the cluster *if people don't* and in SCO two of the eight lines (spoken by different people) use the phrase *PEOPLE DON'T talk like that*. This particular phrase occurs only once in the much larger MAC concordance. It must be noted that the clusters, in SCO, usually start a new passage after a brief pause in speaking This can be seen as the speaker giving a little more thought before he makes the *people don't* statement. This does agree with discourse studies, where negative statements are found to be more circumspect (cf. Cameron: 2001). MAC does not record, where *people don't* is utterance- initial, that many pauses.

*People just* (divergent at a 99.0 % level of significance) is marginal in MAC (0.41% of all uses of *people)* and BNC/C, yet appears 5 times (2.2%) in SCO. In MAC *people just* has a negative semantic association. It is used as an intensifier as in *a lot of PEOPLE JUST tune in for the commercials*. In SCO, this negativity is not to be found. On the contrary, one of the 5 occurrences records *PEOPLE JUST brought whiskey in* – in this way expressing disbelief in something other people do out of kindness.

There is an issue of divergent nesting found when we look at *people like*. However, as no statistical significance is found in the proportional

frequency difference between SCO and MAC for *people like*, this is not discussed here but can be found in Appendix V.

### 6.5.4 PEOPLE divergent in long clusters

There are two *people* cluster groups where we can find a clear divergence of use between SCO and MAC. There is a strong difference in the proportional frequency of *many people* and *when people* – both are far more dominant in their use in SCO compared to MAC as the excerpt of Table 11 below shows:

| 2 word cluster | SCO total | SCO % | MAC % | MAC total | LL |
|---|---|---|---|---|---|
| PEOPLE COME | 5 | **2.2** | **0.466** | 118 | 7.70 |
| MANY PEOPLE | 10 | **4.4** | **2.98** | 769 | 1.36 |
| WHEN PEOPLE | 4 | **1.8** | **0.95** | 240 | |
| WHEN PEOPLE COME | 3 | **1.3** | **0.028** | 7 | |
| NOT MANY PEOPLE | 3 | **1.3** | **0.10** | 25 | |
| HOW MANY PEOPLE | 4 | **1.8** | **0.64** | 164 | |
| WHEN PEOPLE COME IN | 3 | **1.3** | **>0.025** | 1 | |

Table 11(d) *People* long clusters and their component parts where SCO is proportionally more frequent than MAC use.

The following discussion can only be seen as a projection as the low total numbers make statistical testing unreliable. Only *people come* can be said with a 95% reliability to be divergent in SCO when compared to MAC use.

*People come* appears in 3 out of 5 times of the use of *When people come* **and** *when people come in* in SCO, whereas in MAC, *people come* appears only in 1 out of 17 times of *when people come* and *when people*

*come in* is exceedingly rare in MAC[109] : It is recorded 3 times in SCO, yet only once in the much larger MAC corpus (and not once in the BNC/C). This indicates long clusters incorporating *people come* point to lexical nesting properties that appear to be local to Liverpool English.

*Not many people*, appears proportionally 13 times more often in SCO than MAC. While in SCO it is exclusively followed by a verb (*NOT MANY PEOPLE earn; NOT MANY PEOPLE deliver at home now*), it appears in MAC mostly in the idiomatic phrase *NOT MANY PEOPLE know* (7 occurrences out of a total of 25). This phrase is, however, not recorded in SCO.

### 6.5.5 Conclusion: PEOPLE occurrences

*People* is a useful item in this investigation as, compared to the other third-party reference markers, it is a relatively high-frequency term in both SCO and MAC. It has been shown that there is a large amount of agreement between MAC and BNC/C.

Despite *people* being proportionally more frequent, in MAC the word has the same top ten collocates as SCO. Only in the finer detail can we find differences of degree: while the collocates hint at some differences found in 2w clusters (i.e. *people* with *the* or *people* with *don't)*, other preferences of collocates in SCO highlight strong use of a number of key

[109] As before, the occurrence pattern of *people* in MAC appears very close to the patterns found in BNC/C (see Table 12)

words (which will be discussed in the next chapter): *people* with *yeah*, *people* with *like* and *people* with *just*.

There are statistically significant proportional differences in frequencies between MAC and SCO in 2w clusters. While *the people* and *people who* appear proportionally nearly 3 times as often in MAC than in SCO, *people say* is proportionally 5 times more frequent in SCO than in MAC.

The most important find when looking at *people* is the fact that people tends to be found in longer, formulaic, clusters. SCO and MAC share as their most common occurrence of the item *people* the cluster *a lot of people* which can be found to occur in around 3.0% of all uses of *people* in the respective corpora. At the same time, it is also the use of specific long clusters – which incorporate key *people* 2w clusters of each respective corpus – where the main divergence of use can be detected. *People who are* and *of the people* occur only in MAC (and BNC/C) but not in SCO. Conversely, though below the threshold level of statistical validity, we find *not many people* is rare in MAC compared to its use in SCO and the phrase *when people come in*, which appears 3 times in SCO, is virtually non-existent in MAC (or BNC/C).

## 6.6 3rd party referents – difference in degree, not in usage

Third-party reference is a natural feature of casual spoken conversation. Consequently, keywords can be used to test if there are differences of use of such referents in different spoken corpora. And, while there have been found differences in degree, BNC/C data on the whole is very similar to the MAC corpus figures discussed in this chapter.

Comparing MAC with SCO, one difference immediately strikes an observer, looking at all the core words investigated: their proportional occurrence is three times higher in MAC.

Looking at the collocates and short clusters found in *nobody, somebody, someone* and *anybody,* as well as the most frequently occurring long clusters of *people*, conclusive research is hampered by insufficient (SCO) data. Therefore, it can only be said that *nobody* appears always as a subject in SCO, while *nobody* is subject in MAC *not* in 100% of the cases. With the data available, *somebody, someone* and *anybody* are used in SCO in about the same way as MAC (and BNC/C). Though this stands in opposition to a claim that Scouse is a separate dialect, the high level of agreement in the findings for words and phrases which are typical of casual speech does underline the reliability of the corpora and methods of comparison used. The absence of massive differences makes the case for Scouse as a dialect weak. That there are, however, still corpus-specific features in the way the target words occur can be interpreted as *lexical primings* that are characteristic of this speech-community. For example,

the collocations that show that there are certain key words that are used far more frequently in SCO than in MAC (or BNC/C). *People*, while showing a high level of similar use, also presents a numbers significant differences in collocations, 2w and 3-4 word clusters.

To sum up, this chapter presents three things:

- If findings for SCO had always differed from the comparators, this would indicate a potential structural problem with regard to the SCO corpus. However, some key words and clusters of spoken English conversation present a picture where SCO is clearly used in the same way as MAC and BNC/C, and this supports the position adopted in this thesis that SCO represents a valid sample of Liverpool English.
- 3rd party reference markers are an interesting field of investigation and certain differences that have only been noted as trends are worthy of further investigation. This would need a far larger (SCO) corpus however.
- There is some indisputable evidence of divergent use in medium-high occurring clusters in SCO that highlight different semantic association and colligation choices, as the equivalents can either not be found or are extremely marginal in occurrence in both MAC and BNC/C.
- The comparison of some widely used items of casual spoken English undertaken here indicates some tendency of localised use.

This can be seen to support the theory of *lexical priming* (Hoey: 2005) in the context of spoken English variants.

# Chapter 7 Intensifiers and Discourse Particles in their use in casual speech

In order to make a valid comparison between variations of casual spoken English, the focus has to be on a certain set of lexical markers that are likely to be used by the two speech communities under comparison. This is particularly important, given that the subject of my research is spoken language – seen as more open to changes of expression and change over time than written language, which by its very nature is more conservative and bound to conventions.

Like written language, where we find a number of terms which are specific to the written mode, spoken language has a range of lexical items that are predominantly used in speech.

Choosing a representative sample of such items that can provide the basis of a neutral comparison means that each word has to meet certain criteria-

- It has to be a free-standing lexical item or cluster.
- It has to be an expression predominantly appearing in spoken language.
- It has to be a relatively high-profile word that is frequent in both corpora.

- It needs to be found in use by, or recognised by, both groups of speakers.
- It should reflect a function that is performed by both speech communities.

One class of words that appears to meet all these criteria are what one may call the *stress-markers*[110]. Some of these words are also referred to as discourse markers. Watts (1988) points out that Gumperz (1982) sees them as part of a *speech event*. *Stress-markers* or *intensifiers* play an important function in spoken language as they provide the speaker with a ready tool for highlighting the importance the speaker personally gives to certain statements.[111] As many of these words would be described as fulfilling a variety of functions depending in which context they are employed, and as they are discussed here in a corpus-led investigation, I will mostly refer to them as *discourse particles*. *Discourse particles* have attracted a lot of interest and a large array of research has been published about them (e.g. Watts: 1988; Juncker: 1993) as well as Streek (2002). Furthermore, they are described in teaching material[112], and there has been research into discourse particle usage amongst L2 speakers (for example in Fung & Carter: 2007). I will refer back these works during the discussion of each of the core terms. In this list we also have *Discourse Markers* (Schiffrin: 1987) which is seen as the standard work; and also

[110] These words go by a variety of names. Karin Aijmer and Lawrence Schourup refer to them as *Discourse Particles*. Fraser (1999) notes that a host of terms are employed to described words like the ones discussed here. See Fox Tree & Schrock (2002) for a more in-depth discussion.
[111] This is particularly important for the English language which, unlike French or German, makes little use of reflexive pronouns. These can, though they fulfil other functions as well, work as stress-markers.
[112] For a corpus-based example, see Biber, Conrad & Leech: 2002. They refer to them as *Inserts*.

Schourup (1985) who gives an overview of *well, like, now* and *you know* and *I mean*. (See also Schourup: 1999, 2001).

One form of *Discourse Markers, Intensifiers* are seen as a fitting item for linguistic research into spoken language as the following introduction by Rika Ito and Sali Tagliamonte shows:

> This area of grammar (*intensifiers*) is always undergoing meaning shifts (Stoffel 1901:2), partly because of "speaker's desire to be 'original', to demonstrate their verbal skills, and to capture the attention of their audience" (Peters 1994: 271)
>
> The first relevant question that arises is: What is an intensifier? There are two types – intensives and downtoners (e.g. Stoffel 1901, Quirk et al. 1985). (…) we restrict ourselves to those of the first type, in part because they are more frequent (Mustanoja 1960:316), but also because, we believe, they are more interesting. The terminology referring to these types of adverbs is not entirely uniform among scholars.
>
> (Ito & Tagliamonte 2003: 258)

Since the 1990s, the use of real data from corpora became established in the study of this class of words. Partington (1993) looks at diachronic change of intensifiers and says that "this can be explained as part of a wider process of delexicalisation", a point we will come back to later. Among others, Miller and Weinert (1995) and Macaulay (2002) have looked at the usage of *like* and *you know* in Scottish English corpora. Ajimer (2002) looks at London–Lund Corpus occurrences of *now, oh, just, actually* and *sort of*.

Ito & Tagliamonte (2003) and Tagliamonte (2004) focus on discourse particles employed by generations of speakers in York (UK) and Canada respectively and highlight another important aspect to this research, namely, rapid change:

> According to Partington (1993:180) "in this sea of change, processes of expansion and contraction are occurring all the time," which was also observed earlier by Bolinger, as described above.
>
> Given this backdrop, it is not at all surprising to find in spoken data hearty variability in the use of intensifiers (see 1–2), even in the same speaker in the same stretch of discourse, as in (4), undoubtedly reflecting the coexistence of older and newer layers in the process of change. (Ito & Tagliamonte 2003: 261)

I will refer back to their paper where there are direct comparisons between my data and theirs. Likewise, in my research I try to show that there are in fact dominant uses of certain words and clusters and such preferences or non-preferences would be the hallmark of that particular language community. Were I to find solid proof for my hypothesis, the quoted "hearty variability", even as found in a single speaker, would not be that great.

Discourse particles are likely to be used (unlike many nouns) by every speech community. Nevertheless, if my hypothesis is correct that dialects differ in the way they use the same words, speech communities, like individual speakers, should be found to express certain characteristics by their non-use, use, or an apparent over-use of discourse particles, and by

the collocational and colligational environments in which we can find these discourse particles.

As there are a number of lexical items that fulfil this function, it is also valid for this research to see whether or not some terms are used more frequently than others or in a different context when comparing the use in two speech communities.

In this chapter, I will focus on the following words:

- Just
- Like
- Really
- Very
- Well
- Yeah

I shall investigate which words are most likely to collocate with these markers and compare the occurrence of clusters that result.

The order above is purely alphabetical. An alternative way to order these items is by frequency as they appear in the 120.000 words of SCO as shown in Table 1(a):

| Core word | SCO frq. | SCO % | MAC frq. | MAC % | BNC/C frq. | BNC/C % |
|---|---|---|---|---|---|---|
| YEAH | 1651 | 1.60 | 56,818 | 1.72 | 58,708 | 1.46 |
| LIKE | 970 | 0.81 | 26,570 | 0.81 | 21,920 | 0.54 |
| JUST | 546 | 0.46 | 30,739 | 1.0 | 19,693 | 0.49 |
| WELL | 320 | 0.27 | 36,869 | 1.2 | 35,806 | 0.89 |
| REALLY | 289 | 0.35 | 11,471 | 0.27 | 9,128 | 0.23 |
| VERY | 153 | 0.14 | 24,939 | 0.83 | 6,525 | 0.16 |

Table 1(a): The most frequent discourse markers in SCO, comparative frequencies in MAC and BNC/C Percentages in relation to the total corpus.

These raw figures show that there is agreement of proportional frequencies across the corpora only in the cases of *yeah* and *really*. While figures for *just* and *very* are similar in SCO and BNC/C, MAC use is far higher proportionally. *Like* is the only term where proportional use in BNC/C is clearly lower. At the same time, though, *well* is lower in its proportional frequency in SCO than in either MAC or BNC/C.

I will continue to occasionally refer to the BNC/C, but the main comparison will be between SCO and MAC.

| Core word | SCO frq. | SCO % | MAC frq. | MAC % | Log-Likelihood |
|---|---|---|---|---|---|
| YEAH | 1651 | 1.60 | 56,818 | 1.72 | 80.64 |
| LIKE | 970 | 0.81 | 22,858 | 0.81 | 23.29 |
| JUST | 546 | 0.46 | 30,739 | 1.0 | 342.40 |
| WELL | 320 | 0.27 | 36,869 | 1.2 | 1081.89 |
| REALLY | 289 | 0.35 | 11,471 | 0.27 | 40.88 |
| VERY | 153 | 0.14 | 24,939 | 0.83 | 929.83 |

Table 1(b) Log-Likelihood test figures of the core words in MAC : SCO comparison

As Table 1(b) shows, the statistical test to check in how far those particular core words diverge in their proportional occurrence is very much in line with the divergences found when the proportional percentages are compared. The one exception is *very* which, according to the test, should occur with a far higher comparative frequency in SCO.

## 7.1 YEAH

### 7.1.1 Introduction of the term

The intensifier *yeah* is a relevant case in this discussion, as the item *yeah* is prototypically one thing – a form of the word of approval *yes* in its spoken form – but functionally always emphatic or stressing, in short, an intensifier.

*Yeah* is less well investigated than many of the other items in our list. Schiffrin (1987) describes it as an *acknowledgement marker* or *receipt marker* (Schiffrin 1987: 89 and 260). Fung & Carter give a more detailed description of *yeah* as used in the CANCODE student subcorpus:

> In spoken discourse *yeahs* function primarily in interpersonal and structural categories to acknowledge, agree, affirm, and mark continuation. (…) Native speakers (use *yeah* to) exhibit understanding or acknowledgement (interpersonal category), or as a continuer of the progress of the primary speaker's turn (structural category). Syntactically, the environment in which *yeah* occurs is less varied in the student data than in CANCODE. *Yeahs* in the interpersonal category appear mostly in isolation in turn-initial position, whereas use in the structural category tends to correlate with a turn-medial use, combining with other DMs [*Discourse Markers*] to emphasize the propositions made in the prior discourse. (…). (Fung & Carter 2007: 431)

The functions are visible when some of the uses of *yeah* are looked at: *Yeah!* (marking success); *Yeah, right* (jeering); *alright, yeah* (strengthening the qualifier alright).

It is remarkable how coy the dictionaries consulted are with regard to this item. *Yeah* occurs with a high level of frequency amongst the set of intensifiers. All the same, while other words have elaborate entries, *yeah* is dealt with at the bare, minimum level:

> yeah (also yeh)
>
> exclamation & noun informal non-standard spelling of yes.
>
> (Concise Oxford English Dictionary)
>
> Main Entry: yeah; Pronunciation: ‚ye-„, ‚ye˜, ‚ya-„; Function: adverb; Etymology: by alteration; Date: 1902; : yes
>
> (Merriam Webster Dictionary)
>
> yeah (informal) YES. yeah right (spoken) used for saying that you do not believe something someone has just told you.
>
> (Macmillan English Dictionary)

All three indicate *yeah* is a form of *yes*. Only the Concise OED remarks upon the function as an exclamation. It is Macmillan Dictionary that puts stress on the aspect of *informality* – with the latter being the only dictionary giving three other important pieces of information: (1) the cluster yeah right (as mentioned above), (2) an indication that it is mostly found in spoken use – hence this only example of *yeah* as part of a phrase / cluster, and (3) that it is a high-frequency word, commonly used.

### 7.1.2 YEAH is not YES

Fung and Carter (2007) point out that non-native speakers of English do not necessarily make a distinction between *yeah* and yes in their spoken utterances, whereas native speakers do:

> The data also reveal that there is an over reliance on *yes* rather than *yeah* among the Hong Kong subjects, yet *yeah* (which is commonly associated with a discourse-marking role) was found to be the third most frequent word in the pedagogic sub-corpus of CANCODE (…) Its frequency is 0.47 per cent [in the Hong Kong data] in comparison with 0.9 per cent in its British counterpart, with a great contrastive frequency of –0.43 (Table 4). In contrast, its formal form *yes* is widely represented in the student corpus, being the fourth most frequent word (0.94 per cent) in the present data. (Fung & Carter 2007: 431)

The use of the lexical item *yeah* has been checked in comparison to the use of *yes* in a number of corpora. There appears to be strong evidence that they are different lexical items. This is also been noted by Fung & Carter:

> With its backward-pointing role, *yeah* is employed primarily as a solidarity building device to mark agreement which a listener would reasonably be expected to recognize, and also as a reception marker to signal coherence within and between turns. Throughout the British English extracts, the speakers respond to each other at various points using *yeahs*, showing that the speakers are expressing a general acknowledgement of the preceding interactive unit (Jucker and Ziv 1998a). This is a very frequent usage in which they appear singly as an individual turn without indicating any change of speakership. (Fung & Carter 2007: 431f.)

Table 2: Direct comparison of YEAH and YES proportional frequencies and collocate patterns in SCO, MAC, BNC/S and BoE.

| SCO | Total | % | SCO | Total | % | MAC | Total | % | MAC | Total | % |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *YEAH* | *1,651* | *1.6* | *YES* | *101* | *0.1* | *YEAH* | *51,814* | *1.57* | *YES* | *17,994* | *0.55* |
| I | 241 | 14.6 | I | 16 | 15.8 | I | 11,166 | 19.7 | YOU | 6,784 | 37.7 |
| THE | 180 | 10.9 | (PAUSE) | 14 | 13.8 | THE | 7,747 | 13.6 | I | 5,393 | 29.8 |
| YOU | 168 | 10.2 | YEAH | 13 | 12.8 | YOU | 9,393 | 16.5 | IT | 3,375 | 18.8 |
| A | 165 | 10 | OH | 12 | 11.8 | A | 4,924 | 8.7 | OH | 2,843 | 15.8 |
| OH | 163 | 9.8 | SO | 12 | 11.8 | OH | 6,322 | 11.3 | THAT | 2,631 | 14.6 |
| IT | 142 | 8.6 | THE | 12 | 11.8 | IT | 9,947 | 17.5 | THE | 2,489 | 13.8 |
| AND | 136 | 8.2 | A | 11 | 10 | AND | 6,520 | 11.5 | WELL | 2,022 | 11.2 |
| KNOW | 107 | 6.5 | IT | 11 | 10 | KNOW | 2,567 | 4.6 | YEAH | 2,019 | 11.2 |
| THAT | 93 | 5.6 | OF | 9 | 8.8 | THAT | 7,806 | 13.7 | KNOW | 1,916 | 10.6 |
| IT'S | 91 | 5.5 | YOU | 9 | 8.8 | IS | 2,506 | 4.5 | AND | 1,847 | 10.4 |
| **BNC/C** | **Total** | **%** | **BNC/C** | **Total** | **%** | **BoE** | **Total** | **%** | **BoE** | **Total** | **%** |
| *YEAH* | *58,708* | *1.5* | *YES* | *17,876* | *0.52* | *YEAH* | *151,056* | *1.64* | *YES* | *113,876* | *1.24* |
| I | 10,542 | 17.6 | I | 4,444 | 17.4 | THE | 25,602 | 16.9 | THE | 19,669 | 17.3 |
| EAH * | 9,633 | 0.2 | OH | 3,178 | 12.7 | AND | 23,609 | 15.6 | I | 19,607 | 17.3 |
| YOU | 7,601 | 12.7 | YOU | 3,058 | 12 | YOU | 21,139 | 14 | AND | 19,084 | 16.8 |
| IT | 6,648 | 11 | IT | 2,608 | 10.2 | I | 20,748 | 13.7 | YOU | 15,575 | 13.7 |
| AND | 5,801 | 9.9 | THE | 2,406 | 9.5 | TO | 14,751 | 9.8 | OH | 13,343 | 11.7 |
| THE | 5,463 | 9.1 | AND | 2,230 | 8.8 | THAT | 14,448 | 9.5 | IT | 12,792 | 11.2 |
| OH | 4,728 | 7.9 | A | 1,602 | 6.3 | A | 14,445 | 9.5 | A | 11,178 | 9.8 |
| BUT | 4,017 | 6.7 | THAT | 1,528 | 6 | IT | 14,022 | 9.3 | ER | 10,927 | 9.6 |
| A | 3,895 | 6.5 | HE | 1,289 | 5 | OF | 13,351 | 8.8 | TO | 10,803 | 9.5 |
| THAT | 3,669 | 6.3 | YEAH | 994 | 3.9 | IN | 9,711 | 6.4 | THAT | 9,946 | 9 |

*YEAH misspelt: 1.20% + 0.20% =1.40%

tokens used for word list 4,855,134

**Table 2: Direct comparison of YEAH and YES proportional frequencies and collocate patterns in SCO, MAC, BNC/S and BoE.**

Table 2 throws up a number of interesting features of *yes* compared to yeah use. As comparison is made amongst four spoken corpora, a high level of salience regarding yeah use can be obtained[113]. It will be seen that *yeah* is the preferred choice compared to *yes* in spoken English. *Yeah* tends to have *yes* as a collocate and *yeah* tends to take different collocates, with different frequencies, from *yes*. As far as spoken contexts are concerned, *yeah* and *yes* must therefore be treated as different words. Because *yeah* is more frequent in all corpora, a wider range of functions can be assumed to be covered by *yeah*. A more detailed study of its use will be found below[114].

#### 7.1.2.1 Comparison of YES and YEAH collocates

Both *yeah* and *yes* collocate freely and the percentages of co-occurrence for even the top clusters are relatively low. In SCO, MAC and BNC/C corpora, yeah occurs significantly more often than *yes*. Indeed, *yeah* occurs 14 times more often than *yes* in SCO, over three times more often in MAC and over twice as often in the BNC.

Looking at the top collocates of *yes* and *yeah* (Table 3) there are differences in ranking found throughout.

[113] For further comparison, I have also checked the occurrence patterns of *yes* and *yeah* in the BoE UkSpoken. However, data from the BoE have to be discounted. As Table 1 shows, collocates for both *yeah* and YES are the same, leading me to conclude that transcribers heard *yeah* but normalised it to YES in writing.

[114] Hongying (2003: 15) has already made the point that *yes*, *yeah* and *yep* in spoken English function as different similar items. Unfortunately, I did not read his paper on turn-taking until 2010. Fortunately, my own findings support his.

Table 3: YEAH and YES top clusters compared in 4 corpora

| SCO cluster YEAH | tot. | % | SCO cluster YES | tot. | % | MAC cluster YEAH | tot. | % | MAC cluster YES | tot. | % |
|---|---|---|---|---|---|---|---|---|---|---|---|
| YEAH - YEAH - YEAH | 39 | 2.4 | YES OF COURSE | 3 | 3 | YEAH YEAH YEAH | 3879 | 7.5 | YES YOU KNOW | 397 | 2.2 |
| OH YEAH - YEAH | 16 | 1.0 | YES YES FOR | 2 | 2 | YEAH YEAH I | 1201 | 2.3 | YES I MEAN | 316 | 1.8 |
| THAT'S RIGHT YEAH | 12 | 0.7 | YES YOU CAN | 2 | 2 | OH YEAH YEAH | 976 | 1.9 | YES I KNOW | 251 | 1.4 |
| YEAH THAT'S RIGHT | 9 | 0.6 | | | | YEAH. YEAH . AND | 961 | 1.8 | YES , YES , YES | 247 | 1.4 |
| YEAH OH YEAH | 9 | 0.6 | | | | YEAH I KNOW | 767 | 1.5 | YOU KNOW YES | 146 | 0.8 |
| | | | | | | YEAH OH YEAH | 736 | 1.4 | YES IT IS | 126 | 0.7 |
| | | | | | | | | | YES YOU CAN | 120 | 0.6 |
| BNC/C cluster YEAH | tot. | % | BNC/C cluster YES | tot. | % | BoE cluster YEAH | tot. | % | BoE cluster YES | tot. | % |
| YEAH YEAH YEAH | 1.015 | 1.7 | YES, YES , YES | 453 | 2.5 | YEAH. YEAH. YEAH | 19,022 | 21.8 | YES. YES. YES | 14,118 | 12.2 |
| YEAH I KNOW | 990 | 1.7 | OH YES YES | 281 | 1.6 | YEAH. YEAH. AND | 3,643 | 4.2 | OH YES. YES | 4,002 | 3.5 |
| YEAH YEAH I | 601 | 1.0 | YES I KNOW | 238 | 1.3 | YEAH. YEAH. SO | 3,162 | 3.6 | YES. YES. AND | 3,793 | 3.3 |
| YEAH . I MEAN | 544 | 0.9 | YES , THAT 'S RIGHT | 232 | 1.3 | YEAH. I MEAN | 3,117 | 3.6 | YES THAT'S RIGHT | 3568 | 3.1 |
| YEAH BUT I | 540 | 0.9 | YES YES I | 223 | 1.2 | YEAH. YEAH. I | 2,953 | 3.4 | THAT'S RIGHT. YES | 2,277 | 2 |
| YEAH YOU KNOW | 533 | 0.9 | YES I THINK | 212 | 1.2 | RIGHT. YEAH. YEAH | 2,902 | 3.3 | YES. OH YES | 1,877 | 1.6 |
| OH YEAH YEAH | 474 | 0.8 | YES IT IS | 188 | 1.1 | YEAH. YOU KNOW | 2,333 | 2.7 | THAT'S RIGHT YES | 1,585 | 1.3 |
| | | | OH YES I | 180 | 1.0 | | | | | | |

Table 3: YEAH and YES top clusters compared in 4 corpora

In all four corpora, *I*, *the* and *you* are key collocates of *yeah*. This is also true for *yes* to a certain extent. *I* collocates proportionally more often with *yes*. This hints at the fact that major differences of use may only surface when *yeah* and *yes* cluster patterns are compared.

#### 7.1.2.2 Comparison YES vs. YEAH clusters

Use of collocates on their own does not provide conclusive proof that *yeah*, in casual spoken English, is employed in a different way from *yes*. Consequently, the next step is to compare the most frequently occurring clusters of both *yeah* and *yes* in all four corpora.

Table 3 shows that, though there is some overlap, on the whole *yeah* and *yes* appear as part of different sets of clusters. These differences are even more pronounced when the proportional uses are compared. We find, for example, *yeah yeah yeah* occurs proportionally over four times more often than *yes yes yes* in MAC. Oddly, the results are inverted in the BNC where the triple repetition is proportionally used more often with *yes*.

The major and most important difference however lies in the fact that many clusters with *yeah* have no equivalent with *yes*.

In SCO, where the transcription has not been normalised, none of the top clusters has an equivalent. In MAC, where the transcripts do not appear to be normalised, two out of the six top clusters are the same. However, the proportional use is different.

In the BNC/C, three out of seven of top-occurring clusters for *yeah* and *yes* overlap. However, a clear dividing line between *yes* and *yeah* is drawn by the similar clusters *yes that's right* compared to *that's right yeah*. Here, word order is determined by the choice of either *yes* or *yeah* use.

All the corpora have in common the triple repetition of both *yeah* and *yes* as one of the most frequent clusters.

#### 7.1.2.3 Comparison YES vs. YEAH conclusion

At this point, the relevant differences between the clusters can be highlighted. Only the SCO and MAC contain recent recordings of casual, informal BE speech. The BNC/C contains a high proportion of speech recorded in academic environments and structured interviews.

Despite of this [115], the differences in the use of the terms *yeah* and *yes* have become obvious by comparing their proportional occurrence in the most frequent clusters. This is a fact that has to be kept in mind during the following discussion.

So far, we have discovered that *yeah*, as opposed to *yes*, is by far the preferred option in casual speech. In total numbers, the difference in SCO is very strong – the ratio is 16:1 - and though in MAC the ratio is much smaller, the ratio is still 4:1.

[115] Though these are a valid reason not to undertake a like-for-like comparison

| SCO Rk. | *Yeah* collocate | SCO Total | % | MAC Rk. | MAC Total | % |
|---|---|---|---|---|---|---|
| *1* | *YEAH* | *1,651* | *1.6* | *1* | *51,814* | *1.57* |
| 5 | I | 241 | 14.6 | 4 | 11,166 | 19.7 |
| 6 | THE | 180 | 10.9 | 9 | 7,747 | 13.6 |
| 7 | YOU | 168 | 10.2 | 6 | 9,393 | 16.5 |
| 8 | (A) | 165 | 10 | 13 | 4,924 | 8.7 |
| 9 | OH | 163 | 9.8 | 11 | 6,322 | 11.3 |
| 10 | IT | 142 | 8.6 | 5 | 9,947 | 17.5 |
| 11 | AND | 136 | 8.2 | 10 | 6,520 | 11.5 |
| 13 | PAUSE | 128 | 7.8 | | | |
| 16 | KNOW | 107 | 6.5 | 20 | 2,567 | 4.6 |
| 17 | THAT | 93 | 5.6 | 8 | 7,806 | 13.7 |
| 18 | IT'S | 91 | 5.5 | | | |
| 19 | IS | 86 | 5.2 | 23 | 2,506 | 4.5 |
| 20 | LIKE | 78 | 4.7 | 35 | 1,586 | 2.8 |
| 21 | THAT'S | 77 | 4.7 | | | |
| 22 | TO | 75 | 4.5 | 15 | 3,861 | 6.9 |
| 23 | BUT | 70 | 4.2 | 14 | 3,890 | 6.9 |
| 24 | WAS | 68 | 4.0 | 25 | 2,255 | 4.6 |
| 26 | OF | 57 | 3.5 | 19 | 2,734 | 4.9 |
| 27 | EHM | 55 | 3.3 | 30 | 1,927 | 3.4 |
| 28 | JUST | 53 | 3.2 | 50 | 1,165 | 2.0 |
| 29 | SO | 53 | 3.2 | 29 | 1,934 | 3.4 |
| 30 | WHAT | 53 | 3.2 | 26 | 2,101 | 3.7 |
| 31 | HE | 50 | 3.0 | 17 | 3,560 | 6.3 |
| 33 | WELL | 50 | 3.0 | 18 | 3,047 | 5.4 |
| 35 | IN | 47 | 2.85 | 22 | 2,524 | 4.5 |
| 36 | ME | 46 | 2.7 | 68 | 839 | 1.5 |
| 37 | THEY | 46 | 2.7 | 16 | 3,763 | 6.5 |
| 38 | RIGHT | 45 | 2.5 | 31 | 1,769 | 3.0 |

Table 4(a): SCO and MAC *yeah* top collocates. (Rk. = rank)

### 7.1.3 YEAH collocates in the SCO and MAC corpora

Concentrating on the differences of use of *yeah* in SCO and MAC, Table 4(a) shows that *yeah* itself is proportionally used as often by SCO speakers as it is by MAC speakers. Likewise, the most common collocates are similar in their proportional frequency, too. However, the percentages of use for the highest-occurring collocates are mostly lower in SCO.

The exclamation *oh* and the conjunct *and* are amongst the most frequent collocates in both corpora. Colligational features include the fact that *yeah* appears with personal pronouns and determiners with similar proportional frequencies in both corpora.

Other intensifiers (i.e. *right*) and hesitancy markers (i.e. *ehm; well*) account for no less than 2.5% of all occasions in both corpora when *yeah* is used.

The largest differences in proportional use are found for the following items:

| *Yeah* collocate | SCO Total | % | MAC Total | % | LL |
|---|---|---|---|---|---|
| I | 241 | 14.6 | 11,166 | 19.7 | 40.67 |
| THE | 180 | 10.9 | 7,747 | 13.6 | 19.44 |
| YOU | 168 | 10.2 | 9,393 | 16.5 | 66.78 |
| IT | 142 | 8.6 | 9,947 | 17.5 | 118.87 |
| AND | 136 | 8.2 | 6,520 | 11.5 | 27.49 |
| KNOW | 107 | 6.5 | 2,567 | 4.6 | 6.83 |
| THAT | 93 | 5.6 | 7,806 | 13.7 | 125.45 |
| LIKE | 78 | 4.7 | 1,586 | 2.8 | 12.32 |
| TO | 75 | 4.5 | 3,861 | 6.9 | 21.24 |
| BUT | 70 | 4.2 | 3,890 | 6.9 | 27.18 |
| HE | 50 | 3 | 3,560 | 6.3 | 43.85 |
| WELL | 50 | 3 | 3,047 | 5.4 | 27.11 |

Table 4(b) *Yeah* collocates that are most divergent between SCO and MAC.

It must be noted that where *yeah* collocates are proportionally more frequent in SCO, the statistical test shows that significance is only at a 99.0% (*know*) and 99.9% (*like*) level. For all the other collocates, significance is clearly above the 99.99% level. As Table 4(b) shows, *yeah* with *that,* with *it,* and with *you* are the most significant of those, occurring more than twice as often proportionally in MAC than in SCO.

The next section will show whether these differences are reflected in the yeah clusters to be found.

#### 7.1.4.1 Most frequent YEAH clusters – detailed use

Table 5 below looks at the most frequent 2-3w *yeah* clusters in SCO and their MAC equivalents.

The statistical tests in Table 5 show that about half these 2-3w *yeah* clusters appear with significantly different proportional frequencies in the two corpora. Only *Oh yeah*, *I know yeah* and *that's right yeah* show no discernible differences in use. However, it is notable that there are clear differences where a similar phrase with a different word order is used - for example *oh yeah* vs. *yeah oh.*

These differences will be discussed in detail below.

| 2-4w *Yeah* clusters | SCO tot. | MAC tot. | LL |
|---|---|---|---|
| **YEAH YEAH** | 204 | 17278 | 281.25 |
| **YEAH YEAH YEAH** | 41 | 3969 | 76.77 |
| **YEAH YEAH YEAH YEAH** | 10 | 864 | 14.46 |
| OH YEAH YEAH | 20 | 1001 | 4.98 |
| OH YEAH | 136 | 4263 | 0.00 |
| **YEAH OH** | 13 | 2324 | 75.30 |
| **YEAH BUT** | 36 | 3518 | 68.81 |
| YEAH YEAH BUT | 15 | 687 | 2.37 |
| YEAH BUT I | 5 | 503 | 10.17 |
| **YEAH THAT'S** | 29 | 1928 | 20.80 |
| **YEAH IT'S** | 23 | 2669 | 62.53 |
| **YEAH AND** | 20 | 4000 | 138.01 |
| **YEAH YOU** | 17 | 2351 | 64.01 |
| YEAH YOU KNOW | 7 | 552 | 8.08 |
| YOU KNOW YEAH | 2 | 573 | n/a |
| RIGHT YEAH | 21 | 1138 | 7.38 |
| THAT'S RIGHT YEAH | 17 | 651 | 0.70 |
| YEAH THAT'S RIGHT | 10 | 537 | 3.39 |
| YEAH I KNOW | 12 | 771 | 7.74 |
| I KNOW YEAH | 6 | 181 | 0.01 |
| WELL YEAH | 16 | 763 | 3.15 |
| **YEAH WELL** | 7 | 2378 | 102.25 |

Table 5: Most frequent 2-4w SCO *yeah* clusters[116]

[116] Ordered by 2w clusters which form part of longer clusters.

Table 5 presents an interesting insight into the comparative uses of *yeah*. Apart from *yeah* single-word-repetition, *yeah and* and *yeah well* stand out as significantly divergent. We also find that there are strongly significant (as shown by values of LL > 15.13) divergences in a number of 2w clusters - notably *yeah oh, yeah but, yeah you* etc. However, table 5 also gives us those examples where these appear in longer clusters and neither the proportional frequency of use nor the statistical tests indicate a strong divergence of use for 3w clusters like *yeah you know.* Looking at how *yeah* clusters with *oh* serves as a general example:

*Yeah* with *oh*, (cf. Table 5) appears in the top seven 3-word clusters of SCO, MAC and BNC/C. In the MAC, the use for *oh yeah* are very similar to BNC/C[117].

| OH + YEAH cluster | SCO | % | MAC | % | LL |
|---|---|---|---|---|---|
| OH YEAH | 136 | 8.23 | 4263 | 8.22 | 0.00 |
| YEAH OH | 13 | 0.80 | 2324 | 4.48 | 75.30 |
| YEAH OH YEAH | 9 | 0.55 | 311 | 0.60 | 0.08 |
| YEAH OH YEAH YEAH | 5 | 0.30 | 119 | 0.23 | 0.34 |

Table 6: *Yeah* with *oh* clusters in SCO and MAC.

As Table 6 shows, *yeah oh* is an outlier in SCO, as all clusters either incorporating *yeah oh* in a larger unit appear with about the same proportional frequencies in SCO and MAC - as does the reverse 2w cluster *oh yeah*.

[117] The BNC/C records 41,565 instances of *oh* of which only 2,989 occur in the phrase *oh yeah*. This means that *oh yeah* represents only 7.2 % of all uses of *oh* in BNC/C, rather than 25% of all uses of *oh* as in SCO.

In 7.1.5.2 and 7.1.5.3 we will discuss those longer *yeah* clusters where occurrence patterns in SCO diverge significantly from what we find in MAC.

### 7.1.4.2 Repetition clusters in YEAH

In 7.1.2, we have seen that *yeah* (and *yes*) appear in comparator corpora with similar high proportional frequencies of single-word repetition (*yeah yeah).* While we do find clusters in SCO of multiple (not just single) single-word repetition, multiple single-word repetition is still far more common in MAC. It occurs three times as often as a 3-word cluster and as a 4 word cluster in MAC than it does in SCO.

| cluster | SCO Freq. | SCO % | MAC FREQ. | MAC % | Log-Likelihood |
|---|---|---|---|---|---|
| YEAH YEAH YEAH | 41 | 2.53 | 3969 | 7.66 | 76.77 |
| YEAH YEAH YEAH YEAH | 10 | 0.60 | 986 | 1.90 | 19.49 |
| OH YEAH YEAH YEAH | 8 | 0.48 | 284 | 0.55 | 0.12 |
| OH YEAH - YEAH | 16 | 1.00 | 1001 | 1.93 | 9.47 |

Table 7: *yeah* repetition clusters compared

Table 7 shows that the rarer form of *yeah* repetition - *oh yeah yeah yeah* - has no meaningful differences of use between the two corpora. There is, however, notably less use of *oh yeah - yeah* in SCO than there is in MAC. The biggest and most significant differences are, however, in multiple single-word-repetition which, as we already have seen in Table 5, are proportionally used far more often in MAC than in SCO in 2w, 3w and 4w single *yeah* repetition clusters. As we will also see in chapter 8.2,

in MAC there is a tendency to find single-word repetition of discourse markers at a higher proportion than in SCO.

#### 7.1.4.3 YEAH clusters with other intensifiers

Section 7.1.2 has shown that *yeah* acts not simply as a marker of agreement but is, as a discourse particle, employed in different ways. In this section we look at how far *yeah* plays a role as intensifier. *Yeah* can be used to either stress or dampen a (part of a) statement[118].

In Table 5 we have seen that *yeah well* in SCO diverges significantly in its use as to what we find in MAC. It is a single 2w phrase that does not otherwise appear to cluster. That it is, as a phrase, found significantly less than we would expect in SCO, has as its most likely explanation that *well* itself is a rare word in SCO (see chapter 7.2).

| **Cluster** | SCO | % | MAC | % | LL |
|---|---|---|---|---|---|
| YEAH AND | 20 | 1.2 | 4000 | 7.7 | 138.01 |
| YEAH YEAH AND | 5 | 0.3 | 990 | 1.9 | 33.97 |

Table 8: *Yeah and* clusters compared

As Table 8 clearly shows, both *yeah and* and *yeah yeah and* differ significantly in their use. Both are far less frequent in their proportional occurrence in SCO than in MAC. The cluster also appears as a two-speaker utterance (*yeah // and*). In this case, we find the 3w (2 speakers)

[118] These depend very strongly on the intonation pattern. *I know that. Oh god yeah* or *Northend. Yeah. Always been Northend,* presents *yeah* as part of a phrase that gives extra stress; *something like that yeah* seems to indicate agreement yet in a diluted form.

utterance S1 – *yeah* // S2- *and then.* This, again, occurs twice as proportionally frequent in MAC than it does in SCO.

Another interesting point is the distribution of the *yeah you know cluster*. This appears in both SCO and MAC not just as a single-speakers cluster but can also be found to be split between two speakers:

S1 yeah
S2 you know

Looking, therefore, at the clusters in detail we find the following:

| Cluster | SCO Freq. | SCO % | MAC FREQ. | MAC % | Log-Likelihood |
|---|---|---|---|---|---|
| YEAH I KNOW | 12 | 0.70 | 771 | 1.50 | 7.74 |
| I KNOW YEAH | 5 | 0.30 | 341 | 0.66 | 3.87 |
| YEAH // YOU KNOW | 10 | 0.60 | 187 | 0.40 | 2.19 |
| **YOU KNOW / / YEAH** | **8** | **0.50** | **1037** | **2.00** | **26.80** |
| YEAH YOU KNOW | 7 | 0.42 | 552 | 1.07 | 8.08 |
| YOU KNOW YEAH | >5 | n/a | 1537 | 2.97 | n/a |

Table 9: *Yeah* with *know* clusters in single and 2 speaker formats.

First of all, the one cluster that appears to be highly significant in its divergence: *you know//yeah*. On closer inspection, it is shown that every single exchange of this sort consists of the informant ending an utterance in *you know* and in every single case it is the recorder who answers with *yeah*. Only if there had been cases where such an exchange between two informants had been recorded can be classed as reliable. The other recorded 2-speaker cluster, *yeah // you know* shows very little difference in use between the two corpora.

The second issue Table 9 shows is that in SCO we see a tendency to use (single speaker) three-word *yeah* clusters (in this case with *know*) that are used with proportionally far lower frequencies in SCO than in MAC. This is significant at the just above the 99.0% level. Therefore, while *I know yeah* is used at about the same level in both corpora, *yeah I know* appears with a significantly lower percentage of use. In the case of *yeah you know* the difference is even more significant and the 3w cluster *you know yeah* is one of the most frequently occurring 3w *yeah* clusters in MAC while it is too rare in SCO to qualify for a valid statistical comparison.

### 7.1.5 Conclusions for YEAH

Research into *yeah* has brought up a relevant insight into the use of this term in spoken English. *Yes* and *yeah* are still used to express the same thing – agreement. And overall the clusters brought up show that *yeah* is clearly linked to *yes*. It is, nevertheless, in spoken English, there is enough evidence to state that *yeah* becomes a separate term in its own right, a term that is employed as part of different clusters that are more intricate than the use of *yes* (or *no*) would allow.

In comparing the use of *yeah* in SCO, MAC and in BNC/C, it has been found that the term is used largely in the same way. *Yeah* collocates freely and there are only a few fixed clusters. The clusters that are found

for the most part occur in all corpora, to a degree, occur with the same proportional frequencies, too. While there are a number of 2w clusters that diverge significantly between SCO and MAC, the majority of longer clusters occur with no significant differences.

Probably the most significant difference lies in single-word repetition clusters. 2-4w single word repetition of *yeah* is in all cases proportionally more frequent in MAC.

Other, significant, differences can be found in the occurrence pattern of *yeah and* and *yeah yeah and* as well as *yeah I know* and *yeah you know* all four of these clusters appear significantly less often in use in SCO than they do in MAC.

A further point of interest is in the split into single- and two-speaker clusters[119].

To conclude: though the use of *yeah* is very similar in both corpora, there is ample difference to be found in the concrete use of the term when SCO and MAC are compared.

[119] This is an aspect of transcription rarely referred to in corpus linguistics articles. In fact, as far as I am aware, *cross-turn clusters* have not been discussed by anyone before and have been investigated for the first time in connection with lexical priming in this thesis. It is possible that, in this respect, a small corpus is of advantage. It enables one to pull out all available concordance lines of a particular cluster without too much effort and occurrence of clusters across separate turns by separate speakers becomes visible. In the context of this thesis, however, no major differences have been found where SCO is compared with MAC.

## 7.2 Uses of WELL

### *7.2.1 Introduction and literature discussion*

*Well* is widely used as a discourse marker. The COED and the Merriam Webster however, focus on its uses as adjective or adverb (of which, only the latter will be discussed below) while the Macmillan Dictionary indicates that *well* has a function as a discourse marker. *Well* in its discourse *marker* use has been investigated widely. A.H. Jucker gives a comprehensive description of its function:

> In a conversation, the relevant context is continually being negotiated throughout a text or discourse. This is not necessarily a straightforward and linear movement; digressions, mistaken assumptions about partner's context, etc. may occur. It is exactly in these positions that the discourse marker *well* can occur. It signals that the context created by an utterance may not be the most relevant one for the interpretation of the next utterance. (Jucker 1993: 451)

Jucker makes clear that *well* is an integral part of spoken interaction - while already indicating that employing *well* means a degree of uncertainty and also operates as a downtoner ("an utterance may not be the most relevant…").

L.C. Schourup (1985; 1999; 2001) also investigated *well* intensely. He indicates that views vary on how *well* functions:

> A substantial body of research deals with semantic and pragmatic aspects of the discourse marker *well.* (…) *Well* has probably received more attention than any other English discourse marker.1 Most studies have concluded that *well,* as a marker, has an invariant semantic or functional core. There is, however, a lack of consensus regarding how this core should be formulated. (Schourup 2001: 1026)

Schourup (2001) reviews most of the current research about *well.* He states that *well* is semi-lexical and half extra-lingual. *Well*, according to him, acts as a gesture:

> I have argued that it may be more appropriate to view *well* as quasi-linguistic vocal gesture used to 'portray' the speaker's mental state than as a 'full-fledged word' linguistically encoding information about that state. (Schourup 2001: 1026)

The problem with this is that something voiced cannot be a gesture. It may be seen, however, to fulfil the same function as a gesture. For example, *well* may support a point made.

Deborah Schiffrin sees it as a pre-closing device:

> At more global levels of conversational organisation, *well* (alongside with *okay* and *so*) is used as a pre-closing devise (sic), offering its recipient a chance to reinstate an earlier or unexpected topic, or to open another round of talk, prior to conversational closure. (Schiffrin 1987: 102)

Schiffrin's research is based on a conversational (AE) corpus. This gives it the validity that comes from the real use of examples though there always

has to be a question of how far AE patterns of language use are mirrored in BE speech. Schiffrin also found that -

> Use of *well* with answers is sensitive to the linguistic form of the prior question. (...) Answers were marked with *well* more frequently after WH-questions [21%)][compared to appearing] after yes-no questions [10%]. This difference (…) suggests that when the conditions for propositional suffiency of an answer have been relatively delimited by the form of the prior question, well is not as useful for marking the answer as a coherent response. (Schiffrin 1987: 104)

She further found that *well* is used as a face-saver. This, too, is relevant in the light of my spoken Liverpool English research:

> My results thus far suggest that *well* is more likely to be used when a respondent cannot easily meet a conversational demand for a response because the idea content of his or her answer will not fit the options just opened by a prior question.
> Discourse markers tend to occur at the beginning of a turn or utterance. They signal interactively how the speaker plans to steer the dialogue. (Schiffrin 1987: 114)

Like Schiffrin before her, Jucker (1993) highlights the positioning of *well* within a conversation as well as its function as a face-saver device:

> The discourse marker well is used to indicate a shift in the relevant context. It is not the context as set up by the immediately preceding utterance which is most relevant, because the speaker wants to embark on a new topic; because there is a change in perspective (as in reported direct speech); or because it turns out that the interlocutor uses a slightly different context (contradicting assumptions, missing assumptions, etc.). These situations are often face threatening for one of the participants, but *well* does not

> directly signal the face-threatening act but the shift in the relevant context. Therefore it can occur even if there is no conceivable FTA [Face Threatening Act] (reported direct speech); and it does not occur with every single FTA. (Jucker 1993: 452)

The important point here is that *well* is employed even in situations that are not perceived as *face-threatening*, but that *well*, acting like a pacifying formula, already seems to be used to pre-empt any conceivable threat a listener may perceive.

These insights have not been reversed by more up-to-date corpus linguistic research, as this point is also highlighted in the Longman Grammar of Spoken and Written English:

> *Well* has varied uses, but overall has the function of a "deliberation marker", indicating the speaker's need to give brief thought to the point at issue. *Well* also often marks a contrast, (…) and it can introduce an indirect or evasive answer.
>
> (Biber et *al.* 2002: 450)

Michael Hoey, in describing discourse markers to English learners, confirms the above. *Well* is used in spoken English to indicate disagreement. This indicates its use as a face-saving device.

> WELL (...) is used at the beginning of a speaking turn (...) You start your reply with WELL when answering someone who has just said something factually incorrect or made a false assumption. (...)

> You can also begin your answer with WELL if someone asked you a question which assumes something that is not in fact true. (...)
>
> Another use of WELL is to round off a topic near the end of a conversation.
>
> (Macmillan Dictionary. Section L14 (Hoey): 2002)

Very important for my research is the position within an utterance. The comparison below will reveal that position is of importance for *well* clusters in particular.

The following discussion will show in how far positioning of *well* within an utterance is relevant to my research.

| Word / Collocate | Total SCO | % | Total MAC | % | LL | BNC/C | % |
|---|---|---|---|---|---|---|---|
| **WELL** | **328** | **0.3** | **36,869** | **1.1** | **1059.7** | **35,806** | **0.89** |
| PAUSE / - | 42/ 61 | 12.8 / 18.6 | 18,110 | 49.1 | 23.88 | | |
| AS | 98 | 29.9 | 3,701 | 10.0 | 82.50 | 3,716 | 10.4 |
| I | 75 | 22.9 | 13,117 | 35.6 | 16.95 | 11,083 | 31.0 |
| YOU | 64 | 19.5 | 8,592 | 23.3 | 2.13 | 6,864 | 19.2 |
| A | 54 | 16.5 | 3,502 | 9.5 | 13.56 | | |
| YEAH | 49 | 14.9 | 2,979 | 8.1 | 15.07 | 2,484 | 2,484 |
| THE | 43 | 13.1 | 5,344 | 14.5 | 0.44 | | |
| IT | 38 | 11.6 | 8,105 | 22.0 | 19.39 | 5,029 | 14.0 |
| TO | 30 | 9.1 | 3,486 | 9.5 | 0.03 | | |
| THAT | 28 | 8.5 | 5,574 | 15.1 | 11.09 | 3,098 | 8.7 |
| KNOW | 27 | 8.2 | 1,882 | 5.1 | 5.24 | | |
| WAS | 27 | 8.2 | 1,890 | 5.1 | 5.15 | | |
| AND | 26 | 7.9 | 3,240 | 8.8 | 0.28 | | |
| IN | 25 | 7.6 | 1,868 | 5.1 | 3.62 | | |
| HE | 24 | 7.2 | 3,346 | 9.1 | 1.19 | | |
| IS | 22 | 6.7 | 1,797 | 4.9 | 2.00 | | |
| DO | 21 | 6.3 | 1,899 | 5.2 | 0.92 | | |
| EHM | 21 | 6.3 | 1,018 | 2.8 | 11.30 | | |
| EH | 14 | 4.3 | 1,523 | 4.1 | 0.01 | | |
| IT'S | 19 | 5.8 | 2,179 | 5.4 | 0.01 | | |
| THERE | 19 | 5.8 | 1,713 | 4.8 | 0.85 | | |
| NOT | 18 | 5.5 | 1,491 | 4 | 1.50 | | |
| OF | 18 | 5.5 | 1,830 | 5 | 0.17 | | |
| THEY | 18 | 5.5 | 3,221 | 8.8 | 4.54 | 2,081 | 5.8 |
| LIKE | 16 | 4.9 | 954 | 2.6 | 5.20 | | |
| OH | 16 | 4.9 | 2,717 | 7.4 | 3.12 | | |
| REALLY | 15 | 4.6 | 490 | 1.3 | 15.56 | | |
| SO | 15 | 4.6 | 1,243 | 3.4 | 1.25 | | |
| ME | 14 | 4.3 | 713 | 2.4 | 6.77 | | |
| BE | 12 | 3.7 | 1,277 | 3.4 | 0.04 | | |
| HE'S | 12 | 3.7 | 578 | 1.6 | 6.54 | 772 | 2.2 |
| I'M | 12 | 3.7 | 847 | 2.3 | 2.21 | | |
| NOW | 12 | 3.7 | 481 | 1.3 | 9.18 | | |
| THAT'S | 12 | 3.7 | 2223 | 6.2 | 3.54 | 2,190 | 6.1 |
| WHAT | 12 | 3.7 | 2,068 | 5.7 | 2.52 | | |
| FROM | 11 | 3.4 | 236 | 0.64 | 18.31 | 210 | 0.6 |
| HAVE | 11 | 3.4 | 1,840 | 5 | 1.98 | | |
| JUST | 11 | 3.4 | 1,063 | 2.7 | 0.24 | | |
| THEM | 11 | 3.4 | 781 | 2.4 | 1.98 | | |
| NO | 10 | 3.0 | 1,774 | 4.8 | 2.42 | | |
| SAID | 10 | 3.0 | 2,334 | 6.3 | 6.87 | 2,453 | 6.9 |
| SHE | 10 | 3.0 | 2,233 | 6.2 | 5.96 | 1,855 | 5.2 |

Scouse < 5% less  Scouse <2% more  Scouse >2% more  occ. than MAC
(*indicates a break in speech flow)

Table 1: SCO-MAC *well* collocate comparison
(with key BNC/C collocates' figures included).

### 7.2.2 WELL collocates

Being mostly employed as a *discourse marker, well* is significantly rarer in its use amongst Scouse speakers than amongst other UK speakers as represented by the MAC corpus. In MAC, *well* appears once in every 90 words spoken. This is 3.6 times more often than in SCO, where *well* occurs only once in every 323 words spoken. Furthermore, *well* also occurs nearly three times as often in BNC/C as in SCO.

Table 1 illustrates that *well* stands out by being markedly different when its occurrence in SCO is compared with that in MAC (and in BNC/C where the proportional figures show mostly strong agreement with the proportional percentages recorded in MAC). We find seven out of the top ten most frequent collocates of *well* in SCO occur with significantly different proportional frequencies from those in MAC. Although *well* appears in the corpora mainly as a *discourse marker*, it has also other uses. As it would be arbitrary to leave these out, there will also a discussion of the homonym *well* where there is high-frequency use and a notable margin of difference in use between the corpora. The one particular function shows *well* in its adverbial use with the collocate *as* (29.9% of all collocates of *well* in SCO; 10% in MAC; 10.4% in BNC/C). *Well* as a discourse marker appears with different frequencies of use with the collocates *yeah* (14.9% in SCO, 8.1% in MAC and 6.9% in BNC/C) and *from* (3.4% in SCO; 0.64% in MAC; 0.6% in BNC/C).

At the same time, however, *I, it* and *that* are three collocates found to be significantly less frequent in SCO than in MAC.

### 7.2.3 Markers of hesitation used with WELL

The Liverpool English corpus (SCO) picks up certain paralinguistic features by lexicalising them – typing in (pause) for longer pauses; (laughs) and (laughter) for audible laughs. It also records hesitation ((*eh); (ehm)*). Since the purpose of my research is comparison, all this is of little value when no other corpus gives any indication of paralinguistic features. MAC transcribers appear also to indicate pauses, in their case by punctuation (commas and full stops) as well as with hesitation markers (*ehm; erm*) and these features are picked up by *WordSmith*.

Given the issues involving recording and discussing paralinguistic features, all the findings presented below have to be seen as a rough approximation. There is nevertheless an indication that different patterns appear in the two corpora. *Well* interacts strongly with paralinguistic features. This supports Lawrence Schourup's thesis of *well* being a gesture.

Based on the figures presented in (the second line of) Table 1, one clear feature of *well* is that it tends to be followed by a pause. It is the feature that is most likely to follow *well* – more likely than any word. This is true in both SCO and MAC. A pause can either indicate a clause-end

(and the wish to pass the conversational turn to another speaker), and / or hesitation, or it can be a hedging device (gaining time).

Through this, the different uses of *well* in the two corpora is shown: a pause appears in about half of all cases in MAC and in nearly 1/3 of all cases in SCO. The proportional difference of SCO to MAC is 3:5 (30.4% : 49.1%). Though a pause is the most frequent collocative event of *well*, Table 1(l.2) indicates that pauses co-occur with *well* about 20% less often in SCO than in MAC.

The picture is similar when we look at another hesitation marker, the particle (or sound) *ehm* (or *erm*). It appears as a collocate of *well* in 6.1% of all cases in SCO but in less than half as many cases - 2.8% - in MAC.

This might turn out to be important – and an indication that paralinguistic features play part of lexical priming as well. On the other hand, it might be merely the effect of different standards of transcription.

### 7.2.4 WELL two-word clusters

#### 7.2.4.1 WELL two-word clusters by proportional frequency

Only a small number of clusters (two-word clusters and contractions of three-word clusters) can be found with sufficiently high numbers of occurrence. The use of *well* is dominated by two chunks (*as well* and *well I* – discussed below) that have the highest frequencies in both corpora by

a fair margin. All the other chunks are not very frequent in their absolute use as *well* clusters.

| Cluster | SCO total | % | MAC total | % | Log-Likelihood | BNC/C total* | % |
|---|---|---|---|---|---|---|---|
| AS WELL | 90 | 27.0 | 3,183 | 8.6 | 83.60 | 3,260 | 9.1 |
| WELL I | 18 | 5.5 | 5,995 | 16.3 | 31.36 | 5,290 | 14.8 |
| WELL - I (combined with above) | 6 (24) | 1.8 (7.3) | | | (20.20) | | |
| WELL YOU | 13 | 4.0 | 2,437 | 6.6 | 4.03 | 2,223 | 6.2 |
| WELL YEAH | 8 | 2.4 | 399 | 1.0 | 4.05 | 727 | 2.0 |
| WELL – YEAH (combined with above) | 5 (13) | 1.5 (3.9) | | | (14.63) | | |
| YEAH WELL | 15 | 4.3 | 1636 | 4.5 | 0.01 | 2,036 | 5.7 |
| WELL IT'S | 6 | 1.8 | 1,092 | 3.0 | 1.63 | 1,382 | 3.9 |
| WELL THERE | 6 | 1.8 | 501 | 1.5 | 0.48 | 1,675 | 4.7 |
| OH WELL | 6 | 1.8 | 1,413 | 3.8 | 4.24 | 1,675 | 4.7 |
| WELL HE | 6 | 1.8 | 1,136 | 3.1 | 1.94 | 927 | 2.6 |
| EHM - WELL | 5 | 1.5 | | | n/a | 390** | 1.1 |
| WELL THAT | 4 | 1.2 | 2,017 | 5.5 | n/a | 560 | 1.6 |

Table 2: Most frequent 2 word WELL clusters in SCO, proportional % for MAC & BNC/C equivalents.
(* BNC/C WELL total: 35,806 words (0.89% of corpus total)**ER WELL)

Table 2 looks at the highest occurring 2w *well* clusters in SCO, MAC and BNC/C. The significance test has been done by comparing SCO occurrence patterns with MAC occurrence patterns. While the majority of 2w clusters does not diverge significantly (like *well there* or *well he*) there are a number of clear exceptions.

### *7.2.4.2* WELL 2w clusters with different proportional frequencies and uses

The more carefully the *well* clusters in the two corpora are compared, the more striking the differences seem to be. This is shown in the investigation of those two-word clusters where there is a marked difference in proportional occurrences. For this, I shall disregard all those clusters that appear less often than around 1% in SCO[120]: Consequently, the focus will be on the following: *as well, well I, oh well, yeah well* and *well yeah*. As Table 2(b) shows, three of these five clusters are amongst the most frequent common clusters by a fair margin.

| Cluster | SCO total | % | MAC total | % | Log-Likelihood | BNC/C total* | % |
|---|---|---|---|---|---|---|---|
| AS WELL | 90 | 27.0 | 3,183 | 8.6 | 83.60 | 3,260 | 9.1 |
| WELL I | 18 | 5.5 | 5,995 | 16.3 | 31.36 | 5,290 | 14.8 |
| WELL - I (combined with above) | 6 (24) | 1.8 (7.3) | | | (20.20) | | |
| WELL YEAH | 8 | 2.4 | 399 | 1.0 | 4.05 | 727 | 2.0 |
| WELL – YEAH (combined with above) | 5 (13) | 1.5 (3.9) | | | (14.63) | | |
| YEAH WELL | 15 | 4.3 | 1636 | 4.5 | 0.01 | 2,036 | 5.7 |

Table 2(b)[121]: Divergence of use in WELL 2w clusters SCO compared to MAC and BNC/C.

*As well* is proportionally higher in occurrences in SCO compared to MAC and BNC/C, while *well I* is used proportionally far more often in MAC and BNC/C[122].

[120] Unless they were to be found to be very frequent in MAC.
[121] Table 2(b) is an excerpt of Table 2
[122] It should be noted at this point that, apart from the low-frequency 2w clusters *well yeah* and *well that*, BNC/C proportional usage is always similar to MAC usage.

While the 2w cluster *yeah well* appears proportionally with the same frequency in all three corpora, we can also find the inverted form – *well yeah*. The latter presents differing proportional occurrence frequencies. The cluster accounts for a total of 3.9% of occurrences of *well* in SCO (2.4% - *well yeah* ; 1.5% *well* (short pause) *yeah* ). The total in MAC is only 1 % (in BNC/C it is 2%).

The one fixed part of *well yeah* in SCO is that it always starts a clause and nearly always a turn. In MAC, however, it can be found in any position in a turn – though it is, here too, mostly a clause-starter.

If we just look at the proportional frequency (and statistically valid divergence) of *yeah well,* it appears to be used in the same way in all three corpora. Comparing the concordance lines, however, shows a marked difference of usage between SCO and MAC.

In the 20 uses in SCO of *yeah well* (5.8% of all uses of *well*[123]), the functions are split in Table 3:

| word used 1 | turn position | word used 2 | turn position | occ. |
|---|---|---|---|---|
| **YEAH** | end of turn | **WELL** | new turn being taken | 10 |
| **YEAH WELL** | new clause, new turn | | | 6 |
| **YEAH WELL** | new clause | | | 4 |

Table 3: Turn-taking pattern in WELL occurrence in SCO

As pointed out earlier, Jucker, Schiffrin, Biber *et al.* and Hoey all have noted that *well* is often used to indicate turn-taking. Table 3 demonstrates how SCO speakers employ the word *yeah* in conjunction

[123] This includes occasions where myself is the speaker to end a turn with YEAH, and this is followed by WELL by a Scouse speaker.

with *well* in a characteristic way: the term *well*, following *yeah*, introduces a new clause in every single case. In half of all cases in SCO the first speaker gives up a turn ending the utterance in *yeah*, while the next speaker seems to be reluctant to have his / her turn at this stage and therefore starts with *well* . Whatever the formation, *yeah well* is mostly followed by a pronoun, either *"I"* (3 times) or *you* (5 times). Consequently, the cluster most frequently occurring incorporating *yeah well* is *yeah well you* (1.5% of all uses of *well*). This typically (four times out of five) is spoken by more than one person in an exchange. (YEAH. / *well* YOU )

In comparison, amongst the total of 1852 occurrences of *yeah well* in MAC, only 1/3 are split by a change of turn taking (660 occurrences). This conclusion is based on where there is a *full stop* in the transcript. I have however no access to MAC transcription conventions to confirm this.

There appears to be no fixed pattern as regards what follows *yeah well* in MAC. Spot-checks on the position of *yeah well* seems to indicate that *yeah well* occurs most frequently mid-turn, rather than as clause-starter. In MAC, the most frequent clusters are *yeah well I* (289 occ. – 0.8%); *it's*, *yeah*, *well* (150 occ. – 0.4%); *yeah well it's* (120 occ) and *yeah well that's* (105 occ). There is, however, little evidence of the three-word cluster found in SCO, discussed above.

We turn now to the two most frequent clusters in both corpora: *as well* and *well I*. Though amongst the most frequent clusters in the three

corpora, the proportional usage shows how differently *well* is employed in the different corpora. While the former indicates adverbial use of *well,* the latter is the *well* homonym, functioning as a *discourse marker*. *As well* will be included in this discussion because of its prominence in all the corpora and because it shows divergent use in SCO when compared to MAC (and BNC/C).

The cluster *as well* is used on nearly a third of the occasions when the term *well* occurs in SCO. There is not one cluster in MAC that is remotely as frequent. *As well* accounts for 8.6% of all uses of *well* in MAC. The most frequent cluster in MAC with *well* is *well I* which accounts for 16.8% of all uses of *well* (just over one in six of the occurrences of *well*).

By contrast, *well I* accounts for only 7.3% of all uses of *well* in SCO.

In SCO, *as well* is used with the meaning of *also*, in order to provide added information, though there are two cases (out of 90) where it is an elliptical form of *just as well*.

Checking on *as well* in MAC, we find that, although there is a difference, already noted, in the frequency of use, MAC speakers are like SCO speakers in employing *as well* in the majority of cases with the notion of *also*, too. ("*You can buy it as well").*

There are, as Table 4 highlights, within this set of clusters that incorporate *as well*, a number of clusters that fulfil quite a different function. The divergence of use between SCO and MAC is striking with regard to these. For example, in MAC there is the cluster *(you) might as well*. This is one of the most frequent 3-word clusters in MAC, accounting

for 0.78% of all uses of *well*, but it does not at all appear in SCO. Likewise, the most frequent chunk in SCO, *as well you*[124] (as in "*And I liked it that way as well you know*") is extremely marginal in MAC.

The position of *well I* is the same in utterances. *Well I* in SCO occurs at the start of a clause in 71.8% of all uses (23 out of 32). In the 5821 concordance lines of *well I* in MAC, a random sample of 5% of these lines found the same proportion of *well I* at the start of a clause. Marked divergence is found, however, when the most frequent clusters incorporating *well I* are compared. The most frequent clusters in SCO are the following:

| Cluster | Frequency | % of *WELL* uses |
|---|---|---|
| AS WELL I | 6 | 1.8 [23.25] |
| AS WELL I WAS | 2 | 0.6 |
| WELL I HAVE / WELL I'VE | 5 (3+2) | 1.5 [1.12] |
| WELL I MEAN | 4 | 1.2 |
| WELL I WILL / WELL I'LL | 3 (2+1) | 0.9 |

Table 4: SCO clusters incorporating *WELL I*[125]

Yet again, the most frequent 3-word cluster in SCO is marginal in MAC[126] (*as well I* – 60 occ. – below 1%) while the most frequent 3-word cluster in MAC – *well I don't* (488 occ – 1.3%) does not at all appear in SCO (see also Table 6 below to compare BNC/C figures).

[124] *As well you* appears 5 times in SCO, 71 times in MAC. Log-Likelihood figure is 11.69, meaning that the divergence is over 99.9% siginificant.
[125] In brackets: breakdown of the different forms found. Log-Likelihood test figures in square brackets where applicable.
[126] Table 5 shows that WELL *with* "I" usage pattern is similar in BNC/C to that into MAC and dissimilar to that in SCO.

### 7.2.5 WELL - usage in three word clusters

Finally, we look at those 3-word clusters with *well* that have, so far, not been discussed as extensions of two-word chunks.

As before, because of low numbers, no final conclusions can be drawn. Only tendencies can be described. The clusters shown in Table 5 below shows the differences in usage.

| SCO cluster | tot. | % | MAC cluster | total | % | LL |
|---|---|---|---|---|---|---|
| WELL YOU KNOW | 7 | 2.1 | WELL YOU KNOW | 309 | 0.9 | 4.53 |
| AS WELL YOU | 5 | 1.5 | AS WELL YOU | 71 | 0.18 | 23.25 |
| AS WELL I | 6 | 1.8 | AS WELL, I | 35 | >0.1 | 11.69 |
| WELL I HAVE/ I'VE | 5 | 0.9 | WELL I'VE | 335 | 0.9 | 1.12 |
| YOU AS WELL | 4 | 1.2 | YOU AS WELL | 24 | >0.1 | |
| WELL THERE IS | 3 | 0.9 | WELL THERE'S | 287 | 0.78 | |
| WELL I MEAN | 3 | 0.9 | WELL I MEAN | 443 | 1.1 | |
| WELL I THINK | 0 | | WELL I THINK | 394 | 1.0 | |
| I SAID WELL | 0 | | I SAID WELL | 449 | 1.1 | |

Table 5: 3w *Well* clusters most divergent SCO:MAC

| BNC/C WELL + AS | occ. | % | BNC/C WELL + „I“ | occ. | % |
|---|---|---|---|---|---|
| MIGHT AS WELL | 119 | 0.33 | WELL I DON'T | 166 | 0.46 |
| AS WELL AS | 61 | >0.3 | I SAID WELL | 131 | 0.37 |
| AS WELL AND | 51 | >0.3 | WELL I MEAN | 109 | 0.30 |
| YOU MIGHT AS | 43 | >0.3 | WELL I THINK | 106 | >0.3 |
| I MIGHT AS | 41 | >0.3 | I THOUGHT WELL | 105 | >0.3 |
| AS WELL YEAH | 37 | >0.3 | WELL I'M NOT | 39 | >0.3 |
| AS WELL YOU | 37 | >0.3 | WELL I KNOW | 38 | >0.3 |
| IT AS WELL | 35 | >0.3 | WELL I DIDN'T | 34 | >0.3 |

Table 6: WELL *with* AS and WELL *with* „I“ 3w clusters in BNC/C.

We find that there are four 3w *well* clusters where SCO figures are high enough to test divergence for statistical validity. The highest occurring - *well you know* and the lowest occurring - *well I have / I've* do not significantly diverge in their use where SCO and MAC are compared. We find, however, that the medium-high 3w clusters with *well* and *as* occur with significantly higher proportional frequencies of use: *as well I* and, even more so, *as well you* are significantly more frequent in SCO than in MAC.

### 7.2.6 WELL Conclusions

As a conclusion, it can be maintained that the use of WELL represents a good example of how differently an item is employed by the speakers represented in SCO and MAC.

Initially, we saw that the most frequent collocate in SCO is *well* with *as,* which is a collocate in nearly one third of all occurrences of *well*. In MAC, the most frequent collocate is *well* with *I* and it is used only one-sixth of all the times *well* is spoken. Interestingly, the most common collocate of *well* in SCO appears almost exclusively (in 90 out of 98 occurrences) in the 2w cluster *as well*.

In the corpora compared, *well* combines freely, so that two-word clusters rather than three word-clusters can be compared. In fact, the 3w cluster *well you know* - the most frequent three-word cluster in both corpora – which incorporates the 2-word-cluster *well you* is the only

cluster of relatively high frequency that appears to have the same function and nesting in both SCO and MAC. Notwithstanding, *well you know* appears proportionally three times as often in SCO compared to MAC.

The most important finding is, however, that even those *well* 2w clusters in SCO and MAC that do not differ strongly in their proportional frequency of occurrence diverge nevertheless strongly in their patterns of use.

This is shown by the occurrence pattern of the cluster *yeah well. Yeah well* is often found mid-turn in MAC; in SCO, however, *yeah* ends the turn for one speaker and the next speaker picks the conversation up by following on with *well* (which, consequently, is seen by the concordancing software as the cluster *yeah well*). Similar divergences in use have been found for *well you, well it's* and *well I*, too. This can be seen as an indication that *well* appears in some cases with different nesting in SCO compared to MAC. It also indicates that there appears some evidence that priming can be found beyond the unit of single-speaker utterances and can also be seen as covering two-speaker utterances.

# Chapter 8 VERY and REALLY uses compared

## 8.1 VERY – a rare indicator

*Very* is widely perceived to be as a prime example of a word used as an intensifier in spoken English. Unlike other terms discussed above (*just, like, well*, etc.), *very* is not seen as a discourse particle but as having the specific role of intensifying any given utterance. Leech and Svartvik ([1975] 1992: 99) call *very* a *degree expression*.[127] They also note that "you can also intensify meaning by repeating the word *very"* (p.103) and say *very* is used to give emotive emphasis (p.138)[128]. As part of the *Cobuild Series* (Sinclair *et al.*: 1998b) where BoE corpus-based pattern grammar is described, *very* is defined as: "a grading adverb, part of the ' fairly' and 'extremely' group. These adverbs indicate that someone or something has a lot or a little of a quality." (p.353) *Very* is also part of "the 'absolute' and 'mere' group – these adjectives are used to emphasise the quality of something" (i.e. … *the very thought of Laura* …)(p.367). This shows agreement with Leech and Svartvik's view, too.

[127] See Paradis, C. (1997) for a critical discussion about how appropriate this terminology is.
[128] Biber *et al.* (2002) concur with Leech and Svartvik fully, adding the *very* and *so* are equally well used in both conversational & academic corpora.

Partington (1993) sees *very* as highly de-lexicalised (which would bring it in line with *really)*:

> *Very* is highly delexicalized because it combines very widely indeed and is the intensifier with the least independent lexical content. (Partington 1993: 183)

We have come across a number of core words in this chapter that "combine widely" and will see in section 9.2.3 how far our data matches Partington's claim.

That the use of *very* appears to be age-dependent is also highlighted by a study undertaken in New Zealand, using naturally occurring language collected from amongst school children:

> *Very* was reported as a booster from just seventeen schools [out of 150 – MP-S]. It occurred with the following adjectives: *bad, difficult, embarrassed, embarrassing, fun, good, hard, mad, not _ good, not _ well, squashed, shameful, sore, ugly*. *Not _ well* and *not _ good* were the most frequent collocations. Again, notice that the majority of these had negative connotations. It is worth pointing out that in the two most frequent collocations, *very* is not entirely clearly a booster. The expression *not very good* does not mean that the quality of being very good is absent, so much as that the quality of being good is not present to any significant extent. Several of the reports of *very* were marked as the contributions of nonnative speakers. (Bauer & Bauer 2001: 250)

That there seems to be a stronger preference by non-native speakers to use *very* could be seen as yet another indicator that *very* tends to be used more by older speakers, as L2 speakers (who receive formal instruction) tend to learn first more formal and dated forms of any modern language.

We will discuss in how far UK and Liverpool English corpora concur with the collocates presented by Bauer and Bauer.

Bauer and Bauer's findings with regards to *very* being seemingly age-related are apparently supported by work on UK corpora:

> Teenagers of the nineties in London use the degree modifiers found to be most frequent in LLC [London Lund Corpus ; 1975 - MP-S] to a much lesser extent. [*very* is by far the highest occurring *degree modifier* in LCC – MP-S] In fact, only 22% of the total number of degree modifiers in the two corpora occur in COLT [The Bergen *Corpus* of *London Teenage Language*; collected in 1993. MP-S]. (Paradis 1998: 5)

Ito & Tagliamonte researched the use of intensifiers (boosters) in different age groups amongst speakers in the City of York, UK (2003). While, in their recorded speech samples, *very* is the most frequent intensifier (38,0% of all the intensifiers used in York) overall, Ito & Tagliamonte also highlight that (…) "*very* is the most common [intensifier] amongst the older speakers". (Ito & Tagliamonte 2003: 257)[129] I will discuss what that means in the context of this thesis in 8.1.1 below.

[129] They also point out that *very* has been classified as Standard (US) English – in opposition to, for example, *real. Real* had been seen as vulgar.

### 8.1.1 VERY – a signifier of speaker age in SCO?

Ito and Tagliamonte[130] point out that intensifiers are subject to rapid change:

> The most frequent intensifiers, however, are shifting rapidly. *Very* is most common, but only among the older speakers. In contrast, *really* increases dramatically among the youngest generation. (Ito & Tagliamonte 2003: 257)

In that paper, recordings were taken a couple of years before my sample. The age-range is between 17 and 66+ (a median age cannot be inferred from the figures given). For my SCO corpus, informants are aged between 10 and 70. The median age is around 35.

When going through the hits per thousand of *very* in every single file, I find the lowest numbers (that is – the least frequent use) not only amongst the 12-14 year olds[131] (f.: 0.48) and twenty-year olds (m.: 0.92 & 0.98) but also amongst the 30-year olds (m.: 0.68), thirty to fifty year-olds (f./ m.: 0.22) and fifty-year olds (f.: 0.33).

At the same time, the most frequent use of *very* (again per thousand words) can be found amongst twenty year-olds (m.: 2.46) and forty-year olds (f.: 3.19). The oldest informant, a seventy-year old male is right in the middle with 1.58 uses of *very* per thousand. These results show that *very* is not very much used in SCO, regardless of age or sex.

[130] I chose Ito and Tagliamonte's text for the direct comparison here because the data were collected at about the same time, with informants of similar age groups and, most importantly, amongst native UK speakers of a city other than Liverpool.
[131] f. – female; m. – male.

This means that, by the criteria listed by Ito and Tagliamonte, Liverpudlians in everyday speech do not employ standard English intensifier patterns and do use *very* in a way more associated with young people, whether or not the speakers are themselves old (in particular when compared to another intensifier discussed – *really*):

> Finally, (our figures show) that among the youngest generation, there is an exponential increase in use of *really* across nearly all categories. Moreover, there is spread to an additional category, colour. In at least four (value, human propensity, dimension, and physical property), use of *really* is double that of *very*.
>
> (Ito & Tagliamonte 2003: 271)

In the SCO, the differential is even more marked across the board when the use of *very* and *really* is being compared. 126 occurrences of *very* are even less than half than the 264 occurrences of *really*.

### 8.1.2 VERY: frequent collocates

The most obvious difference of the use of the word *very* that can be found (and it is more marked a difference in comparison to MAC than in comparison to BNC/C) is that Scouse speakers do not employ the term *very* (sometimes defined as a "booster") all that often. Focussing on the most frequent collocates of *very* in all three corpora, the differences are small.

| Word | SCO occ. | % | Rk | MAC occ. | % | Rk | LL | BNC/C occ. | % | Rank |
|---|---|---|---|---|---|---|---|---|---|---|
| ***VERY**** | **153** | **0.1** | | **24,671** | **0.83** | | **929.83** | **7565** | **0.2** | |
| A | 24 | 15.7 | 4 | 4,061 | 16.2 | 4 | 0.06 | 1005 | 15.4 | 5 |
| IT'S | 20 | 13.0 | 5 | 4,051 | 16.1 | 5 | 1.12 | 893 | 13.7 | 7 |
| AND | 19 | 12.4 | 6 | 4,072 | 16.3 | 3 | 1.69 | 812 | 12.4 | 8 |
| IS | 17 | 11.0 | 7 | 2,009 | 8.1 | 12 | 1.47 | 517 | 7.9 | 17 |
| IT | 17 | 11.0 | 7 | 4,825 | 19.3 | 2 | 6.59 | 1038 | 15.9 | 4 |
| YOU | 16 | 10.5 | 8 | 3,307 | 13.3 | 7 | 1.07 | 1064 | 16.3 | 2 |
| GOOD | 15 | 10.0 | 9 | 1,588 | 6.4 | 16 | 2.30 | 920 | 14.1 | 6 |
| I | 15 | 10.0 | 9 | 3,221 | 12.9 | 8 | 1.35 | 1046 | 16.0 | 3 |
| THE | 13 | 8.5 | 10 | 3,821 | 15.3 | 6 | 5.75 | 752 | 11.5 | 9 |
| WAS | 13 | 8.5 | 10 | 1,897 | 7.6 | 13 | 0.12 | 617 | 9.5 | 14 |
| EHM | 12 | 7.8 | 11 | n/a | n/a | n/a | | n/a | n/a | n/a |
| ER | | | n/a | 1,507 | 6 | 18 | | n/a | n/a | n/a |
| ERM | | | n/a | 1,152 | 4.6 | 25 | | 179 | 2.7 | 50 |
| OF | 12 | 7.8 | 11 | 2,112 | 8.5 | 11 | 45.89 | 378 | 5.8 | 20 |
| THAT | 12 | 7.8 | 11 | 2,962 | 11.9 | 9 | 80.09 | 581 | 8.9 | 16 |
| TO | 11 | 7.2 | 12 | 2,880 | 11.5 | 10 | 80.32 | 638 | 9.8 | 12 |
| YEAH | 10 | 6.5 | 13 | 595 | 2.4 | 43 | 1.84 | 324 | 5.0 | 24 |
| JUST | 9 | 5.9 | 14 | 307 | >2 | 72 | 0.19 | 127 | 1.9 | <50 |
| WELL | 9 | 5.9 | 14 | 1,200 | 4.8 | 23 | 20.29 | 698 | 10.7 | 10 |
| NOT | 8 | 5.2 | 15 | 990 | 3.9 | 27 | 15.41 | 635 | 9.7 | 13 |
| BUT | 7 | 4.6 | 16 | 1,249 | 5 | 20 | 27.42 | 447 | 6.9 | 18 |
| HE'S | 7 | 4.6 | 16 | 1,158 | 4.6 | 24 | 23.95 | 221 | 3.4 | 33 |
| ISN'T | 7 | 4.6 | 16 | 169 | >2 | n/a | 1.43 | 114 | 1.7 | <50 |
| KNOW | 7 | 4.6 | 16 | 595 | 2.4 | 42 | 5.21 | 289 | 4.4 | 27 |
| REALLY | 7 | 4.6 | 16 | 278 | >2 | 76 | 0.00 | 132 | 2.0 | <50 |
| THEY | 7 | 4.6 | 16 | 1,627 | 6.5 | 15 | 1.06 | 346 | 5.3 | 22 |
| WHICH | 7 | 4.6 | 16 | 508 | 2 | 49 | 3.45 | 61 | 0.9 | >50 |
| NICE | 6 | 3.7 | 17 | 654 | 2.6 | 38 | | | | |
| SHE | 6 | 3.7 | 17 | 691 | 2.7 | 34 | | | | |
| THERE | 6 | 3.7 | 17 | 745 | 3 | 33 | | | | |
| THIS | 6 | 3.7 | 17 | 832 | 3.3 | 30 | | | | |
| DON'T | 5 | 3.3 | 18 | 297 | >2 | 73 | | | | |
| HAVE | 5 | 3.3 | 18 | 800 | 3.2 | 32 | | | | |
| HE | 5 | 3.3 | 18 | 1,158 | 4.6 | 24 | | | | |
| IN | 5 | 3.3 | 18 | 1,746 | 7.2 | 14 | 55.38 | 344 | 5.4 | 21 |
| MUCH | 5 | 3.3 | 18 | 1,506 | 6 | 19 | 44.91 | 603 | 6.2 | 15 |

Table 1: VERY top 18 collocates in SCO and the figures for those collocates in MAC and BNC/C.

* percentage here refers to VERY as part of corpus total

Figures in ***blue*** highlight proportionally higher use in SCO vs. MAC; figures in ***purple*** proportionally higher use in MAC vs. SCO.

Table 1 shows the 18 most frequent collocates occurring in SCO in direct comparison to their occurrence in MAC and BNC/C.

Looking at the statistically highly valid divergences only, we are presented with a curious picture. *Very*, as such, appears proportionally significantly less often in SCO than in MAC, yet the majority of *very* collocates appear with similar frequencies when we look at their occurrences in relation to the total number of *very* in their respective corpora. Significant differences are only found in the medium-high and low frequency collocates of *very*: *well, not* and *he's* appear significantly more frequently in SCO than in MAC, yet most collocates are significantly less often found (proportionally) in SCO than in MAC: *of, that, to, in* and *much*.

### 8.1.3 VERY frequent short clusters

Unlike collocates, where a word is found within 5 words either side, two-word clusters are fixed in their position directly to the right or left of the target word. This section focuses on 2w clusters, as there are only very few 3w clusters in SCO that are recorded more than twice. This would support Partington's claim about "*very* combining widely" (see above).

Table 2: Most frequent 2w VERY clusters in SCO, MAC and BNC/C (BNC/C in order of proportional frequency of occ.)

| ***VERY SCO* top 2w clusters** | **occ.** | **%** | ***VERY MAC* equiv. 2w clusters** | **occ.** | **%** | **Log-Likelihood** | ***VERY BNC/C* top 2w clusters** | **occ.** | **%** |
|---|---|---|---|---|---|---|---|---|---|
| VERY VERY | 18 | 11.8 | VERY VERY | 4626 | 18.8 | 4.57 | VERY GOOD | 919 | 12.1 |
| A VERY | 16 | 10.0 | A VERY | 2654 | 10.8 | 0.01 | VERY VERY | 707 | 9.3 |
| VERY GOOD | 15 | 9.8 | VERY GOOD | 1689 | 6.8 | 1.71 | A VERY | 648 | 8.6 |
| IT'S VERY | 8 | 5.0 | IT'S VERY | 955 | 3.9 | 0.65 | VERY NICE | 614 | 8.0 |
| WAS VERY | 7 | 4.4 | *YOU VERY* | 828 | 3.4 | 0.60 | VERY MUCH | 579 | 7.6 |
| VERY NICE | 6 | 3.8 | VERY NICE | 709 | 2.9 | 0.52 | NOT VERY | 441 | 5.8 |
| VERY CLOSE | 6 | 3.8 | VERY WELL | 798 | 3.2 | 0.21 | IT'S VERY | 441 | 5.8 |
| NOT VERY | 6 | 3.8 | NOT VERY | 542 | 2.1 | 1.66 | VERY WELL | 432 | 5.7 |
| IS VERY | 5 | 3.0 | IS VERY | 763 | 3.1 | 0.01 | WAS VERY | 363 | 4.8 |
| VERY MUCH | 5 | 3.0 | VERY MUCH | 1865 | 7.6 | 4.72 | YOU VERY | 305 | 3.9 |

Table 2: Most frequent 2w VERY clusters in SCO, MAC and BNC/C (BNC/C in order of proportional frequency of occ.)

As Table 2 establishes, the most frequent 2-word cluster, the repetition *very very*,[132] is the most common 2w cluster in terms of the number of occurrences for both SCO and MAC. This appears to support Leech & Svartvik's claim that *you can also intensify meaning by repeating the word* ***very*** (see above). Appendix VI shows that *very* single-word repetition seems to be used in MAC in the colligational environment of *very good*. This is significant as it appears to show that *very good*, in MAC, occurs with single word repetition when it is used in its positive sense. When it appears in its negative sense (*not very good*) there is no *very* repetition (see the MAC concordance excerpts above). In the whole MAC *really* concordance there are only two exceptions: *No it's not very very long* and *now if you're not very very sure*. Amongst SCO speakers, the same seems to hold true as the only time *very very* occurs, it is in an utterance with a positive connotation, but this is, as we have seen, proportionally less frequently occurring than in MAC.

On a broad level (i.e. high / medium / low level of occurrence), the proportion of usage of 2w clusters is similar between the two corpora and statistically valid divergence cannot be found. Indeed, most clusters (e.g. *thank you very much, a very* and *very good*; *or where* very appears in a cluster of several adverbs, for example, *very well* or *very quickly)* appear to be formulaic and differ little in their use across the corpora.

[132] Fung & Carter (p.424: 2007) show that *very very* can be found in CANCODE as a distant collocate of *I mean*.

### 8.1.5 VERY conclusions

While *very* is used proportionally far less by Liverpool speakers than by MAC speakers, *very* is found with mainly the same collocations in both corpora. On the whole the naturally occurring use of *very* shows it to be an integral part of spoken English usage, and the amount of data the small SCO corpus provides show little indication of divergent usage of *very* between the SCO and the comparators.

## 8.2 The use of *REALLY* in casual speech

*Really* has seen much less attention devoted to it than *like* or *well*. In both MAC and SCO, however, it can be found as one of the most frequently used words[133]. Carter and McCarthy (2004) point out that Loewenberg (1982) classes *really* as a signal for hyperbole. Paradis (2003) looks at two spoken corpora (COLT and LLC) to describe use of *really* as threefold:

> Firstly, in the case of truth attesting really, the evidence reflects the [REALITY] concept evoked by *really*. The evidence is factual in nature and *really* is primarily a carrier of a content-based message.(...) *Really* takes scope over propositions in order to provide factual evidence for the truth of the proposition. The content proper of *really* [REALITY] is foregrounded. (…)

[133] cf. Duguid (2009: 4) who notes that proportional use of *really* in British newspapers has doubled between 1993 and 2005. *Really* is, however, proportionally less frequent in BNC/C, an older corpus.

> Secondly, in the case of emphasizing *really*, the evidence of truth is indirect via subjective emphasis made by the speaker. Content-wise *really* is bleached and backgrounded, the schematic function of subjective stance is in the foreground. *Really* takes scope over situations denoted by stative verbs and adjectivals (sic) that may be attitudinally emphasized. (…)
>
> Finally, in the case of *really* as a reinforcer, the evidence of truth conveyed is indirect through *really* as a degree operator. Truth is a prerequisite for the reinforcement of a scalar property. The expression of scalar meanings is always subjective. Similar to the emphasizing reading, the content proper of *really* is bleached and backgrounded, and the schematic function of degree and subjective stance is in the foreground.
>
> (Paradis 2003: 15)

Unfortunately, Paradis does not highlight how these three uses stand in proportional occurrence to each other or how they are distinguished in specific examples. It remains to be seen how far the predominant clusters with the core word *really* in SCO reflect "reality" or "subjective emphasis" or act as a "reinforcer" and how far use of really in SCO would appear to be different from that in MAC (and in BNC/C).

Aijmer and Simon-Vandenbergen put *really* into "the grammatical field of expectation and say that *actually, really*, in fact belong to the core [of the same] lexical field." (Aijmer & Simon-Vandenbergen 2004: 1797). However, actually *is* much less frequent than *really:* In SCO there are only 54 occurrences of *actually* - 0.05% of the corpus total. The figure for BNC/C is similar: 3,309 occurrences equal 0.08% of the corpus total. This is significantly less than any of the other discourse particles discussed here.

Bauer and Bauer (2002) looked at what they call *boosters* (*really* but also, as we have seen, *very* etc.) amongst New Zealand youngsters but note that they found more questions than answers, in particular as the *Wellington Corpus* seems to show a very strong use of *really* while *very* is largely absent. This may be down to a generational shift, however, as other research shows. Ito and Tagliamonte point out that the frequency of *really* use justifies more attention:

> This intensifier [*really]* vies for the highest frequency position; it occurs 30% of the time in our data. (…) It is much less frequent than *very* in Bäcklund's (1973) study of contemporary written American and British English. More recently, Labov (1985:44) observes that *really* is "one of the most frequent markers of intensity in colloquial conversation" in American English. In British English, *really* has not received much attention, but it is reported to be the most common premodifier of adjectives among teenagers in London (Stenström 1999). (Ito & Tagliamonte 2003: 265)

Fung and Carter describe *really* in their pedagogic corpus as "interpersonal, indicating an attitude" (Fung & Carter 2007: 418) and comment:

> *Really* and *obviously* enable the speakers to express certainty towards the propositional meanings of the utterances. (Fung & Carter 2007: 419)

Fung and Carter stand here in agreement with Biber *et al.* who list *really* as a frequent *stance adverbial* and point out that –

> It can be difficult to tell whether a word is a stance adverbial or a circumstance adverbial (…). The adverb *really* is particularly tricky to analyse. Some instances seem clearly to have the epistemic stance meaning of 'in reality' or 'in truth' especially when the adverb is in initial or final position (…) But in medial position, the meaning is less clear. (Biber *et al.* 2002: 385)

When taken with Ito and Tagliamonte's claims above and diachronic developments of use we have described amongst other discourse particles, all of this hints that *really* is another example of the process of bleaching and that it has become more prominent in its use only relatively recently.

### 8.2.1 REALLY and how it occurs

There are a number of lexical items that are used in English to put stress on a particular statement, which is something all speakers presumably need to do. The term *really*, like all the other discourse particles investigated in this chapter, fits this description. Unlike all the other discourse particles discussed here, however, *really* is the only one that is found with proportionally higher recorded use in SCO than in either MAC or BNC/C. *Really* appears proportionally 1.3 times more often in SCO than in MAC, and 1.5 times more often than in BNC/C:

| Core word | SCO frq. | SCO % | MAC frq. | MAC % | BNC/C frq. | BNC/C % |
|---|---|---|---|---|---|---|
| REALLY | 289 | 0.35 | 11,471 | 0.27 | 9,128 | 0.23 |
| Log-Likelihood SCO:MAC | | | 40.88 | | | |

Table 1: *Really* in SCO, MAC and BNC/C occ.

Compared with other discourse markers, *really* is one of the proportionally less frequent words in all three corpora. However, in line with Ito & Tagliamonte's claim, *really* usage appears to be rising over time. This claim seems to be supported by the fact that the BNC/C is based on the oldest recordings, MAC is more recent, while SCO is the most recent.

As the Log-Likelihood test shows, though the difference of proportional frequencies of occurrence of *really* between SCO and MAC are not as great as in the case of *very*, it is a difference of strong significance.

| Rk * | Word (SCO) | Total | % | Word (MAC) | Total | % | Rk | Log-Likelihood | Total BNC/C | % |
|---|---|---|---|---|---|---|---|---|---|---|
| | REALLY | 289 | 100 | REALLY | 11,475 | 100 | | 40.88 | 9,128 | 100 |
| 1 | I | 70 | 24.2 | I | 3,661 | 32.0 | 1 | 5.70 | 2592 | 28.4 |
| 2 | INAUD | 68 | | | | | | | | |
| 3 | IT | 56 | 19.4 | IT | 3,603 | 31.4 | 2 | 15.10 | 1922 | 21.1 |
| 5 | PAUSE | 45 | | | | | | | | |
| 6 | YOU | 41 | 14.2 | YOU | 2,482 | 21.6 | 3 | 8.25 | 1595 | 17.5 |
| 7 | AND | 40 | 14.0 | AND | 1,966 | 17.1 | 6 | 1.91 | 1312 | 14.4 |
| 8 | A | 39 | 13.5 | A | 1,532 | 13.6 | 7 | 0.00 | 1078 | 11.8 |
| 9 | IT'S | 37 | 12.8 | IT'S | 1,615 | 14.1 | 8 | 0.33 | 939 | 10.3 |
| 10 | THE | 36 | 12.5 | THE | 2,177 | 18.98 | 4 | 7.21 | 1038 | 11.4 |
| 11 | IS | 29 | 10.0 | IS | 1,191 | 10.38 | 10 | 0.03 | 717 | 7.9 |
| 12 | WAS | 28 | 10.0 | WAS | 1,114 | 10.00 | 11 | 0.00 | 874 | 9.6 |
| 13 | YEAH | 27 | 9.3 | YEAH | 710 | 6.19 | 19 | 3.90 | 434 | 4.8 |
| 14 | NOT | 26 | 9.0 | NOT | 1,000 | 8.70 | 13 | 0.03 | 710 | 7.8 |
| 15 | KNOW | 25 | 8.7 | KNOW | 721 | 6.21 | 16 | 2.24 | 681 | 7.5 |
| 16 | LIKE | 25 | 8.7 | LIKE | 537 | 4.68 | 20 | 7.53 | 568 | 6.2 |
| 17 | EHM | 22 | 7.6 | ER/ ERM | 608/ 556 | 10.14 | 11 | 1.95 | 268/ 260 | 5.8 |
| 18 | TO | 22 | 7.6 | TO | 1,841 | 16.04 | 8 | 15.61 | 1125 | 12.3 |
| 19 | DON'T | 21 | 7.3 | DON'T | 797 | 6.95 | 15 | 0.04 | 744 | 8.2 |
| 20 | GOOD | 21 | 7.3 | GOOD | 398 | 3.47 | 22 | 8.82 | 455 | 5.0 |
| 21 | THAT | 21 | 7.3 | THAT | 2,116 | 18.4 | 5 | 24.99 | 938 | 10.3 |
| 22 | BUT | 19 | 6.6 | BUT | 809 | 7 | 14 | 0.09 | | |
| 23 | ME | 19 | 6.6 | IN | 718 | 6.2 | 17 | 0.04 | | |
| 24 | THEY | 19 | 6.6 | THEY | 1,007 | 8.7 | 13 | 1.71 | | |
| 25 | OF | 16 | 5.5 | OF | 1,100 | 9.6 | 12 | 5.72 | | |
| 26 | WELL | 16 | 5.5 | WELL | 650 | 5.66 | 19 | 0.01 | | |
| 36 | THINK | 12 | 4.2 | THINK | 518 | 4.5 | 21 | 0.08 | | |

Table 2: Top collocates for REALLY in SCO, MAC and BNC/C

**Rk.* – Rank of collocate in respective corpora. LL compares SCO : MAC based on total occurrences of *really*.

Looking first at the collocations of *really* in Table 2, there seem at first to be few differences – word co-occurrences with *really* are broadly similar in all three corpora. The log-likelihood test proves this point: only *it* (LL value 15.10), *to* (LL value 15.61) and *that* (LL value 24.99) stand out in appearing as *really* collocates with a significantly lower proportional frequencies. This appears to show that *really* collocates conform in SCO even more than *well* collocates seem to do[134].

### 8.2.2 Occurrence Differences found in the corpora

So far, we have no evidence to support the claim that there are systematic differences between the two sets of speakers (SCO and MAC). If we look at the *really* 2w clusters in Table 2 below, we can see that there is only a slight tendency to be significantly divergent in use for only

[134] Given that we have two relatively up-to-date corpora which are displaying a high degree of convergence, we are able to test the following claim (based on the LSWE Corpus) made by Biber *et al.:*

Both British and American English conversation commonly uses *really* to modify adjectives, especially … *good, nice, bad* and *funny*. (Biber *et al.* 2002: 196)

Table 1 shows that *good* indeed is amongst the top 25 most con-current collocates of all three corpora, occurring between 3.47% of all uses of *really* in MAC and 7.3% of all uses in SCO. The BNC/C proportional frequency of occurrence is in between at 5.0%. Biber *et al.*'s claim is therefore confirmed as far as *really* with *good* is concerned. The collocations with *nice, bad* etc. are, however, far less common in MAC or SCO than with *good*.

Three general facts seem apparent where both SCO and MAC corpora show similar patterns of use:
*Really* with "I" is the most common collocation – it accounts roughly for 1/3 of all such clusters with *really*.
*Really* with "I" **usually** expresses something with a negative connotation in both corpora.
When looking at possible combinations with which *really* with "I" can be found, the clusters that most often come up use a form of *do* or *can* + *not* .

Though it is not relevant to the main focus of this study, I therefore conclude that in Spoken English *really* with *"I"* expresses something negative in the majority of cases.

two clusters: *really good* (significantly different above the 99.9% level) and *it was really* (significantly different above the 99.9% level). Many other 2w *really* clusters, however, can be found to occur with absolutely minimal difference where SCO is compared with MAC. *Really really, not really, I really, don't really* etc. are examples of this.

| Cluster | SCO Freq. | MAC Freq. | LL |
|---|---|---|---|
| **Really really** | 22 | 812 | 0.11 |
| **Really good** | 17 | 285 | 9.34 |
| **Not really** | 17 | 680 | 0.00 |
| **It's really** | 13 | 788 | 2.64 |
| **Was really** | 12 | 407 | 0.27 |
| **It was really** | 6 | 37 | 11.56 |
| **I really** | 11 | 455 | 0.02 |
| **It really** | 11 | 612 | 1.38 |
| **Don't really** | 9 | 364 | 0.00 |
| **I don't really** | 6 | 199 | 0.18 |
| **I really don't** | 0 | 69 | n/a |
| **Are really** | 7 | 103 | 4.90 |
| **Is really** | 5 | 297 | 0.91 |
| **Really is** | 5 | 252 | 0.30 |

Table 3: Most frequent *really* 2w / 3w clusters in SCO with MAC equivalents.

### *8.2.3 Most divergent* really *clusters*

Concentrating on those clusters where significant differences have been found, we see the following:

| Cluster | SCO Freq. | SCO % | MAC Freq. | MAC % | LL |
|---|---|---|---|---|---|
| **Really good** | 17 | 5.88 | 285 | 2.48 | 9.34 |
| **It was really** | 6 | 2.08 | 37 | 0.32 | 11.56 |

Table 4: Most divergent 2w/3w *really* clusters where SCO is compared to MAC (% as of total *really* occurrences)

Table 4 shows that significant differences can only be found in two clusters - both of which are clearly far more proportionally frequent in SCO than they are in MAC.

The first of which, really good, appears in the following nestings:

| ***really good* cluster** | SCO tot. / % | MAC tot. / % |
|---|---|---|
| A REALLY GOOD +N | 3 /1.0 | 34 / 0.30 |
| IT'S REALLY GOOD | 4 / 1.38 | 37 / 0.32 |
| IT WAS REALLY GOOD | 2 /0.69 | 21 / 0.18 |
| REALLY REALLY GOOD | 0 / 0 | 30 / 0.26 |

Table 5: Most frequent occurrence patterns of *really good* in 3w clusters

Table 5 looks at the most frequent clusters that incorporate *really good.* Though the numbers in SCO are too low to give statistically reliable data, we can still see that the most frequent clusters in SCO are very rare in MAC, while *really really good*, making use of repetition, occurs in MAC but not in SCO. There is also a clue why the cluster *it was really* is disproportionally higher in its occurrence in SCO compared to MAC: 2 out of 6 times, it incorporates the two most prominent *really* clusters in SCO to form *it was really good.* All other forms, in SCO are *it was really* + a variety of adjectives.

### 8.2.4 I REALLY CAN'T

Amongst the Spoken Liverpool English clusters of *"I"* with *really* and with *negative,* there is one in particular that is in use in Liverpool but seems to have no significance in the comparators. It is the second most frequent three-word cluster for *really* with *"I"* in SCO (3 out 67 = 4.5%)[135]. No statistically reliable comparison can be made based on such low numbers. However, as the differences found are fairly strong, they would be seen as significant if projected onto corpora of a larger size.

The phrase *I really can't* is the chosen negative form in the majority of cases in SCO: It appears 3 times, (and is used by 2 speakers) out of a total of 33 three-word clusters of *really* with *"I"* in SCO[136].

This stands in marked contrast to MAC where this cluster is barely used. When we look at all the recorded clusters of *really* with "I" in MAC, we find 16 occurrences out of a total of 3165 (0.51%) and 4 occurrences out of 2851 in BNC/C (= 0.14%).

To express this within the wider picture of *really* usage: *I really can't* appears in 1.4% of all clusters containing *really* in the SCO corpus (289 occurrences), but accounts for just 0.14% of all uses of *really* in MAC (11,475 occurrences of the word) and for 0.044% of all uses of *really* in BNC/C (9.128 occurrences). That means it appears over ten times more

[135] See Appendix VII.1 for a list of frequent *really* with *I* clusters in SCO.

[136] Some further, anecdotal, evidence: While I have never heard *I really can't* uttered unprompted by a non-Merseyside speaker, I have heard it (unprompted) as part of conversations by Merseyside speakers.

often among Liverpool speakers than among  speakers in MAC and 30 times more often than among speakers in BNC/C.

Looking at the wider context where this phrase is used, Liverpool speakers say *I really can't* in two out of three cases when they refer back to a statement made earlier – in which they have used *I can't*. See the following example:

| | |
|---|---|
| 413. Ja | can't be arsed now |
| 414. Mi | no - come on you gotta do it |
| 415. Ja | no - i really can't |
| 416. Mi | loud |
| 417. Ja | no can't be asked |

I infer that *I really can't* is a single, freestanding phrase that refers back to something already expressed earlier. It is, therefore, context-bound. By contrast, in MAC, the cluster *I can't really* is usually followed by a verb (*remember* (twice), *get, doubt*), appearing to have no cohesive functions, and is not context-bound.

Even though the total occurrence numbers in SCO are low and this can be seen an obstacle to evaluating the *I really can't* occurrence pattern, its existence in SCO still implies that there should be a far higher rate of occurrence in MAC and BNC/C. Given that *I really can't* and *I can't really* occur the same amount of times in SCO, something similar could have been expected to be recorded in the comparators. The very fact that this phrase is extremely marginal in MAC and BNC/C highlights its importance for characterising SCO.

Table 6: REALLY *with* "I" 3word cluster comparison SCO; MAC and BNC/C ranked by frequencies in the respective corpus. Normalised to 10.000 occurrences.

Table 7(a): REALLY *with* "I" 3word cluster comparison SCO; MAC and BNC/C ranked by frequencies in the respective corpus. Normalised to 10.000 occurrences.

| SCO 3w cluster | Total | Per 10k | MAC 3w cluster | Total | Per 10k | BNC/C 3w cluster | Total | Per 10k |
|---|---|---|---|---|---|---|---|---|
| I DON'T REALLY | 6 | 207 | I DON'T REALLY | 214 | 187 | I DON'T REALLY | 164 | 180 |
| I REALLY CAN'T | 3 | 104 | I REALLY DON'T | 59 | 51 | I REALLY DON'T | 62 | 68 |
| | | | I'M NOT REALLY | 50 | 44 | I'M NOT REALLY | 36 | 39 |
| I CAN'T REALLY | 3 | 104 | I CAN'T REALLY | 46 | 40 | I CAN'T REALLY | 35 | 37 |

Table 6: REALLY *with* "I" 3word cluster comparison SCO; MAC and BNC/C ranked by frequencies in the respective corpus. Normalised to 10.000 occurrences.

| SCO 4w cluster | Total | Per 10k | MAC 4w cluster | Total | Per 10k | BNC/C 4w cluster | Total | Per 10k |
|---|---|---|---|---|---|---|---|---|
| I DON'T REALLY KNOW | 3 | 104 | I DON'T REALLY KNOW | 66 | 58 | I DON'T REALLY KNOW | 52 | 57 |
| WE REALLY DON'T HAVE | 2 | 69 | I'M REALLY REALLY REALLY | 29 | 25 | I DON'T REALLY WANT | 25 | 27 |
| SO I DON'T REALLY | 2 | 69 | I REALLY DON' T KNOW | 26 | 23 | I REALLY DON'T KNOW | 21 | 23 |
| | | | WELL I DON'T REALLY | 12 | 11 | I REALLY WANT TO | 8 | 8.8 |
| I DON'T KNOW REALLY | 2 | 69 | I DON'T KNOW REALLY | 14 | 12.3 | I DON'T KNOW REALLY | 4 | 4.3 |

Table 7(a): REALLY *with* "I" 3word cluster comparison SCO; MAC and BNC/C ranked by frequencies in the respective corpus. Normalised to 10.000 occurrences.

### *8.2.5.1 REALLY* with DON'T

The figures for *really* with *don't* are as follows: MAC has a total of 797 occurrences, equalling 6.95% of all occurrences of *really.* SCO has a total of 21 equalling 7.3% of all occurrences of *really.* So, usage is roughly equal. If we look a single 3w and 4w clusters in use, however, SCO figures are below 5. No statistically reliable comparison can be made based on such low numbers. However, as the differences found are fairly strong, they would be seen as significant if projected onto corpora of a larger size.

A clear difference in use between the corpora can be found when we look at the clusters of *don't* with *really* that exclude *know. So I don't really* and *I don't really have* are two 4-word clusters that account for nearly 10% of all the *don't* with *really* clusters in SCO. Neither *so I don't really* nor *I don't really have* are clusters that come up in MAC.[137] There, the only clusters of *really* with *don't* or with *don't* and *have* are *so you don't really* and *we really don't have*. These only account for 0.60% of all clusters with *don't* with *really.* Furthermore, the former appears in only 1.4% of all *really* with *don't* clusters in MAC and the latter is even less frequent. Admittedly, these are phrases with very low occurrence numbers in SCO, but the fact that certain clusters should appear in the far smaller Liverpool corpus but not in the far larger MAC hints that

[137] REALLY *with* SO *and* I seem to collocate in SCO: *So I really think* and *So I don't really think* are further examples of combinations making use of their collocation in SCO. None of the clusters with *SO* appear in MAC.

there is a specific use of *really* with *don't* which is worth further investigation. While we can see in Table 3 that *I don't really* occurs with a frequency not significantly different where SCO and MAC are compared, the nesting of this 3w cluster diverges clearly. *So I don't really* occurs 69 times per 10k words in SCO, but is very rare in MAC. *So I don't really* occurs 3 times in MAC – 2.5 times per 10k words. Instead, we find in MAC cluster *well I don't really*, though this too in MAC is extremely marginal - 11 times within 10k words of all clusters with *really* (0.3 % of *really with* "I" use in MAC*)*.

#### 8.2.5.2 I DON'T REALLY KNOW

In MAC, the preferred 4w cluster with *really* is *I don't really know*[138]. It accounts for 8.3% of all clusters of *really* with *don't*. In SCO, the dominant cluster is also *I don't really know* which accounts for 14.3 % of all clusters containing *really* with *don't*. This means, the cluster is proportionally occurring nearly twice as often among Liverpool speakers as among speakers across the UK. In MAC, even with variations of word-order (see Table 6), the total comes to 13.4% of all such clusters – the equivalent figure in SCO (*I don't really know* and *I don't know really*) would be 23.8%. This suggests differing colligational patterning. Taken together, variants of *I* with *don't* with *know* use are occurring

[138] I will return to the uses of this particular cluster in chapter 11.3.2

proportionally twice as often throughout for Scouse speakers when compared to speakers across the UK.

The major difference amongst the corpora lies in the choice of alternative word order for the phrase, *I + don't + really + know,* as shown in Table 3(b) below:

| Cluster | Occurrence in SCO per 10k | Occurrence in MAC per 10k | Occurrence in BNC/C per 10k |
|---|---|---|---|
| I DON'T REALLY KNOW | 104 | 58 | 57 |
| I REALLY DON'T KNOW | 0 | 23 | 23 |
| I DON'T KNOW REALLY | 69 | 12.3 | 4.3 |

Table 7(b) I *with* DON'T, KNOW *and* REALLY usage comparison (normalised to 10.000)[139].

Even with the highest used choice in all three corpora, *I don't really know,* SCO speakers are found to employ it nearly twice as often as MAC (or BNC/C) users. The second most frequent variation for MAC and BNC/C appears less than half as often again and is I *really don't know*. This variation is, however, not recorded in the Liverpool corpus at all. What is recorded instead is *I don't know really* which is only in marginal use (in the case of BNC/C: very marginal use) in the comparators yet appears in SCO two-thirds as often as the highest used choice, *I don't really know*. This also means *I don't know really* appears proportionally more often in SCO than *I don't really know* appears in either MAC or BNC/C. Taken together, this means that both the clusters *I don't really*

[139] if were to project these figures onto corpora 3 times the size, we would see the following:

| Cluster | *SCO total x3* | *MAC total x3* | Projected LL |
|---|---|---|---|
| I DON'T REALLY KNOW | 9 | 198 | 2.52 |
| I REALLY DON'T KNOW | 0 | 69 | n/a |
| I DON'T KNOW REALLY | 6 | 42 | 10.39 |

*know* and *I don't know really* are (each and taken together) proportionally far more frequently occurring in SCO than in the comparators.

### 8.2.6 Repetition of REALLY

Repetition of *really* is a noticeable (though not dominant) feature found in both SCO and MAC. In MAC, there are 576 occurrences of the bi-gram *really really* – 5% of all clusters with *really.*[140] (The amount of repetition does, of course, bend the statistics for sum totals). Single repetition of *really* amongst Liverpool English speakers stands at 2%. In total, there are only six occurrences of *really really* in SCO and *really really* is the only recorded form in SCO. The comparator MAC, however, records multiple repetitions of *really*.

While MAC records one speaker that uses 7 times *really* in consecutive order in an utterance and consequently brought about misleading statistics[141], the fact remains that MAC concordance lines show a number of occurrences where a speaker says *really really really* or even *really really really really.* As Concordance 1 below shows, this is a pattern also found within a typical excerpt of the occurrences of *really* single-word repetition in BNC/C:

[140] In the BNC/C, single repetition *really* clusters amount to 1461 entries – 6.8% of the total of all occurrences of *really.*

[141] Appendix VII has a draft of an earlier chapter before I discovered this anomaly.

| | | |
|---|---|---|
| Perhaps The thick ones really, | really | hurt me! Yeah but look t |
| Promise. Really, | really | promise? Really, really promi |
| Really, really promise? Really, | really | promise. Oh! Alright |
| . This paint I want it really really | really | well washed out of these brushes |
| well not over her but, but really, really | really | chummy yeah, and I thought and |
| Yeah. I mean Zed makes really really | really | really nice chilli. Mm. |
| catch a bus? really, really, | really, | really late |
| Oh dear I'm really really | really | really really looking forward to it. |
| Oh dear I'm really really really | really | really looking forward to it. Not th |
| Oh dear I'm really really really really | really | looking forward to it. Not that you |
| I'm really, really, | really | changed dramatically from not |
| n's boxing day thing was really really | really | shit? It was actually wasn |
| my house and she was really, really, | really | in a happy mood! And er you |
| bus? really, really, really, | really | late well we |
| Okay.I'm really really | really | gonna do this. Are these |
| Mmm really? Really | really | Yes provided yes but I would r |

Concordance 1: REALLY single-word multiple repetition as found in the BNC/C (excerpt)

This demonstrates that both MAC and BNC/C have a characteristic use of *really* that is not found in SCO, where all speakers restrict themselves to a single repetition of the term *really*. How this affects the usage of the term is shown by Table 5:

| REALLY | SCO total | per 10k | MAC total | per 10k | BNC/C total | per 10k |
|---|---|---|---|---|---|---|
| x2 | 6* | 207.6 | 406* | 354.1 | 401 | 439.3 |
| x3 | 0 | n/a | 15* | 17.3 | 54 | 59.2 |
| x4 | 0 | n/a | 8* | 9.2 | 13 | 14.2 |
| x5 | 0 | n/a | 6* | 6.9 | 0 | n/a |

Table 5: occurrence pattern of multiple single-word repetition of REALLY in SCO, BNC/C and MAC. Normalised by occurrence per 10.000 words out of the total of REALLY occurrence.
(* - based on count found in the concordance lines).

In Table 5 we can look at single-word repetition of *really* where SCO and MAC counts from the concordance lines are directly compared: *really*

*really* appears in both corpora[142]. It can be seen that the proportional occurrence of *really really* is lower in SCO than in MAC. Furthermore, Table 5 shows that SCO is the only corpus where the only repetition occurring with *really* is *really really*. Compared to either MAC or BNC/C, where multiple repetition of *really* can be found, even a corpus as small as SCO would be expected to record at least a small number of *really really really* use. Chunks like *really really really* appear to be rare in their use in Liverpool and this can be assumed to be the reason for their non-appearance in SCO.

### 8.2.7 REALLY Conclusions

Some claims made in the literature about the term *really* appear of little relevance here. *Really* in SCO does not seem to be expectation-led or indicating hyperbole. The fact that *really* is far more frequent in all three corpora than *actually* or *in fact* make in-depth corpus investigation of the item *really* feasible. Comparing the more recent corpora - MAC and SCO - with older data support the claim for a relatively recent preference for the use of *really* in spoken discourse.

Comparing *very and really* use in MAC and Scouse, Scouse speakers a present a "younger feel" as all ages appear to use the intensifier *really* more often than they use the intensifier *very* – a development in spoken English usually connected with younger speakers only. If Scousers use an

[142] And also in BNC/C though it must be noted that BNC/C figures are raw counts and repeated concordance lines are not eliminated.

intensifier, other options seem to be preferred, making the use of *very* marginal in SCO compared to MAC. While other research – notably by Bauer & Bauer (2001) and Ito & Tagliamonte (2003) – indicates that the use of *really* has changed over generations, my research appears to indicate that there is also a regional quality to its use.

Multiple single-word repetition of the intensifier *really* seemed at first a clear feature in MAC spoken English. However, I have shown that just a very small number with a high count of single one-word repetition can give a wrong impression about the real use. It is true, however, that MAC and BNC/C record instances of *really* multiple repetitions while SCO records only a single repetition of *really*.

On the whole, figures of *really* in SCO are too low to make many valid claims with regards to divergent occurrence patterns. Many findings have to be accepted as mere projections and, hopefully, access to a larger corpus will eventually validate these findings.

Under the given caveat, the cluster *I don't really know* occurs twice as frequently – proportionally – in SCO than in the comparators. SCO speakers also employ two further variants of this phrase (with different word order) which are rare in their use in both MAC and BNC/C. This hints at a collocational "accent", where a wider use in SCO is documented, that cannot be found in the general English corpora, yet given the low available figures in SCO, no reliable conclusions can be made.

This is also true for *I can't really* which occurs proportionally as often in SCO as in the comparators, the alternative variant – *I really can't* is a cluster that stands out in its comparatively high use amongst Liverpool speakers. The latter cluster is rare in both MAC and BNC/C. It is therefore possible that this is a Liverpool-specific phrase. Likewise *it was really (good)* is the only phrase where we have enough examples in SCO to make a statistically reliable claim that this cluster is significantly more frequently used in SCO than in MAC.

# Chapter 9 The uses of JUST and LIKE

While we have compared two *discourse particles* in the previous chapter that contrasted markedly in their patterns of use where SCO and MAC were compared, in this chapter I am looking at the items *just* and *like*. While these are found to be used by themselves as discourse markers in both corpora, the high level of co-occurrence is one reason to discuss these two terms in one chapter. Linguists like Tagliamonte (2005) have also highlighted further parallels between the words.

## 9.1 JUST – frequent with pronouns

*Just* is one of the most frequent words in both SCO and MAC and fits the description of *discourse particle.* However, as Tagliamonte points out:

> While an inordinate amount of media attention as well as academic research has been devoted to *like*, the use of *just* is barely mentioned. However, this form has also been increasing in recent years and has apparently garnered the same type of stigma as *like*. Indeed, when we examine our corpus, we find that *just* is one of the most frequent forms used among the young people. (Tagliamonte 2005: 1904)

This implies that *just* is associated with a certain youthfulness of language use when used frequently as a discourse particle.

In this section, comparison is made of *just* occurrence in SCO with that in MAC. For this, I first choose to look at the way just collocates. This is then followed by an examination of *just* and its most frequent clusters. Finally, I will look at those collocates and clusters of *just* where the difference in occurrence between SCO and MAC is the most marked.

Aijmer (2002) says that *just* has three main functions:

> …*Just* is used as a restrictive adverb paraphrasable as 'exactly' or 'only' (i.e. *just beyond Swindon)* (…) In addition, *just* has a temporal meaning *(just now*). (i.e. *I've only just discovered that* …) (…) The discourse particle *just* differs from the restrictive adverb because it signals involvement in the discourse event (i.e. *You've got a cold – No. Just a bit sniffy*) (Aijmer 2002: 155)

The difficulty here is that, while the temporal meaning of *just* is fairly straightforward to discover, both *just a bit* and *just beyond* could be seen as *just* in an adverbial function.

It is relevant, however, that Aijmer points out that the emphatic function accounts for 2/3 of the total occurrences of *just* – seven times more frequently than temporal uses of *just*. (cf. Aijmer 2002: 157)

As Table 1 shows, *just* occurs in the SCO corpus about twice as often as *really* and about half as often as *like*. There are 546 occurrences of *just* (0.46% of the corpus total). *Just* is proportionally significantly

used more in MAC: 30.739 occurrences mean it comprises just under 1.0% of all words. The BNC/C proportional frequency is closer to SCO – 0.49% of the total corpus.. For the purposes of the analysis that follows it is, however, immaterial how we categorise the word *just.*

*Just*, though it can be used for a range of meanings, only appears to occur in its function of discourse marker in SCO. This could be down to the relatively small size of the corpus. It may indicate a colligational choice, indicating a stronger bleaching of the meaning of *just* amongst SCO speakers. In what follows, the focus of the comparison is on the clusters that appear in SCO which also fulfil the same function in MAC.

### 9.1.1 Collocates of JUST in SCO and MAC

Unlike the other discourse markers discussed, the collocates of *just* in both SCO and MAC corpora do not differ to a great degree. The most prominent collocate (by a wide margin) of *just* in spoken English in both corpora is *I. I* collocates with *just* in nearly a third of all the occasions when *just* is uttered. Similarly, other high-frequency collocates of *just* differ very little when the corpora are compared. Divergences become apparent, however, in medium-high frequency collocates of *just.*

| Word | SCO Rank | % of total | Total | MAC Rank | % of total | Total | LL |
|---|---|---|---|---|---|---|---|
| *JUST* | | *0.46* | *546* | | *0.93* | *30,739* | *342.40* |
| I | 2 | 29.7 | 175 | 2 | 31.0 | 9,613 | 0.10 |
| IT | 8 | 15.1 | 89 | 3 | 27.3 | 8,405 | 28.09 |
| THE | 4 | 20.5 | 121 | 4 | 23.6 | 7,244 | 0.46 |
| YOU | 7 | 16.3 | 96 | 5 | 22.6 | 6,960 | 6.59 |
| IT'S | 12 | 9.3 | 55 | 11 | 9.3 | 3,397 | 0.48 |
| AND | 5 | 19.5 | 115 | 7 | 17.8 | 5,477 | 2.99 |
| A | 9 | 14.4 | 85 | 8 | 17.6 | 5,399 | 1.27 |
| TO | 10 | 12.7 | 75 | 9 | 15.6 | 4,800 | 1.27 |
| THAT | 16 | 7.8 | 46 | 10 | 15.3 | 4,706 | 19.93 |
| THAT'S | 58 | 2.5 | 15 | 33 | 4.1 | 1,346 | 3.77 |
| OF | 27 | 5.3 | 31 | 12 | 9.9 | 3,029 | 11.25 |
| THEY | 14 | 9.0 | 53 | 13 | 7.7 | 2,374 | 2.52 |
| IN | 19 | 6.9 | 41 | 15 | 7.7 | 2,352 | 0.01 |
| HE | 18 | 6.9 | 41 | 16 | 6.1 | 1,874 | 31.67 |
| WAS | 17 | 7.6 | 45 | 17 | 6.0 | 1,849 | 3.64 |
| EH | 45 | 2.9 | 17 | 18 | 5.9 | 1,820 | 9.16 |
| EHM | 23 | 5.6 | 33 | 33 | 4.0 | 1,325 | 0.01 |
| ON | 22 | 5.9 | 35 | 19 | 5.7 | 1,759 | 4.75 |
| IS | 20 | 6.8 | 40 | 23 | 4.9 | 1,504 | 30.33 |
| YEAH | 15 | 9.0 | 53 | 24 | 4.8 | 1,481 | 20.39 |
| CAN | 95 | 1.4 | 8 | 25 | 4.8 | 1,479 | 17.29 |
| SO | 25 | 5.6 | 33 | 26 | 4.7 | 1,467 | 1.67 |
| BUT | 34 | 3.4 | 20 | 27 | 4.7 | 1,438 | 1.28 |
| NO | 56 | 2.5 | 15 | 28 | 4.6 | 1,430 | 4.92 |
| THERE | 21 | 6.6 | 39 | 29 | 4.5 | 1,420 | 6.31 |
| LIKE | 11 | 11.7 | 69 | 30 | 4.5 | 1,390 | 51.83 |
| KNOW | 24 | 5.6 | 33 | 31 | 4.5 | 1,382 | 2.57 |
| WHAT | 28 | 5.1 | 30 | 32 | 4.4 | 1,338 | 1.48 |
| NOT | 39 | 3.1 | 19 | 34 | 4.2 | 1,301 | 0.77 |

Scouse < 5% less
Scouse <2% more
Scouse >2% more

Table 1: JUST most frequently occurring collocates in SCO and MAC. Percentages relative to the total number of *just*.

Table 1 shows that the highest occurring collocates of *just,* namely *I, the, you* and *and* show little difference in proportional frequency of occurrence, and *just* collocates *I, the, and, to* and *it's* show the least statistical significant differences.

We now turn to the cases where SCO and MAC differ with the highest degree of statistical significance (Table 2):

| Word / Collocate | Rank SCO | Freq. SCO | % | Rk. MAC | Freq. MAC | % | LL |
|---|---|---|---|---|---|---|---|
| ***JUST*** | | *546* | *0.46* | | *30,739* | *0.93* | *342.40* |
| **LIKE** | 11 | 69 | 11.7 | 30 | 1,390 | 4.5 | 51.83 |
| **HE** | 18 | 40 | 6.9 | 16 | 1,874 | 6.1 | 31.67 |
| **IS** | 20 | 41 | 6.8 | 23 | 1,504 | 4.9 | 30.33 |
| **IT** | 8 | 89 | 15.1 | 3 | 8,405 | 27.3 | 28.09 |
| **THAT** | 16 | 46 | 7.8 | 10 | 4,706 | 15.3 | 25.22 |
| **YEAH** | 15 | 53 | 9.0 | 24 | 1,481 | 4.8 | 20.39 |
| **CAN** | 95 | 8 | 1.4 | 25 | 1,479 | 4.8 | 17.29 |
| **OF** | 27 | 31 | 5.3 | 11 | 3,029 | 9.9 | 11.25 |

Table 2: Rank, frequency and prop. percentage of *just* collocates that are most divergent

Table 2 shows that amongst *just* collocates, *that* is proportionally significantly less used in SCO than in MAC[143] while *like* and *yeah* both collocate with *just* approximately twice as frequently in SCO as in MAC. The rankings indicate how many other collocates are more frequent than the word in question. An interesting find is that the words *it* and *that* (variously described to as pronouns or referrers - their function depends very much on the context) as well as the preposition *of* are found in SCO

[143] This is similar to what we have seen with previously discussed discourse markers.

with a proportional frequency that is about half the proportional frequency of these *just* collocates in MAC.

### *9.1.2 JUST* 2-word clusters

When we look at *just* 2w clusters, we find a high degree of convergence between SCO and MAC but also a larger number of short clusters that diverge along the same lines as we have found in 9.1.1, when we looked at collocates.

| *JUST 2w cluster* | SCO tot. | % | MAC tot. | % | log likelihood |
|---|---|---|---|---|---|
| I JUST | 93 | 17.0 | 2607 | 8.5 | 35.50 |
| JUST I | 6 | 0.25 | 501 | 1.6 | 1.05 |
| IT'S JUST | 34 | 6.2 | 1669 | 5.4 | 0.60 |
| JUST LIKE | 30 | 5.5 | 567 | 1.8 | 24.95 |
| YOU JUST | 26 | 4.8 | 1618 | 5.2 | 0.27 |
| WAS JUST | 22 | 4.0 | 1005 | 3.3 | 0.88 |
| IT JUST | 18 | 3.3 | 1011 | 3.3 | 0.00 |
| JUST A | 15 | 2.7 | 1595 | 5.2 | 7.48 |
| WE JUST | 15 | 2.7 | 528 | 1.7 | 2.79 |
| JUST GO | 15 | 2.7 | 406 | 1.3 | 6.25 |
| IS JUST | 15 | 2.7 | 463 | 1.4 | 4.38 |
| HE JUST | 15 | 2.7 | 560 | 1.8 | 2.17 |
| JUST THE | 12 | 2.7 | 87 | 0.28 | 27.09 |

Table 3: 13 most frequent SCO *just* 2w clusters and their MAC equivalents.

We can see, in Table 3, that not many 2w clusters diverge between the two corpora. Where they do, however, the 2w *just* clusters appear significantly more often, proportionally, in SC O than they do in MAC.[144]

[144] With the possible exception of *just a* which is significant just above the 99.0% level and proportionally appears twice as often in MAC than in SCO.

The most interesting 2w cluster is *just* with *I*. *I* as a collocate appears at the same relative level in the two corpora, and so does the 2w cluster *just I*. The significant divergence is found, however, when we look at the highest occurring *just* with *I* cluster, *I just*. This appears proportionally twice as often in SCO as it does in MAC.

Two further 2w *just* clusters diverge significantly: *just like* which appears proportionally three times as often in SCO than in MAC, and *just the*, appearing proportionally nearly ten times as often in SCO than in MAC. Below, we will see whether these differences can still be found when (and if) these 2w clusters are found as constituent part of 3w clusters.

### 9.1.3 JUST 3w clusters with A

Looking at the totals for collocates (cf. Table 1) *just* with *a* occurs slightly more often in MAC than in SCO.

| SCO | tot. | % | MAC | tot. | % | BNC/C | tot. | % |
|---|---|---|---|---|---|---|---|---|
| JUST LIKE A* | 5 | 0.91 | JUST LIKE A | 71 | >0.3 | JUST LIKE A | 50 | >0.3 |
| JUST FOR A | 4 | 0.74 | JUST FOR A | 58 | >0.3 | JUST FOR A | 42 | >0.3 |
| IS JUST A | 3 | 0.55 | IS JUST A | 60 | >0.3 | IS JUST A | 27 | >0.3 |
| JUST HAD A | 3 | 0.55 | JUST HAD A | 89 | 0.3 | JUST HAD A | 76 | 0.39 |
| IT'S JUST A | 0 | n/a | IT'S JUST A* | 312 | 1.0 | IT'S JUST A* | 166 | 0.84 |
| (*Highest occ. such cluster in respective corpora) | | | JUST A LITTLE | 160 | 0.5 | JUST A LITTLE | 89 | 0.45 |
| | | | JUST HAVE A | 126 | 0.4 | JUST HAVE A | 94 | 0.48 |
| | | | JUST A BIT | 84 | 0.3 | JUST A BIT | 69 | 0.35 |

Table 4: Divergent proportional usage of *just* with *a* cluster in SCO, MAC and BNC/C.

The most frequent *just* with *a* 3w clusters shown in Table 4 demonstrate a clear divergence of use for the target *just* with *a* between Liverpool spoken English and UK spoken English as represented in the respective corpora, though SCO, MAC and BNC/C have in common that their respective highest occurring 3w cluster appears in around 1% of all uses of *just* and that all the clusters with *a* recorded for SCO can also be found in the comparators. Yet these common aspects are minor compared with the divergences in the proportional frequencies found between the most frequent SCO 3w *just* with *a* clusters and their MAC (and BNC/C) equivalents.

A difference can be found, when the three highest occurring clusters of *just* with *a* are compared. Table 4 shows that *it's just a*, the most frequently occurring of such clusters in MAC (312 occ. / 1.0% of all uses of *just)* and BNC/C (166 occ. / 0.84%) does not occur at all in SCO. Conversely, *just like a*, the highest occurring cluster in SCO, is marginal in the comparators. It appears more frequent in SCO than in MAC with above 95% statistical significance (LL value of 6.11) .The same is true for *just for a* and *is just a* which appear 740 and 549 times in every 100.000 (100k) uses of *just* respectively in SCO but appear less than 300 times in 100k in MAC and BNC/C. Once more, *just* occurrence pattern in SCO shows uses that are only in the margins in the comparators.

### 9.1.4 JUST 3w clusters with LIKE

Throughout the earlier part of the discussion it has become obvious that *just* with *like* form clusters in a number of variations amongst Scouse speakers, but that this is not the case in MAC. In this section, I look at the clusters that are in both corpora. *Just* with *like* is a very infrequent combination in any of the three corpora, making it not possible to draw firm and final conclusions:

| CLUSTER | Freq. SCO | % | Freq. MAC | % | LL | Freq. BNC/C | % |
|---|---|---|---|---|---|---|---|
| JUST LIKE A | 5 | 0.91 | 71 | 0.24 | 6.11 | 50 | 0.25 |
| WAS JUST LIKE | 5 | 0.91 | 44 | >0.2 | 9.74 | 43 | 0.22 |
| IT'S JUST LIKE | 3 | 0.54 | 63 | >0.2 | n/a | 60 | 0.30 |
| JUST LIKE JUST | 2 | 0.32 | n/a | n/a | n/a | n/a | n/a |

Table 5: JUST *with* LIKE 3w cluster comparison in SCO, MAC & BNC/C.

Table 5 shows a difference for every single *just* with *like* cluster. The most frequently used 3w clusters in SCO, are proportionally much less used in both MAC and BNC/C – namely *just like a*, *was just like* and *it's just like*. Where SCO records more than 5 instances of use, the divergences are statistically significant above the 99.0% level only.

### 9.1.5 JUST 3w-clusters with "I"

We saw, in the discussion of Table 2 above, that *I* is the most frequent collocate in both corpora. Here, we will look at the use of *just* with *I* in detail to see whether the items co-occur in ways similar or divergent between SCO, MAC and BNC/C.

As Tables 6 and 7 show, *I was just* is one of most frequent *I* with *just* clusters in all three corpora. It accounts for 0.85% of all uses of *just* in SCO, for 1.1% of all uses of *just* in MAC (and for 1.45% in BNC/C). At the same time, it is also the *just* cluster with the least difference of use.

| Common cluster | SCO Total / % | MAC Total / % | LL |
|---|---|---|---|
| KNOW I JUST | 6 / 1% | 52 />0.25 | 11.83 |
| I JUST THOUGHT | 6 / 1% | 80 / 0.25% | 7.87 |
| I WAS JUST | 5 / 0.85% | 325 / 1.1% | 0.11 |
| I JUST HAD | 5 / 0.85% | 32/>0.25% | 12.30 |
| JUST – I - JUST - | 4 / 0.7% | 70/ >0.25% | |
| I JUST COULDN'T | 4 / 0.7% | 23/ >0.25% | |
| AND I JUST | 4 / 0.7% | 152 / 0.5% | |
| I JUST I | 3 / 0.5% | 186 / 0.59% | |

Table 6: JUST *with* "I" clusters in SCO and MAC. Percentages as of total JUST occurrences.

| JUST cluster in the BNC/C | 19,695 items JUST total | % | LL SCO:BNC/C |
|---|---|---|---|
| KNOW I JUST | 56 | 0.26 | 6.99 |
| I JUST THOUGHT | 95 | 0.48 | 3.03 |
| I JUST HAD | 35 | 0.17 | 7.90 |
| | | | frq. dist. |
| I WAS JUST | 286 | 1.45 | 100% |
| IT WAS JUST | 228 | | |
| JUST SORT OF | 167 | | |
| IT'S JUST A | 166 | | |
| JUST HAVE TO | 156 | | |
| AND I JUST | 151 | 0.77 | ~50% |
| IT'S JUST THAT | 130 | | |
| I JUST DON'T | 119 | 0.60 | |
| IT'S JUST THE | 119 | | |
| YOU KNOW JUST | 117 | | |

Table 7: JUST *with* "I" most frequent clusters in BNC/C, indicating JUST *with* "I" relative usage in BNC/C

There are, however, clusters that are markedly different in its pattern of use: First of all, there is *I just thought*. This cluster is amongst the most frequent clusters of *just* with *I* in SCO. However, whereas it accounts for 1.0% of 3-word *just* clusters in SCO, it only accounts for 0.25% of such clusters in MAC. In the BNC/C, the relative distribution of clusters here is in line with MAC (see Table 6). What Aijmer (2002) says about *just* and *indirectness* suggests the potential relevance of this:

> … *Just* collocates with 'I think' and it modifies the assertion (…) Assertions, questions, suggestions, criticism or requests are face-threatening acts whose effects may not be welcome to the hearer. It may involve the risk for the speaker to ask a too direct question, make a request abruptly, assert something without simultaneously using 'redressive action' (…) *Just* is a way of avoiding or softening the … face-threatening act of requesting by conveying that there is only one thing the speaker is wondering about. *Just* is associated with something small and unimportant. (Aijmer 2002: 169)

A higher use of the *just* could therefore be interpreted as evidence of a stronger effort undertaken by SCO speakers to avoid face-threatening utterances. I will come back to this pattern of *strategic hedging* later.

*I just thought* is a cluster that appears with 99.0% significance more often in SCO than in MAC. Compared to BNC/C, however, the difference is barely significant.

There are, however, two more clusters where there is a reliably significant proportionally higher use of *just* with *I* in SCO: *Know I just* and *I just had.*

| Cluster | SCO Frq | % | MAC Frq | % | LL | BNC/C Frq | % |
|---|---|---|---|---|---|---|---|
| KNOW I JUST | 6 | 1.1 | 52 | 0.19 | 11.83 | 56 | 0.28 |
| YOU KNOW I JUST | 4 | 0.7 | 24 | 0.09 | n/a | 31 | 0.15 |
| I JUST HAD | 5 | 0.85 | 32 | 0.12 | 12.30 | 35 | 0.18 |
| I JUST COULDN'T | 4 | 0.7 | 25 | 0.096 | n/a | 3 | 0.015 |

Table 8: Divergent proportional usage of JUST *with* I cluster in SCO and MAC. (See also Table 4)

Looking more closely at the first of the *just* with *"I"* clusters in Table 8, we find it appears two-thirds of the time as part of the four-word cluster *you know I just* (4 occ. / 3 speakers = 0.7% of all clusters with *just*). In MAC, this cluster is very rare. More broadly speaking, the 3w and 4w clusters of *just* with *I* in Table 8 are proportionally five to ten times less frequent in MAC. As Table 6 shows, BNC/C proportional figures are close to MAC and therefore are similarly divergent from SCO. A phrase like *you know I just* again points to SCO speakers employing greater face-saving terminology. *I just had* and *I just couldn't* are two 3w *just* clusters where the higher proportional use in SCO must be noted but total figures are to low for these clusters to appear in longer clusters.

While the majority of *just* with *I* clusters used by speakers occur with similar frequencies, the relatively low-frequency *just* clusters with *"I"* in SCO appear to point toward a slightly different use of the term. *Just* with *know* and *I* appears to be frequent amongst Scouse speakers, infrequent amongst other UK English speakers.

### 9.1.6 JUST Conclusions

*Just* is an interesting item to study when differences of lexical use are the focus. The main collocates of *just* – *the, to* and *it* – show a broadly similar use and the same clusters appear in both corpora with similar rates of usage where *just* is employed as a downtoner. However, there appears to be a significant number of clusters with personal pronouns – *just* with *he, it* and *that* – where there is a marked difference between the use by Liverpool speakers and speakers from across the UK.

The differences in use come into a still stronger focus when we look at *just* with *a* or *of* (lower percentage of collocates in SCO) as well as *just* with *yeah* or *like* (higher percentage of collocates in SCO).

All the four collocates mentioned have in common that they form clusters that are a preferred choice for SCO speakers but not for MAC (and BNC/C) speakers. When the proportional level of occurrence is compared to MAC and BNC/C, it shows that fairly commonly used 3w clusters in SCO are very marginal in both MAC and BNC/C.[145]

We have shown that *I just* and the longer clusters incorporating it, *know I just* and *I just had* all appearing with significantly higher proportional frequencies in SCO compared to MAC.

[145] The opposite is also the case: fairly common 3w JUST clusters in the comparators are sometimes not be found in SCO.

Likewise, *just like* and the longer clusters incorporating it, *just like a* and *was just like* are significantly more prominent in their use in SCO than they are in MAC.

## 9.2 A view on the many uses of LIKE

*Like* is a lexical item that has been, and still is, increasingly *delexicalised* or *bleached*. Its origin lies in the Germanic *lik* – meaning body. In modern German it is still used in that sense – *Leiche* [laixΘ] – corpse or *Laich* [laix] – spawn. However, there is also one use where the German word has a direct English equivalent – one of the uses of *like* and *gleich* [glaix], meaning 'equal', 'the same'.[146]

Today, *like* in English is also used as an intensifier. This puts it in line with other intensifiers like *really* (coming from *real*) and *very* from *vrai* – true. A similar bleaching-process can be observed in other languages, too. For example, in Spanish, with *-le*, now used as a suffix intensifier, originally coming from *leismo,* the process of change appears very similar:

> The development of *le* into a verbal intensifier can be understood in the framework of a diachronic process of semantic bleaching. Bleaching or semantic reduction is the loss of features of meaning associated with a form (Bybee et al. 1994: 19). Bleaching of *le* culminates in loss of its argument and pronominal status. As *le*'s argument and pronominal status is eroded, it functions less as an active participant and more as the

[146] It is interesting to note how similar the German and Scouse pronunciation of this word is -[laix] in German; [lai$^{k}$xh] in Scouse.

> location in which the event occurs. In intensifier usage, *le* no longer refers to a participant in the event. Instead, it is a verbal affix, somewhere between derivation and inflexion. (Cacoullos 2002: 286)

According to Streeck (2002):

> *Like* has taken on, first, a role as "discourse marker", specifically as "focus marker", a type of unit that marks subsequent talk as salient. (...) *like* is a prime example of a linguistic unit that, because it has undergone multiple stages of grammaticalisation, relexicalisation, and expansion of use, affords members of the speech community a wide range of things to do with it. (Streeck 2002: 583)

There is very good evidence that the use of *like* found in casual Liverpool English speech qualifies as what the OED terms as *colloq.*, a "meaningless interjection" – where the OED quotes magazine articles to exemplify this "colloquial speech":

> **1961** *New Statesman* 22 Sept. 382/2 'You're a chauvinist,' Danny said. 'Oh, yeah. *Is that bad like?'* **1966** *Lancet* 17 Sept. 635/2 As we say pragmatically in Huddersfield, *'C'est la vie, like!'*

There are three observations that can be made about these quotes.

Number 1 – these are both quotes from the 1960s, the height of Liverpool's fame as a hub of popular culture. In Pace-Sigge (2002) I showed that it was around this time that key pronunciations in Liverpool English speech became more fixed and were already clearly identifiable as *Scouse*. It could be said that certain lexical uses of that time also

became very popular and, being connected by Liverpudlians with "good times", their heyday, stayed in the local idiom.

Hoey (2005) speaks of how primings are created:

> Primings can be receptive as well as productive. **Productive primings** occur when a word or word sequence is repeatedly encountered in discourses (…) in which we are ourselves expected (…) to participate and when speakers (…) are those we like or wish to emulate. **Receptive primings** occur when a word or word sequence is encountered in contexts in which there is no probability (…) of us ever being active participant – party political broadcasts [etc] or where the speaker or writer is someone we dislike or have no empathy with … (Hoey 2005: 11f.) (bold in the original)

Hoey's claims could be reasonably broadened, so that a more positive context in which a word or word sequence is encountered will increase the likelihood of productive priming[147]. This seems to be the case here. This theory would support the idea of *priming* in speech – where speakers, subconsciously, pick up a certain usage because of its positive connotations.

Number 2 – the connection with Huddersfield. Earlier, I already pointed to the possibility that Scouse, rather than being an autonomous dialect, could also be a constituent part of a North-West of England group of dialects.

Number 3 – the two examples quoted above have *like* postpositioned. This may be of relevance when looking at the use of *like* as intensifier. As

[147] This is equally true for receptive priming. This, for example, can be used to explain that while news report on "nationalist movements", few people in modern-day Britain would call themselves a *nationalist*.

an end-particle, *like* appears to be an evaluative marker. Anecdotal evidence says that Liverpool speakers tend to end a clause in casual speech with *like* (*it was a boss match like)* – this could be seen as specific use by Liverpool speakers; it is adding stress to the preceding clause and indicating familiarity with the listener and the subject. (An substitute for this use of *like* would be *you know* which is also found post-positioned).

There are further uses of the word *like*. One is as a filler (*It was like – oh, I don't know)*. Schiffrin (1987) makes no mention of *like* as discourse marker. Schourup (1985) on the other hand points out that

> One frequent use of the form was preceding numerical expressions (i.e. *like one more week)* (…) but in other cases *like* precedes non-numerical expressions (i.e. *like every other night*). (Schourup 1985: 38)

Schourup also says that *like* appears to "introduce direct discourse" and serves as *interjection*. He concludes:

> (…) *like* in conversation, at least among younger speakers,(…) [demonstrates] the spread from its originally quite restricted range of occurrence to an item which in general indicates a possible loose fit between overt expression and intended meaning. With this, *like* is particularly suited to conversation where speakers frequently find themselves in the position of having to formulate what they say without time for the considered eloquence possible [when writing] (Schourup 1985: 61)

This highlights again the *bleaching* of the term while positioning it firmly as a feature of spoken (rather than written) English. Schourup, it has to be pointed out, based his work on US American data. Miller & Weinert

make strong reference to Schourup when they use data from 8 – 16 year old Scottish young people and come to the conclusion that -

> LIKE constructions - clause-initial and clause-final LIKE - have different discourse roles. In general LIKE is a non-introducing, non-contrastive focuser which may focus on new or given information. In addition, clause-initial LIKE is concerned with the elucidation of previous comments, whereas clause-final LIKE is concerned with countering objections and assumptions.
>
> (Capitals in original), (Miller & Weinert 1995: 392)

Both Schourup and Miller & Weinert underline how much *like* appears to be in use in younger people. This is an issue taken up by Tagliamonte in her Canadian study:

> The consistent, highly frequent result for pre-noun phrase position, (sic) in particular suggests that it may be developing some kind of function in the grammar. As mentioned earlier, *like* is typically associated with young people. (…) The 15- to 16-year olds are using more tokens of *like* than any other age group. At the two ends of the scale—the 10- to 12-year olds and the 17- to 19-year olds, however, the use of *like* is much lower.
>
> (Tagliamonte 2005: 1904)

These findings are similar to Ito & Tagliamonte's work conducted in York (2003). Tagliamonte seems to be in agreement with Schourup about the grammatical function that *like* seems to develop. She also indicates that the use of *like* is a feature of a particular, young, age group, something that is not common amongst young children and also less used amongst young adults. It is in fact specifically a teenage language indicator. In this chapter, I do not, however, look at *like* use by age group, though, as data

are available to look at use by age-distribution, this line of enquiry would have been available[148].

In all other respects, in the investigation that follows, I examine in how far my corpora support the research findings described above.

### 9.2.1 Comparison of the top collocates of LIKE

*Like* is an item with multiple functions of meaning in spoken English. Therefore we find *like* in the role of a comparator: *and stuff like that* or *a bit like*. It also is being used to express preference*: I like it* or *you don't like*. *Like* can also be used as a discourse particle*: I mean like*. To decide whether the target word is employed as a discourse particle rather than a preference marker, *like* has to be looked at in the wider context. In *It was* ***you know like*** *in the middle ages* puts *like* in the role of the comparator while in *I was acting it out,* ***you know like*** *on the floor,* the item *like* is a discourse particle, probably employed as a downtoner.

[148] One conclusion that could be drawn from Schourup, Miller & Weinert and Tagliamonte is that strong use of *like* gives spoken utterances a "young" feel. I will investigate this in more detail in the *very* subsection, chapter 8.6.

Table 1: Top collocates of *like* in SCO, MAC and BNC/C. Percentages (apart from top-line) are in relation to the total occ. of the core term *like.*

| Word | Total: SCO | % | Word | Total: MAC | % | Word | Total: BNC/C | % |
|---|---|---|---|---|---|---|---|---|
| ***LIKE*** | ***970*** | ***0.81*** | ***LIKE*** | ***22,858*** | ***0.69*** | ***LIKE*** | ***21,920*** | ***0.54*** |
| I | 254 | 26.2 | I | 8,338 | 36.5 | I | 6769 | 30.1 |
| THAT | 206 | 21.2 | THAT | 7,016 | 30.2 | YOU | 5731 | 26.1 |
| THE | 192 | 19.8 | IT | 6,591 | 28.9 | THAT | 5579 | 25.5 |
| AND | 178 | 18.4 | YOU | 6,431 | 28.1 | IT | 4266 | 19.5 |
| YOU | 176 | 18.3 | THE | 5,091 | 22.3 | AND | 4214 | 19.2 |
| IT | 157 | 16.2 | TO | 4,891 | 21.7 | A | 4018 | 18.3 |
| A | 132 | 13.6 | A | 4,555 | 20.2 | THE | 3961 | 18.1 |
| IT'S | 122 | 12.6 | AND | 4,548 | 20.1 | TO | 3072 | 14.0 |
| TO | 102 | 10.5 | OF | 2,228 | 9.9 | IT'S | 2283 | 10.4 |
| WAS | 98 | 10.0 | THEY | 2,013 | 8.8 | DON'T | 2082 | 9.5 |
| KNOW | 84 | 8.7 | DON'T | 1,896 | 8.3 | OF | 1800 | 8.2 |
| THEY | 78 | 8.0 | HE | 1,869 | 8.2 | KNOW | 1758 | 8.0 |
| YEAH | 74 | 7.6 | WHAT | 1,739 | 7.6 | WAS | 1520 | 6.9 |
| JUST | 70 | 7.2 | WOULD | 1,736 | 7.6 | HE | 1471 | 6.6 |
| OF | 66 | 6.8 | THIS | 1,722 | 7.5 | BUT | 1429 | 6.5 |
| IN | 55 | 5.7 | IN | 1,673 | 7.3 | IN | 1419 | 6.5 |
| DON'T | 53 | 5.5 | KNOW | 1,592 | 7.0 | THIS | 1355 | 6.3 |
| EH | 52 | 5.3 | WE | 1,575 | 6.9 | THEY | 1309 | 6.1 |
| IS | 50 | 5.1 | YEAH | 1,426 | 6.2 | WHAT | 1246 | 5.8 |

Table 1: Top collocates of LIKE in SCO, MAC and BNC/C. Percentages (apart from top-line) are in relation to the total occ. of the core term LIKE.

*Like* is a fairly frequent item in all three corpora. There are 970 occurrences in SCO (0.81%) and 22,858 occ. in MAC (0.69%) while the BNC/C reports 21,920 uses of *like* (0.54%). This means that *like* is significantly (see Table 2 below) *more* frequent, proportionally, in SCO than it is in MAC. Table 1 compares the most common collocates of *like* and also a number of words that appear to be frequently used with this word – notably, indications of pause (*ehm, erm,* etc.). While there are certain collocates (like *they* or *and*), where the proportion of occurrence is about the same, there are also some striking differences to be seen.

The BNC/C collocates on the whole are more closely aligned to MAC than they are to SCO.

| Word / Collocate | Freq. SCO | % | Freq. MAC | % | LL |
|---|---|---|---|---|---|
| *LIKE* | *970* | *0.81* | *22,858* | *0.69* | 23.29 |
| THAT | 206 | 21.2 | 7,016 | 30.2 | 70.26 |
| TO | 102 | 10.5 | 4,891 | 21.7 | 63.98 |
| IT | 157 | 16.2 | 6,591 | 28.9 | 61.87 |
| YOU | 176 | 18.3 | 6,431 | 28.1 | 38.00 |
| A | 132 | 13.6 | 4,555 | 20.2 | 21.10 |
| WAS | 98 | 10.0 | 1,419 | 6.1 | 18.93 |
| DON'T | 53 | 5.5 | 1,896 | 8.3 | 10.20 |

Table 2: *Like* collocates most divergent

Though *like* occurrence in the respective corpora has the lowest figure of all the LL tests undertaken for *discourse particles* in this thesis, the value is still above the level that indicates 99.99% significance. *Like*

collocates[149], show, however, a higher degree of divergence as there are more collocates that are significantly lower in their proportional frequency of use in SCO compared to MAC. The one exception is the *like* collocate *was* which appears 1.6times more frequent in SCO than it does in MAC.

### 9.2.2 LIKE usage: divergence in 2-4w clusters

When we look at the clusters of *like*, we find two things: in spoken usage, a large number of 2w clusters recorded appears to be often an integral part of 3-4w clusters and, furthermore, *like* is often employed with *vagueness markers*. This becomes very clear when the highest occurring 3-4w clusters in SCO and MAC are compared. The most obvious differences can be found in the top of the Table 3. This table shows the clusters ordered by frequency of occurrence in the two corpora and, where the same cluster appears in both corpora, the same background colour is used. In order to highlight which of these clusters differ significantly on statistical terms, 3-4w *like* clusters are being compared pairwise in Table 4(a) and those which diverge most significantly are highlighted in bold type.

[149] Where proportional frequency and statistical testing is based on the total number of *like* in each of the corpora.

| SCO top LIKE cluster | occ. | (%) | MAC top LIKE cluster | occ. | (%) |
|---|---|---|---|---|---|
| LIKE YOU KNOW | 16 | 1.7 | WOULD LIKE TO | 578 | 2.5 |
| STUFF LIKE THAT | 16 | 1.7 | I DON'T LIKE | 564 | 2.4 |
| I WAS LIKE | 15 | 1.6 | SOMETHING LIKE THAT | 555 | 2.3 |
| IT WAS LIKE | 14 | 1.5 | I WOULD LIKE | 446 | 2.0 |
| THINGS LIKE THAT | 11 | 1.2 | THINGS LIKE THAT | 365 | 1.6 |
| SOMETHING LIKE THAT | 11 | 1.2 | WOULD YOU LIKE | 327 | 1.4 |
| I LIKE THE | 11 | 1.2 | LIKE THAT AND | 322 | 1.4 |
| I LIKE THAT | 10 | 1.1 | LIKE THAT I | 280 | 1.2 |
| I DON'T LIKE | 9 | 0.9 | I LIKE THE | 269 | 1.1 |
| AND STUFF LIKE | 9 | 0.9 | LIKE YOU KNOW | 264 | 1.1 |
| ANYTHING LIKE THAT | 9 | 0.9 | YOU KNOW LIKE | 264 | 1.1 |
| YOU KNOW LIKE | 9 | 0.9 | OR SOMETHING LIKE | 254 | 1.1 |
| A BIT LIKE | 9 | 0.9 | IF YOU LIKE | 245 | 1.1 |
| AND STUFF LIKE THAT | 8 | 0.8 | LIKE TO SEE | 234 | 1.0 |
| I MEAN LIKE | 7 | 0.7 | I LIKE THAT | 218 | 1.0 |
| LIKE THAT AND | 6 | 0.6 | AND THINGS LIKE | 210 | 0.9 |
| OR SOMETHING LIKE | 6 | 0.6 | YOU LIKE TO | 194 | 0.8 |

Table 3: Highest occ. clusters with LIKE in SCO and MAC. Percentages relative to occ. of LIKE in respective corpus

| LIKE *3w-4w* cluster | SCO occ. | SCO % | MAC Occ. | MAC (%) | LL |
|---|---|---|---|---|---|
| LIKE YOU KNOW | 16 | 1.7 | 264 | 1.1 | 1.73 |
| **STUFF LIKE THAT** | 16 | 1.7 | 81 | 0.4 | 22.30 |
| **I WAS LIKE** | 15 | 1.6 | 30 | 0.2 | 41.25 |
| IT WAS LIKE | 14 | 1.5 | 148 | 0.7 | 6.63 |
| THINGS LIKE THAT | 12 | 1.3 | 365 | 1.6 | 0.82 |
| SOMETHING LIKE THAT | 11 | 1.2 | 555 | 2.3 | 8.08 |
| I LIKE THE | 11 | 1.2 | 269 | 1.1 | 0.01 |
| I LIKE THAT | 10 | 1.1 | 218 | 1.0 | 0.06 |
| **I DON'T LIKE** | 9 | 0.9 | 564 | 2.4 | 11.88 |
| **AND STUFF LIKE** | 9 | 0.9 | 59 | 0.3 | 9.37 |
| ANYTHING LIKE THAT | 9 | 0.9 | 117 | 0.6 | 2.50 |
| YOU KNOW LIKE | 9 | 0.9 | 264 | 1.1 | 0.45 |
| A BIT LIKE | 9 | 0.9 | 96 | 0.4 | 4.18 |
| **AND STUFF LIKE THAT** | 8 | 0.8 | 55 | 0.3 | 7.84 |
| I MEAN LIKE | 7 | 0.7 | 122 | 0.5 | 0.55 |
| LIKE THAT AND | 6 | 0.6 | 322 | 1.4 | 5.28 |
| OR SOMETHING LIKE | 6 | 0.6 | 254 | 1.1 | 2.44 |

Table 4: Pairwise comparison and LL of the most frequent SCO 3-4w *like* clusters in SCO and their MAC equivalents[150].

[150] See Appendix VIII for BNC/C figures.

In the following sections, I will look at all those areas of use where we find divergence between SCO and MAC proportional frequencies. I shall, first of all, look at the clusters where there is more convergence than difference (*like* functioning as a preference marker) then move to the usage where divergence is strongest (*like* as comparator).

### 9.2.3 LIKE as preference marker

If we look at *like* in its function as *preference marker* (for example *I like tea but I don't like black coffee*). In every single highly frequent *like* 3w cluster do we find a high degree of convergence where *like* is employed in this way, yet there is, in SCO, one interesting outlier.

| LIKE PREF. CLUSTER | SCO | % | MAC | % | LL |
|---|---|---|---|---|---|
| I LIKE THE | 11 | 1.2 | 269 | 1.1 | 0.01 |
| I LIKE THAT | 10 | 1.1 | 218 | 1.0 | 0.06 |
| I LIKE THAT I | 2 | 0.17 | 38 | 0.20 | n/a |
| I DON'T LIKE | 9 | 0.9 | 564 | 2.4 | 11.88 |
| LIKE I DON'T | 5 | 0.5 | 34 | 0.18 | 4.97 |
| I DON'T LIKE IT | 2 | 0.17 | 54 | 0.24 | n/a |

Table 5: *Like* to express preference compared in SCO and MAC.

Table 5 clearly demonstrates that *I like* 3w and 4w clusters have a high degree of convergent use in SCO and MAC. The same is true for *I like that (I)*. However, the negation of this phrase, *I don't like*, is significantly

(above the 99.9% mark) higher in its recorded proportional frequency in MAC - 2.6 times higher to be precise[151]. It must be noted (though the statistical evidence indicates on a low level of probable divergence) that the same holds for these three words when they cluster in a different word order - *like I don't*. Though this cluster is rarer in both corpora, it still occurs proportionally twice as often in SCO as it does in MAC.

### 9.2.5 *LIKE and the personal pronoun THEY*

The use of *like* with *they* presents diverse use. Table 6 below shows the most frequent *like* with *they* clusters in MAC and BNC/C. It shows how closely aligned in frequency the clusters in both corpora are.

| Cluster | Freq. BNC/C | % | Freq. MAC | % | Freq. SCO | % |
|---|---|---|---|---|---|---|
| THEY DON'T LIKE | 42 | 0.19 | 60 | 0.22 | 2 | 0.2 |
| THEY LOOK LIKE | 35 | 0.16 | 45 | 0.17 | 1 | 0.1 |
| LIKE THAT THEY | 33 | 0.15 | 60 | 0.22 | 2 | 0.2 |
| THEY WERE LIKE | 32 | 0.15 | 32 | 0.12 | 1 | 0.1 |
| THEY LIKE TO | 5 | >0.1 | 36 | 0.14 | 1 | 0.1 |

Table 6: *Like* with *they* top clusters in BNC/C and MAC compared to SCO

Moving on to the most frequent *like* with *they* clusters in SCO we see that the use of the clusters *like they have* and *they have like* are not closely aligned at all. As Table 7 below shows, *like they have,* although it appears in MAC, is very marginal, occurring only eleven times (0.048 % of all uses with *like*). *They have like* is even more

[151] The same ratio of use can be found in the BNC/C where *I like the* occurs 163 times (0.74% of aa uses of *like*) and *I don't like* appears 597 times (2.72%).

marginal, appearing only 3 times (0.013%). Similarly, in the BNC/C, *like they have* occurs only 9 times (0.045%), and *they have like* occurs only 5 times (0.023%). This gives the impression that these clusters are not very likely to be heard in English casual conversation. This is in contrast with occurrences in SCO. Again, these clusters are not highly frequent, yet *they have like* can be heard 515 times in 100.000 words in SCO as opposed to only 41.4 times in 100.000 words in MAC[152], and even less often in BNC/C. In other words, Liverpool speakers use this cluster up to twelve times more often. A phrase that is marginal in the MAC and BNC/C corpora is clearly recognisable in its use in SCO.

| **Cluster SCO** | **Freq./ (%)** | **example** |
|---|---|---|
| THEY HAVE LIKE | 5 (0.52%) | They have like three tickets like |
| AND THEY'RE LIKE | 5 (0.52%) | and they're like They're nice |
| THEY SAY LIKE | 3 (0.32%) | and they say like me ald fella |
| **Cluster MAC** | **Freq./ (%)** | **Log-Likelihood** |
| THEY HAVE LIKE | 3 (>0.1%) | n/a |
| AND THEY'RE LIKE | 5 (>0.1%) | 18.57 |

Table 7: LIKE *with* THEY top clusters (with examples) in SCO

As the quotations in Table 7 show, *like* is post-positioned, appearing to be discourse particles that are employed to give the speaker time to formulate the utterance.

I have also to add that these are only the most frequent occurrences of this pattern. In total, there are 93 *like* with *they* occurrences in SCO (nearly 10% of all uses of *like)* and most of them follow the " *like* is post-positioned " pattern above.

[152] *They have like* appears 3 times in MAC, 5 times in SCO; *and they're like* appears 5 times in both MAC and SCO; *they say like* does not occur in MAC.

To conclude: this appears to be a strongly divergent use of *like* (as discourse particle / filler) in our comparison of Liverpool Casual Spoken English with Casual Spoken English as represented in the MAC corpus. It is important to note that it appears to be fairly flexible as well – the core is a cluster of *like* with *they* in combination with a verb or conjunction.

While clusters of *like* with *they* found as standard in MAC and BNC/C are used with similar percentages in SCO, the far more frequent clusters (listed in Table 7) are typical of SCO only and are barely found in either of the other (much larger) corpora.

### 9.2.6 LIKE and past tense use

Focussing on *like* and terms that indicate past tense amongst the most frequent clusters, there is evidence that the past tense markers are used in a significant number of times in connection with *like*. Future tense markers, by contrast, are if recorded, very infrequent.

David Brazil points out in what way this can be seen as important:

> ... an oral narrative [is] a discourse type that is a not untypical outcome of a common kind of social activity: a single offering in the sort of anecdote-swapping session that makes up a significant part of many people's relaxed, everyday conversation.
>
> (Brazil 1995: 24)

Storytelling and therefore reference to past events are expected to occur in casual speech. As described later in Brazil's book, in the course of the story-telling, the speaker will switch between tenses, moving from past-tense to present tense to make action more tense and actual.

Trying to locate past-tense use, however, creates the complication of deciding between *like* as a filler (I was like – frightened) and *like* as used for comparison (I was like a frightened rabbit) when just looking a short clusters.

The use of *like* as filler + tense markers can only be determined by looking at the larger context. In the three corpora, *I was like* is used to buy time (*like* in the function of a filler) during storytelling (*…luckily* ***I was like*** *not in the rave part; … and if* ***I was like*** *having him).* It tends to be followed by a brief pause is typical of the usage of a filler ; *it was like* is mostly employed in the function of *like* a comparator (***it was like*** *that club;* ***it was like that*** *here*).

When we compare the most frequent clusters in SCO, it appears as if many of them use words to indicate past tense.

| CLUSTER | FRQ SCO | % | FRQ MAC | % | LL | FRQ BNC/C | % |
|---|---|---|---|---|---|---|---|
| I WAS LIKE | 15 | 1.5 | 30 | 0.13 | 41.25 | 64 | 0.3 |
| IT WAS LIKE | 14 | 1.4 | 148 | 0.65 | 6.63 | 172 | 0.8 |
| HE WAS LIKE | 2 | 0.2 | 37 | 0.16 | n/a | 61 | 0.3 |
| WAS LIKE THAT | 5 | 0.5 | 35 | 0.15 | 4.78 | 45 | 0.2 |

Table 8: *was like* cluster comparison (percentages proportional to LIKE total) in SCO, MAC and BNC/C.

To be more specific: *It was like* and *I was like* together make up about three percent of all uses of *like* in SCO (see Table 8). Compare this to the

occurrences for MAC: *I was like* as a cluster is marginal (just over 0.1%). The most frequently recorded cluster in MAC, using past tense, is *it was like* which has no high occurrence either. It is used in 0.65% of all clusters with *like* and 0.8% of all *like* occurrences in BNC/C. This means Liverpool speakers (where the recorded frequency of use is 1.4%) employ it around twice as often than speakers across the UK.

On the whole, Liverpool speakers tend to use *like* with the past tense *marker was* significantly more often in connection with the term *like* (as filler and stress-indicator) than other UK speakers do.

### 9.2.3 LIKE with vague terms

In this section, we look at *like's* use as comparator. The top clusters in MAC that include vagueness markers, *something like that* (and also *things like that),* have a clearly higher frequency than the top cluster with a vagueness marker in SCO (*stuff like that*). [153] [154]

The second, probably more important, observation is that all the most frequent clusters of *like* as a discourse particle in both corpora co-occur with a term for an object – *stuff; thing*; *something; anything*[155].

[153] That SCO use of *like* 3w clusters is different is also supported by the evidence presented in Appendix VIII which shows that BNC/C 3w cluster frequencies align more closely with MAC than with SCO.

[154] Seeing the relative preference of the word *stuff,* it comes probably as no surprise that the Scouse Stand-up comedian *Alexei Sayle* called his late 1980s BBC show *Stuff.*

[155] It is very interesting to see how these terms are defined in *Macmillan English Dictionary* and how their frequencies indicated there are in line with the above findings!

Table 8 (below) looks at all the clusters of *like* with vagueness markers and undertakes a log-likelihood test for each pairwise comparison. Thus, while we find that while the clusters *or something like* and *something like that* are clearly more frequent in MAC (twice as frequent), the fact remains that these clusters are amongst the most frequent with *like* in SCO as well and there is no significant divergence of use.

Streek (2002) gives an explanation for the use of *something like* -

> In combination with *something*, *like* can be used to append various kinds of components to units of talk. Each time, then, like postpones the choice point at which the speaker must commit to a grammatical frame for the rest of the sentence. (Streek 2002: 586)

*Like* in this position is, in other words, a preposition.

| LIKE *vague* cluster | SCO occ. | SCO % | MAC Occ. | MAC (%) | LL |
|---|---|---|---|---|---|
| LIKE YOU KNOW | 16 | 1.7 | 264 | 1.1 | 1.73 |
| *STUFF LIKE* | 20 | 2.1 | 103 | 0.6 | 27.40 |
| *STUFF LIKE THAT* | 16 | 1.7 | 81 | 0.4 | 22.30 |
| *AND STUFF LIKE* | 9 | 0.9 | 59 | 0.3 | 9.37 |
| *AND STUFF LIKE THAT* | 8 | 0.8 | 55 | 0.3 | 7.84 |
| THINGS LIKE THAT[1] | 12 | 1.3 | 364 | 1.6 | 0.82 |
| ANYTHING LIKE THAT | 9 | 0.9 | 117 | 0.6 | 2.50 |
| SOMETHING LIKE THAT | 11 | 1.2 | 555 | 2.3 | 8.08 |
| OR SOMETHING LIKE | 6 | 0.6 | 254 | 1.1 | 2.44 |
| OR SOMETHING LIKE THAT | 5 | 0.52 | 234 | 0.88 | 2.90 |

Table 9: Comparative use *Like* with vagueness markers in SCO and MAC

One difference lies in the use of another lexical item by Liverpool speakers that appears to strongly substitute for another used

by other UK speakers. Liverpool speakers use *stuff like that* and *and stuff like* as frequently as speakers across the UK would say *things like that* and *and things like*.

This is most clearly shown in Table 9. In MAC and SCO, the clusters *things like that* and *anything like that* are used with about the same proportional frequencies and there is little significant divergence between them. The differences seem stronger for the uses of *like* with *something* and *anything*. This appears to be connected to the fact that in SCO 2w, 3w and 4w clusters incorporating *stuff like* (including *and stuff like*) are significantly more prominent in their use than in MAC. This is, to a degree[156] also relevant for the use of the phrase *something like that* which is, proportionally, significantly less used in SCO compared to MAC.

The percentages in Table 9 show that, while *things like that* occurs with the same frequency in both corpora in clusters with *like*, the use of *stuff* in such clusters is marginal in MAC. There, the use of *things like that* is 4 times as frequent as *stuff like that*. Similarly, the use of *and things like* occurs 4 times as often as the use of *and stuff like*. By contrast, the first phrase uses *stuff* more frequently in the Liverpool corpus than *things*, while *and stuff like* is used more than twice as often than *and things like*. Table 8 also shows that clusters of *like* with *stuff* and *things* do appear in the BNC/C and are similar in its proportional

[156] degree of likelihood of above 99.0%

frequency (although slightly lower) as in MAC. The Liverpool speakers use *stuff like that* proportionally six times as often and *and stuff like* proportionally nearly four times as often as speakers in BNC/C.

This appears to show an inversion of use when Liverpool speakers are compared to UK speakers as a whole. It also indicates that both forms exist in parallel use amongst Liverpool speakers, whereas most UK speakers appear to be largely restricted to one phrase.

While speakers in all three corpora employ the phrase *things like that* with the same proportional frequency, the alternative formulation *stuff like that* is also used – yet it appears to be significantly preferred by SCO speakers. Comparing the use of vagueness markers *(something; stuff; things)* with *and,* we see that *and things like* seems, while recorded in all three corpora, to be the preferred cluster found in MAC and BNC/C. In SCO, *and things like* is used noticeably less often (three times less than MAC, half as often than BNC/C). In SCO, on the other hand, we find the cluster *and stuff like* is proportionally nearly three times as often used than in MAC and proportionally nearly four times than in BNC/C.

Looking at longer clusters, we see that there is a preference in SCO to use certain vagueness markers that are less prominent in their use in MAC or BNC/C. Unlike *or something like that*, which is highly frequent in all three corpora, *or anything like that* is proportionally used nearly as often as *something like …* in SCO, while it is used considerably less

than *something like* … in the comparators. Both clusters show the use of *like* in a long cluster where it is used to describe a *thing* without wishing to be more specific. In casual spoken English, this can be seen as a salient feature of use, particularly, it would appear, favoured by SCO speakers. Furthermore, the related cluster, *and everything like that* occurs as often as *or anything like …* (4 times in total, the equivalent of 368 times in every 100.000 occurrences of *like*) but is barely at all recorded in either MAC or BNC/C. Another cluster that uses this collocational format is *and stuff like that.* In SCO *and stuff like that* is by far the most frequent *like* 4w cluster, appearing 8 times. This is half of the occurrences of the most frequent 3w cluster with like in SCO, *stuff like that. And stuff like that* is proportionally 4 times more frequent in SCO than in either MAC or BNC/C[157].

While *like* with *vagueness marker* and with *that* is the most common occurrence of the word *like* and shows its use as comparator, it is the difference in the vagueness markers (in 3-5w clusters with *like)* that shows the divergence of use between SCO and MAC or BNC/C. This hints at preferred nesting for a certain set of terms displayed by SCO speakers.

### 9.2.6 Conclusions about the use of LIKE

When looking at *like*, we find that the term is very frequently employed in its function of discourse marker in spoken English. At the

[157] See Table 11 for a complete breakdown of all figures.

same time, *like* also plays a role as comparator or to express preference or as filler, while the speaker is trying to piece together a coherent statement. Given the high frequency of use of *like* and the many ways it can be employed, it is a key item to investigate difference of use between speech communities.

*Like* appears to be delexicalised to a point where it can be employed as a functional verb, an adjective, a filler and a downtoner. In this last role it can, appear in a meaning that is close to its original meaning (i.e. can be substituted with "*It is all equal*", in other words, *it does not matter*).

When looking at the use of *like* as a comparator, a word to express preference, or as a discourse marker in MAC and BNC/C corpora (standing for use by speakers from across the UK) and SCO corpus (for Liverpool English speakers) the following findings have been made:

*Like* appears mainly as a comparator. We therefore find the clusters *something like that* and *things like that* in MAC, whereas in SCO, it is *stuff like that*. Given that there are other clusters where the corpora differ in a similar way (*and stuff like* in SCO *and things like* in MAC) this seems to indicate a case of verbal substitution – *stuff* for *things.*

A strong difference occurs when *like* is used in connection with *they* as a filler or intensifier. These are marginal in MAC – and, where they

occur, *like* is pre-positioned. However, 10% of all uses of *like* appear in SCO with *they* – and here *like* is post-positioned. Though the total numbers are low for all three corpora, the use of these two words together in a 3w cluster is entirely different in SCO from its use in MAC. As with some of the 4w clusters discussed in 10.2.8, this points towards different colligational use and an entirely different set of semantic associations that Scouse speakers connect with *like* in combination with *they* and *and*.

The use of *like* as clause-final marker is prevalent in *like* with *past tense* use. Consequently, we find the clusters *I was like* and *it was like* are ten times more frequent in SCO than in MAC (or BNC/C).

In this context, it is worth noting that in SCO, the use of *like* is found to be a clause-final discourse marker overall in a far higher proportion of cases than in MAC or BNC/C. This in itself is an important marker, as Miller and Weinert point out:

> The two major LIKE constructions - clause-initial and clause-final LIKE - have different discourse roles. In general LIKE is a non-introducing, non-contrastive focuser which may focus on new or given information. In addition, clause-initial LIKE is concerned with the elucidation of previous comments, whereas clause-final LIKE is concerned with countering objections and assumptions. (Miller & Weinert 1995: 392)

"Countering objections and assumptions" may sound rather strong, given that their discussed data has items like -

> (27) A2: mostly in Edinburgh *like*? -
>
> (27-A2) asks for (dis)confirmation of R's assumption (Ibid.)

It may therefore be more appropriate to speak about clause-*final* LIKE as a way of mitigating or softening any previous statement. That this type of *like* usage is predominant in SCO hints that Liverpool speakers try to avoid strong, unalterable, finite statements.

Work on *very, really* and other intensifiers and discourse particles like *just, well,* and in some of its uses, *like* in spoken English confirms that Spoken Liverpool English (in SCO) provides a sample of how collocational, colligational and nesting pattern differences can be a seen as another way of describing (and differentiating between) dialects.

# Chapter 10 Clusters

## 10.1 Introduction

Up to this point, we have looked at individual words and how they collocate with other words. These collocates form clusters and we have looked at several instances in the two corpora where individual words have been used differently, with different collocates in SCO, when compared to MAC (and, partly, BNC/C).

Clusters are an essential part of structuring language. This is true for both language production and language processing. Outside the realm of language studies word clustering is researched by brain specialists, cognitive scientists and information theorists (who look at artificial language systems) amongst others.

The concept (also known as *chunking)* has been discussed for over half a century and has been established as a principle that can be found not just amongst humans:

> Pioneering work in the 1940s and 1950s suggested that the concept of 'chunking' might be important in many processes of perception, learning and cognition in humans and animals. (Gobet *et al.* 2001: 236)

Gobet *et al.* make a specific link, based on their research results, between chunking and individual words:

> Words associated by generative links[158] form groups, which approximate more formal syntactic categories. (Gobet *et al*, 2001: 241)

This links in with the work done by linguists like Biber *et al.* (2002), Wray (2002a & 2002b), Hoey (2005), Millar (2009) and others with reference to prefabricated chunks and retrieval from memory.

### 10.2.1 Frequent cluster groups in SCO

In this chapter, the most frequently occurring clusters in SCO will be looked at.[159] To chose the key clusters, a went through a process involving a number of steps. First of all, the most frequent clusters were determined with the help of WordSmith. As a next step, the focus was on words appearing more than once. These words and the 3-word to 5-word clusters they are found in will be, below, referred to as *groups*. The reason for this cluster length is twofold: First of all, two-word clusters, collocates, have been intensively discussed throughout this thesis. Secondly, beyond the length of five or six words, clusters do not appear in relevant numbers. Strangert (2004) presents a graph showing that in her work (on pauses between meaningful chunks) there is a visible drop in

[158] Generative links associate two nodes that have similar descendant test links.
[159] All percentage figures are based on the WordSmith calculations – that means they give percentage of occurrence of a cluster compared to the total number of single words in the corpus. Consequently, all values appear below the 1% mark. (See also Appendix II).

occurrence numbers beyond six-word chunks. Corpus linguists who work with spoken language corpora present similar evidence:

> Six-word recurrent chunks are of very low frequency in CANCODE, and it does appear that six is a practical cut-off point beyond which such chunks seem rare.
>
> (O'Keefe *et al*. 2007: 64)

In Table 1 the fifty most frequent (3-5 word) clusters in SCO highlight which particular words re-occur in different combinations. The frequency of clusters found in normalised per 100.000 words[160].

| Cluster | Freq. | per 100.000 | core term |
|---|---|---|---|
| I DON'T KNOW | 97 | 81.5 | **KNOW** |
| YOU KNOW WHAT | 62 | 52.1 | **KNOW** |
| A LOT OF | 60 | 50.4 | **OF** |
| WHAT I MEAN | 55 | 46.2 | **MEAN** |
| KNOW WHAT I | 49 | 41.1 | **KNOW** |
| YOU KNOW WHAT I | 47 | 39.5 | **KNOW** |
| KNOW WHAT I MEAN | 46 | 38.8 | **KNOW/MEAN** |
| YOU KNOW WHAT I MEAN | 45 | 37.2 | **KNOW/MEAN** |
| YOU HAVE TO | 34 | 28.6 | **TO** |
| I DON'T THINK | 31 | 26.0 | **THINK** |
| USED TO BE | 27 | 22.7 | **TO** |
| A BIT OF | 22 | 18.5 | **OF** |
| A COUPLE OF | 21 | 17.6 | **OF** |
| I HAVE TO | 19 | 16.0 | **TO** |
| I USED TO | 19 | 16.0 | **TO** |
| STUFF LIKE THAT | 16 | 13.4 | **STUFF/LIKE** |
| TO BE HONEST | 16 | 13.4 | **HONEST** |

Table 1: SCO selection of most frequent 3-5w clusters
(full table in Appendix IX.1)

A first step to determine which groups clusters are predominant in SCO is to look at the most frequent 3 to 5 word long cluster-groups in SCO. Here, it becomes apparent that certain expressions keep re-occurring

[160] Compare Table 1 with the most frequent long clusters found in CANCODE - see Appendix IX.4

throughout in clusters of different length, and in clusters where they occur with different words.[161] So we have, for example *you know what* (62 occ.) as well as *what I mean* (55 occ.) as 3w clusters. There is also the 5w cluster *you know what I mean* (45 occ.). In other words, 45 of the 55 uses of *what I mean* are found in the longer cluster *you know what I mean*. It is only the difference in occurrence numbers that hint that the shorter clusters also appear in different combinations.

As discussed above, certain words are very free in the way that they collocate or do not collocate with other words – and their high total frequency means they will appear in clusters of all corpora. These include words like *the, you, I, and,* etc.[162]. However, the fourth column in Table 1[163] highlights the *core term* in the clusters and it becomes apparent that some words can be frequent, while they are restricted to appearing mostly in one cluster (like *mean*). Alternatively, a core term can be frequent and appear in a variety of clusters (like *know*) and there are also constructions with *of* and to that appear time and again. Though these are words of a different kind, it is the collocational patterns rather than the grammatical features of the words examined that this chapter focuses on.

[161] One re-current feature in a certain number of clusters is the PAUSE. Given that pauses can be found throughout in spoken English (though, in my own observation, hesitance markers do not seem to be prominent in Mandarin. This cast doubt whether pausing is universal in human speech.), they can be seen as a salient feature. An argument could be made about how far speakers are primed to pause after certain spoken key terms. However, while my Liverpool English corpus highlights PAUSE as a feature that will appear in any cluster, other corpora do not indicate PAUSE as a feature that appears in a cluster. Consequently, comparisons with regard to this feature are not available.

[162] All of these words are, in fact, more frequent in the BNC/C than in SCO. Earlier chapters show that these words are proportionally more frequent in MAC than in SCO, too.

[163] See Appendix IX.3 for a full version of this table.

| Key cluster (3-5w) | SCO Freq. | SCO token per 100.000 | BNC/C Freq. | BNC/C token per 100.000 | Keyness |
|---|---|---|---|---|---|
| DO YOU WANT | 20 | 16.8 | 1920 | 47.6 | -31.03 |
| WHAT I MEAN | 55 | 46.2 | 619 | 15.4 | 45.68 |
| KNOW WHAT I | 49 | 41.2 | 515 | 12.8 | 44.99 |
| YOU KNOW WHAT I | 47 | 39.5 | 375 | n/a | 60.81 |
| KNOW WHAT I MEAN | 46 | 38.6 | 372 | n/a | 58.64 |
| YOU KNOW WHAT I MEAN | 45 | 37.8 | 328 | n/a | 64.07 |
| STUFF LIKE THAT | 16 | 13.4 | 69 | n/a | 35.45 |
| I WAS LIKE | 15 | 12.6 | 62 | n/a | 34.22 |

*Table 2(a): 3-5 word SCO cluster Keyness when compared to BNC/C clusters*[1]

Table 2(a): 3-5 word SCO cluster Keyness when compared to BNC/C clusters[164]

[164] See Appendix IX.3 for Table 2(b) where BNC/C negative Keyness compared to SCO is shown.

The next step is to look in what way the most frequent 3w-5w clusters found in SCO differ from the most frequent clusters appearing in other Spoken English corpora. To have a first overview whether the use of these clusters in SCO are markedly different from what other English Spoken corpora provide, I compare them to those found in the BNC/C. To find the most frequent clusters within a (sub-)corpus, the full corpus is needed to make the calculations. The initial comparison was made between SCO and BNC/C, as I had had no access to the full MAC corpus. The next step was to compare Key clusters found in SCO with the occurrence of these clusters in both BNC/C and MAC.

Table 2(a) shows the seven clusters that are positively Key when SCO is compared to BNC/C. Furthermore, there is the one cluster which is negative Key, i.e. Key in BNC/C but not in SCO: *do you want*. I will come back to this cluster later.

Looking at Tables 1 and 2(a), we see that clusters related to the key term *know* are in prominent use in SCO. The other three key terms are *mean*, *stuff* and *like*. Table 1 also shows that the *to* and of clusters are highly frequent, as are the clusters around the core terms think and *honest*.

Of the above key terms, *of* proved to be least revealing. When *of* clusters were compared with their occurrence pattern in BNC/C, a very

high degree of agreement is shown[165]. Not only do we find the same clusters with roughly the same proportional frequencies here, but also the relative frequency of use of the clusters to each other (the *rank*) is in high agreement. Sinclair *et al.* (1998b) class these in as **N *of* pl-n** (The "Gang" Group) and it seems a very stable construction in the English language.[166]

There are certain groups of recurring clusters in SCO which are listed here. They are ordered by the relative frequency of clusters with these core terms.

1) The *KNOW* group.

In SCO, *know* occurs 949 times.

Carter and McCarthy (2006)[167], in their discussion of CANCODE, point out that *I don't know* is the highest occurring cluster in spoken English corpora. This is also true for SCO. While *I know* and *I don't know* is used to start a variety of utterances, in SCO such clusters are often used to seek reassurance. In those cases, the *KNOW* group often comes together with the *MEAN* group. A third use that is relatively frequent is represented by phrases like *you know what*.[168]

[165] See Appendix IX.2 for Table 1(b) that compares SCO and BNC/C most frequent OF clusters.

[166] It would be interesting to see whether the pattern can be found to be stable when compared in a variety of spoken and written corpora. This, however, would be an entirely different project – one which John Sinclair already flagged up: Sinclair, J. M. (1991: cpt. 6).

[167] In *The Cambridge Grammar of English* the use of I DON'T KNOW is referred to with very good graphs comparing spoken to written use in An Introduction to Corpora in English Language Teaching: http://www.ed2go.com/elt_demo/3ce_demo/L03.htm
See also Tables 2-4 for CIC corpus most frequent clusters (O'Keefe *et al* 2007) in the Appendix IX.4 of this chapter.

[168] In SCO, YOU KNOW WHAT is mostly part of the cluster YOU KNOW WHAT I MEAN. As a three-word phrase it is either an attention-marker (e.g.YOU KNOW WHAT happened) – 8 times in SCO; or a question (e.g. YOU KNOW WHAT to do) - 5 times in SCO.

2) The *MEAN* group.

*Mean* occurs 243 times in SCO. As shown above, *mean* is strongly linked in its SCO occurrence to the *KNOW* group. Apart from that *mean* is also used for further explanation.

3) The *LIKE* group.

*Like* appears 970 times in SCO. Earlier chapters have already highlighted that *like* is a key term in Liverpool Spoken English. It occurs more often with *stuff* or *things* than with *know*.

4) The *THINK* group.

*Think* occurs 269 times in SCO. The group also includes *thought* (86 occurrences). Again, *think/thought* collocations and clusters have been discussed earlier. *Think* usually occurs with "I" and is used in narratives.

5) The *TO* group.

*To* appears 152 times in SCO. Sinclair *et al* (1998a) points out that **V *to*** clusters appear in a variety of simple and complex patterns.

6) The *HONEST* group.

In SCO, *honest* appears only 18 times. 16 of these occurrences are in the cluster *to be honest*.

These six groups stand out as appearing and re-appearing all over the main clusters within SCO.

### *10.2.2 A broad comparison of SCO's most frequent clusters with those in BNC/C*

In 10.2.1, we looked at which clusters are very frequent in their use in SCO as well as *key* compared to BNC/C clusters in order to determine which cluster groups to focus on for the direct comparison. Table 3 compares some of the most prominent clusters in SCO with their proportional occurrence pattern in the BNC/C. To make comparisons easier, the percentage of clusters found in each corpus is normalised per 100.000.

Table 3: SCO highest frequency clusters compared to BNC/C frequencies

| SCO or BNC/C cluster | SCO Freq. | Per 100.000 | BNC/C Freq. | Per 100.000 | log likelihood |
|---|---|---|---|---|---|
| I DON'T KNOW | 97 | 81 | 4651 | 115.5 | 13.07 |
| YOU KNOW WHAT I MEAN | 46 | 39 | 326 | 8.1 | 67.17 |
| I DON'T THINK | 31 | 26 | 1937 | 48 | 14.20 |
| STUFF LIKE THAT | 16 | 15 | 68 | 1.5 | 35.74 |
| TO BE HONEST | 16 | 15 | 109 | 2.7 | 24.29 |
| THINGS LIKE THAT | 12 | 10 | 266 | 6.5 | 1.80 |
| THAT'S RIGHT YEAH | 12 | 10 | 307 | 7 | 0.82 |
| SOMETHING LIKE THAT | 11 | 9 | 466 | 11.6 | 0.59 |
| SORT OF THING | 10 | 8 | 396 | 9.7 | 0.26 |
| YOU KNOW LIKE | 8 | 7 | 236 | 6.3 | 0.14 |
| I JUST THOUGHT | 6 | 5 | 91 | 2.3 | 2.88 |

*Shaded area: SCO higher*

Table 3 : SCO highest frequency clusters compared to BNC/C frequencies

Looking first at the two highest occurring SCO clusters, it is very clear that *I don't know*, though the most frequent cluster in both corpora, is still markedly less frequent in SCO than in BNC/C. (The difference is significant above the 99.9% level) There is also one other (high occurring) cluster that is used nearly twice as often in BNC/C than SCO: *I don't think*.

Conversely, there are three clusters that are more commonly used in SCO but are significantly found to be rare in BNC/C: *you know what I mean* (nearly five times more frequent); *stuff like that* (ten times more frequent); and *to be honest* (also more than five times more frequent in SCO).

This initial comparison with a control corpus indicates that looking at high-frequency clusters may yield results that encourage further research.

### 10.2.3 A closer comparison: with MAC

In this section, I will compare SCO cluster frequency of occurrence with the MAC equivalent. The point of comparison will be the main cluster groups in SCO identified in section 10.2.1 – and I shall adhere to the order presented there, which is based on frequency of the target word within the cluster (i.e. *know* has highest; *honest* has lowest frequency within SCO). The initial focus will be on the highest-occurring clusters in SCO and whether these clusters are in similar use within MAC. To

achieve a higher level of validation, the clusters in question will also be compared with those of the BNC/C, in respect of their occurrence patterns.

As a further step, the highest-frequency clusters within each of the groups will be compared across all three corpora. This second, different angle on occurrence patterns will achieve two ends: one, it will re-affirm the prominence of expressions that are found to occur far more or far less frequently in either SCO or the UK-wide corpora. Two, it will show to what extent different expressions give different prominence to core words in different corpora.

## 10.3 The *KNOW* group

In spoken English, the *KNOW* group yields the largest set of clusters. *Know* prominently appears in the phrase *you know*. Macaulay (2002) points out that *you know* has received wide attention and gives an overview of the most important work on this phrase. Using a corpus similar to SCO, with informants from the Scottish towns of Ayr and Glasgow, Macaulay comes to the following, tentative, conclusions:

(…)

(2) Speakers are more likely to use *you know* in conversations with an acquaintance than in interviews with a stranger. (…)

(5) The use of *you know* is not more common in one social class than the other. However, middle-class speakers are more likely to use *you know* medially in an utterance for purposes of self-repair or elaboration, while working-class speakers mainly use *you know* at the end of an utterance.

(6) The use of *you know* does not appear to be primarily based on assumptions of shared knowledge, but rather to form part of the speaker's discourse style and the rhythmic organization of utterances, particularly when it is used at the end of an utterance.

(Macaulay 2002: 765f.)

Point (2) highlights the informality and relaxed attitude that is reflected by the use of *you know*. In (5) and (6) Macaulay refutes the notion that *you know* reflects either class or the "assumption of shared knowledge". It has a stronger discursive character and reflects the flow of (spoken) language. We have seen earlier (in chapter 8), that *yeah* and *like* seem to have similar functions in casual spoken Liverpool English. While Macaulay's findings are relevant to this thesis, I will focus here on 3w clusters with *know* as core word. The comparison is of occurrence patterns that may differ between Liverpool and UK speakers. A first comparison of the most frequent clusters with *know* already highlights that, despite clear overlap, there are high-frequency clusters in each corpus that do not appear at the same rate of frequency in the other cluster.

Table 4: Highest frequency clusters KNOW in SCO compared to MAC clusters

| Word Cluster SCO & MAC | Freq. : SCO | % : SCO | per 100.000 : SCO | per 100.000 : MAC | Freq. : MAC | log likelihood |
|---|---|---|---|---|---|---|
| YOU KNOW | 619 | 0.5198 | 519.8 | 605.6 | 19,984 | 14.69 |
| *I DON'T KNOW* | *97* | *0.0814* | *81.45* | *135.4* | *4469* | 29.01 |
| YOU KNOW WHAT | 62 | 0.0520 | 52.1 | 34.5 | 1141 | 8.74 |
| KNOW WHAT I | 49 | 0.0411 | 41.1 | 16.5 | 545 | 29.30 |
| YOU KNOW WHAT I MEAN | 46 | 0.0386 | 38.6 | 8.9 | 292 | 60.66 |
| YOU KNOW YOU | 21 | 0.0176 | 17.6 | 44.3 | 1462 | 24.15 |
| *DON'T KNOW WHAT* | *17* | *0.0142* | *14.2* | *10.1* | *333* | 1.76 |
| **DO YOU KNOW** | **16** | **0.0134** | **13.4** | **32.6** | **1075** | 16.77 |
| LIKE YOU KNOW | 14 | 0.0117 | 11.75 | 10.2 | 336 | 0.27 |
| YOU KNOW WHEN | 13 | 0.0100 | 10.8 | 14.2 | 463 | 0.86 |
| YEAH I KNOW | 12 | 0.0100 | 10 | 21.6 | 712 | 8.86 |
| *I DON'T KNOW WHAT* | *12* | *0.0100* | *10* | *16.6* | *571* | 4.11 |
| THAT YOU KNOW | 11 | ~ | 9.2 | 9.2 | 304 | 0.00 |
| YOU KNOW LIKE | 8 | ~ | 6.7 | 5.2 | 170 | 0.50 |
| YOU KNOW THAT | 7 | ~ | 5.8 | 24.5 | 810 | 23.85 |
| *DON'T KNOW WHETHER* | *7* | ~ | *5.8* | *12.1* | *399* | 4.57 |
| *Proportional % freq. higher* | | | | *Proportional % freq. lower* | | |

Table 4: Highest frequency clusters KNOW in SCO compared to MAC clusters

### 10.3.1 SCO's Most frequent KNOW group clusters compared

In SCO there are *KNOW* group clusters that also occur in MAC. The major difference is the relative frequency. In Table 4, the normalised use (per 100,000 words) of the most frequent clusters that are found in both corpora are juxtaposed.

There are a greater number of clusters listed that are used proportionally more often in MAC than in SCO. While MAC does contain the same *KNOW* group clusters as SCO, the two corpora only match with regard to the normalised frequency of one cluster: *that you know* is used to the same extent in both corpora. Two further clusters are close enough in normalised frequency in the two corpora to be considered insignificantly different: *like you know* and you *know like.*

While a number of clusters in SCO (particularly the most common ones) are most of the time part of longer expressions (for example, *you know what* and know *what I* are mainly part of *you know what I (mean)* this is not the case for MAC. *Know what I* in SCO appears 49 times, meaning it is almost always constituent part in the 45 occurrences of the longer expression[169]. In MAC, however, it occurs almost twice as often as *you know what I* mean and that cluster is therefore clearly also part of other expressions. This can be seen as evidence that the phrase *you know what I mean* is used differently in SCO.

[169] See also O'Donnell (2009) on how segments of larger clusters that appear with high frequency as shorter cluster should be weighted.

Fox-Tree and Schrock (Fox Tree & Schrock: 2002) point out that -

> *You know* and *I mean* occur frequently in conversation because their functions are tied to the naturalistic, unplanned, unrehearsed, collaborative nature of spontaneous talk (…) speakers are motivated to invite addressees to fill out their inferences by saying *you know* or to forewarn upcoming adjustments by saying *I mean.*
>
> (Fox Tree & Schrock 2002: 323)

This confirms that *you know* and *I mean* have discourse functions. However, in their paper they also argue (as the quote above indicates) that the two phrases have different places and functions within conversation. Their argument, however, is rendered immaterial by the phrase *you know what I mean* which incorporates both expressions. Consequently, any evidence of preference for the phrase *you know what I mean* can therefore be seen as a difference that is both colligational and reflective of a different semantic association.

Within the *KNOW* group *you know what I mean* has the largest proportional difference and also the most significant difference between SCO and MAC. It is used 38.6 times within every 100,000 words in SCO, while it is used 8.9 times / 100,000 words in MAC. This may indicate that the literature cited by Brinton (2003) is more relevant to the use of this longer phrase:

> *I mean* also expresses of range of speaker attitudes. For example, it may function as a "softener" (Crystal and Davy 1975), as a "compromiser" (James 1983) softening the assertive force, or as a mitigator of "the strength of an evaluative statement" by making the speaker less committed (Erman 1986: 143; 1987: 119). It has been argued that as a

> "cajoler" *I mean* increases, establishes, or restores harmony between interlocutors; it is interactive, cooperative, and hearer-oriented, thus contributing to intimacy.
>
> (Brinton 2003: 2f.)

This would possibly make more sense in the context of SCO use, as we have observed other tendencies to *soften* or *mitigate* a statement.[170]

This could also be taken as a possible explanation for the fact that the expression *you know that*, though not amongst the most frequent clusters in the group, is used over four times more often in MAC. *You know that* appears more assertive than *You know what I mean*. With 24.5 times in MAC as opposed to 5.8 times per 100,000 words in SCO, a difference of proportional frequency is clear and significant.

Table 3 shows also (highlighted through the use of italics and bold type) the use of *know* with the negative and as a straightforward question. In this comparison, *know* with a negative ("Don't know whether…"; "I don't' know what") is used more frequently in MAC, though the differences are not significant. Where it is significantly diverging however, is as an apparently straightforward question: *do you know* occurring more than twice as often (32.6 times in 100,000 words in MAC; 13.4 times in SCO).

[170] Brinton (2003: 9) actually quotes samples that contain *you know what I mean* in his paper but does not discuss their occurrence in particular.

### 10.3.2 The most frequent KNOW group clusters

Another way of exploring the data is, as noted before, to see what particular expressions (clusters) are the highest occurring in the three available corpora (SCO, MAC & BNC/C). We can expect to find certain clusters amongst the most frequent in all corpora, e.g. the expressions *you know* and *I don't know* (both of which are significantly more frequent, proportionally in MAC than in SCO). However, the comparison of the most frequent *KNOW* group clusters also show expressions that have a strong presence in only one or two corpora.

In this case, I have chosen the first 15 most frequent *KNOW* group 3+w clusters for direct comparison. This results in the inclusion of all the expressions incorporating *know* that appear more than 10 times within 100.000 words in SCO and BNC. *Know* is considerably more frequent in MAC than in SCO (800 times per 100k in SCO; 1164 times per 100k in MAC) and the top clusters tend to be more frequent in MAC. Still we can still see that the findings shown in 10.3.1 are confirmed: *you know what I mean* is significantly more prominent in SCO, *you know that* is significantly less prominent.

Table 5(a): Highest frequency KNOW group clusters in SCO, MAC, BNC/C

| Cluster SCO | Freq. | per 100 k | Cluster MAC | Freq. | per 100 k | Cluster BNC/C | Freq. | per 100k |
|---|---|---|---|---|---|---|---|---|
| I DON'T KNOW | 97 | 81.45 | I DON 'T KNOW | 4469 | 135.4 | I DON'T KNOW | 4657 | 115.5 |
| YOU KNOW WHAT | 62 | 52.1 | YOU KNOW I | 1472 | 44.4 | DO YOU KNOW | 1037 | 25.7 |
| WHAT I MEAN | 55 | 46.2 | YOU KNOW YOU | 1462 | 44.3 | YOU KNOW WHAT | 970 | 24 |
| KNOW WHAT I | 49 | 41.2 | I KNOW I | 1152 | 34.6 | YOU KNOW I | 879 | 21.8 |
| YOU KNOW WHAT I | 48 | 40.3 | YOU KNOW WHAT | 1141 | 34.5 | DON'T KNOW WHAT | 844 | 20.9 |
| KNOW WHAT I MEAN | 46 | 38.6 | DO YOU KNOW | 1075 | 32.6 | YEAH I KNOW | 838 | 20.7 |
| YOU KNOW WHAT I MEAN | 46 | 38.6 | YOU KNOW THE | 1048 | 32.8 | YOU KNOW THE | 728 | 18.1 |
| YOU KNOW YOU | 21 | 17.6 | YOU KNOW IT | 891 | 27 | I DON'T KNOW WHAT | 614 | 15.2 |
| DON'T KNOW WHAT | 17 | 14.3 | KNOW YOU KNOW | 874 | 26.4 | KNOW WHAT I | 515 | 12.8 |
| DO YOU KNOW | 16 | 13.4 | DON'T KNOW WHAT | 871 | 26.4 | YOU KNOW AND | 506 | 12.5 |
| LIKE YOU KNOW | 14 | 11.75 | YOU KNOW AND | 863 | 25.9 | YOU KNOW YOU | 504 | 12.5 |
| YOU KNOW WHEN | 13 | 10.9 | YOU KNOW THAT | 810 | 24.5 | YOU KNOW THAT | 500 | 12.4 |
| YOU KNOW AND | 12 | 10 | YEAH I KNOW | 712 | 21.6 | I DIDN'T KNOW | 482 | 12 |
| YEAH I KNOW | 12 | 10 | DON 'T KNOW I | 619 | 18.76 | I KNOW BUT | 456 | 11 |
| I DON'T KNOW WHAT | 12 | 10 | AND YOU KNOW | 601 | 18.2 | I KNOW I | 455 | 11 |

Table 5(a): Highest frequency KNOW group clusters in SCO, MAC, BNC/C

Table 5(a) gives a complete overview of the 15 most frequent 3-5w *KNOW* group clusters in the three corpora under discussion. The same colours are given to the same clusters, and the most frequent cluster of each corpus in the *KNOW* group is at the top of the list. Table 5(a) also shows through colour coding the expressions which are the same in at least two of the three corpora.

The differences between the Liverpool SCO and the two comparators are obvious. All three have *you know what* as amongst the most frequent 3-word expressions in the *KNOW* group. Yet it is considerably more frequently used amongst Scouse speakers than others (52.1 per 100k compared to 34.5 in MAC; 15.5 in BNC/C). Still more important is the strong presence in SCO of the expression identified as a keyphrase earlier, *you know what I mean*, which does not occur in the top fifteen clusters of the other two corpora. This is also by far the longest cluster in the high-frequency selection.

Another phrase also stands out as being significantly more frequent in use in SCO but only rarely used in MAC and BNC/C: *like you know*.

Table 5(b): Highest frequency KNOW group clusters in a by occurrence rank.

| Cluster | Rank SCO | Freq. | per 100 k | Rank MAC | Freq. MAC | per 100 k | LL | Rank BNC/C | Freq. BNC/C | per 100k |
|---|---|---|---|---|---|---|---|---|---|---|
| I DON'T KNOW | 1 | 97 | 81.45 | 1 | 4469 | 135.4 | 29.01 | 1 | 4657 | 115.5 |
| YOU KNOW WHAT | 2 | 62 | 52.1 | 5 | 1141 | 34.5 | 8.74 | 3 | 970 | 24 |
| KNOW WHAT I | 4 | 49 | 41.2 | n/a* | 545 | 16.8 | 29.30 | 9 | 515 | 12.8 |
| YOU KNOW YOU | 8 | 21 | 17.6 | 3 | 1462 | 44.3 | 24.15 | 11 | 504 | 12.5 |
| DON'T KNOW WHAT | 9 | 17 | 14.3 | 10 | 871 | 26.4 | 7.73 | 5 | 844 | 20.9 |
| DO YOU KNOW | 10 | 16 | 13.4 | 6 | 1075 | 32.6 | 16.77 | 2 | 1037 | 25.7 |
| LIKE YOU KNOW | 12 | 14 | 11.75 | n/a* | 51 | 1.55 | 29.89 | n/a* | 319 | 7.7 |
| YOU KNOW AND | 13 | 12 | 10 | 11 | 863 | 25.9 | 29.81 | 10 | 506 | 12.5 |
| YEAH I KNOW | 13 | 12 | 10 | 13 | 720 | 21.6 | 21.83 | 6 | 838 | 20.7 |
| I DON'T KNOW WHAT | 13 | 12 | 10 | n/a* | 571 | 16.9 | 16.63 | 8 | 614 | 15.2 |

Table 5(b): Highest frequency KNOW group clusters in a by occurrence rank. (*not amongst the top 15 clusters)

Furthermore, confirming earlier findings, is the use of the question *do you know*. In SCO this is used not only less often than in MAC but also less often than in BNC/C. In fact, all the high-frequency *KNOW* group clusters that are found in SCO and also in either of the other two corpora are found with their normalised frequency lower in SCO. This also highlights that there is more overlap in use between the comparators (MAC and BNC/C) than between them and SCO. If this is true of clusters in general, it presumably provides strong support for the claim that SCO is different. In principle, one would expect BNC and MAC to be similar since they sample the same range of types of speakers (there is a time difference, but not a large one). There are more high frequency clusters in MAC that are also found in BNC/C. Though there is still a difference in their relative frequencies, both corpora record these clusters more often than does SCO.

The expressions *I know I* and *you know I* are relatively prominent in their use in both MAC[171] and BNC/C. They do not occur, by contrast, amongst the most-used expressions in SCO.

Table 5(b) slightly shifts the focus. 5(b) lists only those 3 to 5 word clusters that appear 10 times or more per 100,000 words in SCO and are also amongst the most frequent 15 *KNOW* group clusters in the other corpora. Table 5b uses the ranking within the top 15 clusters of all three corpora but only lists those clusters that appear in at least two of the corpora.

[171] Cf. chapter 5. YOU KNOW I in MAC:MED is usually followed by *mean; I KNOW I* is usually followed by *know*.

Table 5(b) shows that 9 out of the 15 most frequent clusters are common to all three corpora. It also shows that only *I don't know* and *you know and* can be found with the same proportional frequency in all three corpora. The divergence in frequency for the other clusters in SCO when compared to MAC and BNC/C suggests that these clusters have different patterns of use.

It has to be said that MAC and BNC are not the same. They differ in a number of factors (time of collection, material used for *casual conversation* etc.) and this impacts on how spoken language is used. The BNC/C is therefore not a perfect comparator. However, as keywords and keyphrases are compared it is shown that BNC/C is overall more in agreement with MAC than with SCO.

All in all, a comparison of the KNOW group clusters confirms my earlier research. There are areas of overlap as can be expected; certain expressions appear in both SCO and MAC (and BNC/C). At the same time there are phrases that strongly differ in their frequency of use – up to the point where their use is absolutely marginal even in the far more substantially sized MAC. As a result, we have seen that the phrases *like you know* and *you know what I mean* are clearly identifiable as Scouse idiosyncrasies. Conversely, the functional question *do you know* appears about only half as often in spoken Liverpool English as in the UK spoken English corpora. This is comparable to expressions that seem to relate

personal opinion rather strongly (*you know I*)[172] which are far less prominent in SCO then in either of the other corpora.

## 10.4 The *MEAN* Group

It might be expected that the *MEAN* group would be less relevant for the purposes of comparisons between corpora. The information gathered this far seems to indicate that *mean* is mainly a high-frequency word because of its use in the key-cluster *you know what I mean.* However, when *MEAN* group clusters are checked within SCO corpus, there is strong evidence that *mean* is used in SCO in a way that differs from its use in either MAC or BNC/C spoken corpora.

Schourup (1985) and in particular Brinton (2003) note the large variety of functions *I mean* has been described as having. This chapter adds to this discussion and aims to show how context dependent (nesting) the classification of *I mean* functions are. This would indirectly support Brinton's findings that look at *I mean*'s diachronic development:

> On the macro-level, this study suggests that the evolution of *I mean* is best understood as a process of grammaticalization. Beyond the fact that pragmatic markers are not major class items, what distinguishes the development of *I mean* as grammatizalization rather than lexicalization is the apparent regularity of the change. (Brinton 2003: 18)

[172] Discussed in detail in chapter 5.3

| SCO and MAC cluster | Freq. SCO | per 100k SCO | per 100k MAC | Freq. MAC | LL[173] |
|---|---|---|---|---|---|
| WHAT I MEAN | 55 | 45.8 | 16.5 | 517 | 43.83 |
| KNOW WHAT I MEAN | 46 | 38.3 | 8.4 | 265 | 67.00 |
| YOU KNOW WHAT I MEAN | 45 | 37.5 | 7.7 | 242 | 70.02 |
| I MEAN I | 9 | 7.5 | 67.3 | 2113 | 93.94 |
| I MEAN LIKE | 8 | 6.7 | 4 | 112 | 2.87 |
| I MEAN THEY | 7 | 5.8 | 17.8 | 560 | 11.27 |
| DO YOU MEAN | 6 | 5 | 12 | 378 | 5.27 |
| BUT I MEAN | 6 | 5 | 27.1 | 853 | 29.24 |
| WHAT YOU MEAN | 6 | 5 | 6.3 | 199 | 0.20 |
| I MEAN IT | 5 | 4.1 | 5 | 150 | 0.03 |
| I MEAN IT'S | 3 | 2.7 | 21 | 651 | n/a |
| | | % freq. higher | % freq. lower | | |

Table 6: Highest frequency clusters MEAN in SCO compared to MAC clusters

[173] It may have been noticed that the LL figure for *you know what I mean* in 10.4 is higher than in 10.3. This is due to the fact that the log-likelihood has been calculated on the basis of the total numbers recorded for each term (in this case *mean* instead of *know*) in the respective corpora. That the LL figures are still roughly similar underlines the strongly divergent frequency of use of this phrase.

### 10.4.1 SCO's most frequent MEAN group clusters compared to MAC

In this section, the list of the highest-occurring *MEAN* group clusters in SCO is taken and then their frequencies are compared to the same clusters in MAC. In this way, keyness of any given cluster in each corpus is highlighted.

Table 6 highlights that *mean* in SCO is almost exclusively used in conjunction with its collocate *"I" – I mean.* The only exceptions are *mean* when used as a literal question – *do you mean*[174], which is not used significantly different in SCO compared to MAC.

The first comparison shown in Table 6 confirms that the *MEAN* group in SCO is mostly in use as a constituent part of the phrase *you know what I mean* – and this is clearly much stronger in use in SCO than in MAC.

When we look at all the clusters that are used with higher frequency in MAC, a very clear divergence of use is visible. This is very significant in the case of the cluster *I mean I*, which, proportionally, appears nearly nine times as frequently in MAC as it appears in SCO.

The rhetorical question *you know what I mean*, postpositioned like a tag, with the apparent function to check whether the listener still follows the speaker, is the predominant use of *mean* in SCO.

[174] In SCO, DO YOU MEAN is always used as a question. WHAT YOU MEAN is usually within a phrase of confirmation as in I KNOW WHAT YOU MEAN (3 of the 6 uses).

To see how far *mean* is used in a different context, SCO clusters are compared to those in MAC and BNC/C in the next section.

### 10.4.2 The most frequent MEAN group clusters

In this section, the most frequent 3-5 word *MEAN* group clusters are compared in SCO and MAC and BNC/C. Again, this is being used to highlight Keyness of the phrases in the respective corpora.

Table 7 (a/b) is constructed on the same principles that were followed for Table 5(a/b) (the *KNOW* group) with the same colours being given to the same clusters, and the most frequent cluster of each corpus in the *MEAN* group at the top of the list.

TABLE 7(a): Highest frequency MEAN group clusters in SCO, MAC and BNC/C

| SCO | freq. | per 100k | MAC | freq. | per 100k | BNC/C | freq. | per 100k |
|---|---|---|---|---|---|---|---|---|
| WHAT I MEAN | 55 | **45.8** | I MEAN I | 2113 | **67.3** | I MEAN I | 1182 | **29.4** |
| KNOW WHAT I MEAN | 46 | **38.3** | I MEAN IT | 1206 | **38.4** | BUT I MEAN | 764 | **19** |
| YOU KNOW WHAT I MEAN | 45 | **37.5** | I MEAN YOU | 924 | **29.4** | WHAT I MEAN | 624 | **15.6** |
| YOU KNOW WHAT | 45 | **37.5** | BUT I MEAN | 853 | **27.1** | I MEAN YOU | 519 | **12.9** |
| I MEAN I | 9 | **7.5** | I MEAN THAT | 667 | **21.2** | I MEAN IT'S | 485 | **12** |
| I MEAN LIKE | 8 | **6.7** | I MEAN IT'S | 651 | **20.7** | KNOW WHAT I | 371 | **9.5** |
| I MEAN THEY | 7 | **5.8** | I MEAN THE | 640 | **20.4** | KNOW WHAT I MEAN | 369 | **9.4** |
| DO YOU MEAN | 6 | **5** | I MEAN WE | 618 | **19.7** | I MEAN IF | 362 | **9.4** |
| BUT I MEAN | 6 | **5** | I MEAN IF | 566 | **18** | I MEAN IT | 349 | **9.4** |
| WHAT YOU MEAN | 6 | **5** | I MEAN THEY | 560 | **17.8** | I MEAN THE | 322 | **8.1** |
| I MEAN IT | 5 | **4.1** | WHAT I MEAN | 517 | **16.5** | WELL I MEAN | 303 | **7** |
| I MEAN IT'S | 3 | **2.7** | YEAH I MEAN | 463 | **14.7** | KNOW I MEAN | 296 | **7** |
| | | | WELL I MEAN | 459 | **14.6** | I MEAN THEY | 286 | **6.9** |
| | | | I MEAN THERE | 426 | **13.6** | YOU KNOW WHAT | 282 | **6.9** |
| | | | DO YOU MEAN | 378 | **12** | DO YOU MEAN | 277 | **6.9** |
| | | | I MEAN HE | 376 | **12** | I MEAN HE | 274 | **6.9** |
| | | | I MEAN THAT'S | 352 | **11.2** | YOU KNOW WHAT I MEAN | 272 | **6.8** |

TABLE 7(a): Highest frequency MEAN group clusters in SCO, MAC, BNC/C

This overview of spoken *mean* use in Table 7(a) flags up that *mean* very strongly collocates with either *"I"* or *you.* It is a quality that appears in all three corpora. *Mean* has a strong tendency to collocate with *"I"* and *you* and this can be seen as a form of *nesting*.

Table 7(a) shows, in SCO, clusters incorporating *what I mean* produce findings similar to those in the KNOW group. We can see that in MAC and BNC/C, while total frequencies differ, many of the most frequent clusters that have appeared in the older BNC/C are still appearing in the more up-to-date MAC corpus. The contrast between SCO usage and the comparators is stronger still. The only cluster appearing within the five most frequent *MEAN* group clusters in all three corpora is *I mean I*. This becomes clearer when we look at Table 7(b).

TABLE 7 (b): Highest frequency MEAN group clusters in SCO, MAC and BNC/C ordered by cluster & rank.

| SCO | freq | per 100k | | MAC | freq | per 100k | | BNC/C | freq | per 100k |
|---|---|---|---|---|---|---|---|---|---|---|
| 1- WHAT I MEAN | 55 | **45.8** | | 11 - WHAT I MEAN | 517 | **16.5** | | 3 - WHAT I MEAN | 624 | **15.6** |
| 2- KNOW WHAT I MEAN | 46 | **38.3** | | -/- KNOW WHAT I MEAN | 265 | **8.4** | | 7 - KNOW WHAT I MEAN | 369 | **9.5** |
| 3- YOU KNOW WHAT I MEAN | 45 | **37.5** | | -/- YOU KNOW WHAT I MEAN | 242 | **7.7** | | 18 - YOU KNOW WHAT I MEAN | 272 | **6.9** |
| 4- YOU KNOW WHAT | 45 | **37.5** | | -/- YOU KNOW WHAT | 295 | **9.3** | | 14 - YOU KNOW WHAT | 282 | **6.9** |
| 5 - I MEAN I | 9 | **7.5** | | 1 - I MEAN I | 2113 | **67.3** | | 1 - I MEAN I | 1182 | **29.4** |
| 6- I MEAN THEY | 7 | **5.8** | | 10 - I MEAN THEY | 560 | **17.8** | | 13 - I MEAN THEY | 286 | **6.9** |
| 7 - DO YOU MEAN | 6 | **5** | | 15 - DO YOU MEAN | 378 | **12** | | 15 - DO YOU MEAN | 277 | **6.8** |
| 7 - BUT I MEAN | 6 | **5** | | 4 - BUT I MEAN | 853 | **27.1** | | 2 -BUT I MEAN | 764 | **19** |
| 8 - I MEAN IT | 5 | **4.1** | | 2 - I MEAN IT | 1206 | **38.4** | | 9 - I MEAN IT | 349 | **9.4** |
| 9 - I MEAN IT'S | 3 | **2.7** | | 6 - I MEAN IT'S | 651 | **20.7** | | 5 - I MEAN IT'S | 485 | **12** |
| | | | | 3 - I MEAN YOU | 924 | **29.4** | | 4 - I MEAN YOU | 519 | **12.9** |
| | | | | 7 - I MEAN THE | 640 | **20.4** | | 10 - I MEAN THE | 322 | **8.1** |
| | | | | 9 - I MEAN IF | 566 | **18** | | 8 - I MEAN IF | 362 | **9.4** |
| | | | | 13 - WELL I MEAN | 459 | **14.6** | | 11 -WELL I MEAN | 303 | **7** |
| | | | | 16 - I MEAN HE | 376 | **12** | | 16 – I MEAN HE | 274 | **6.9** |

TABLE 7 (b) : Highest frequency MEAN group clusters in SCO, MAC, BNC/C ordered by cluster & rank[2]

The *I mean I* cluster appears less than 10 times per 100.000 words in SCO (7.5 times) whereas it is the most frequent 3-5w cluster in both MAC (67.3 times) and BNC/C (29.4 times).

The most striking difference concerns, as seen in 10.4.1, the cluster *you know what I mean*. This is amongst the most frequent clusters in spoken Scouse (37.5 times within every 100.000 words). Furthermore, we find that shorter clusters are to a high degree constituent parts of this 5w cluster. In MAC and BNC/C the frequency for this cluster is roughly the same and comparatively low: 7.7 times /100.000 words in MAC and 6.9 times /100.000 words in BNC/C.

O'Donnell (2009) points out that adjusted frequency lists "highlight chunks of potential value" and, as explained in detail in chapter 4.3.4, the constituent parts of *You know what I mean* support his analysis[175].

Table 7(b) shows clearly that the proportional frequencies of the constituent parts are not only markedly lower but that these, shorter, clusters also appear to a stronger degree in other clusters than *You know what I mean.*

By contrast, the 3w cluster *I mean you* is frequent both in MAC (29.4 per 100k; ranked third most frequent) and BNC/C (12.9 per 100k; ranked fourth most frequent). This cluster, however, appears less than 3 times in 100k words in SCO - meaning it is below the threshold of what is counted.

[175] In fact, a separate study that I have undertaken looking at the occurrence pattern of *You know what I mean* in a larger number of spoken English corpora highlights the fact that SCO is unique in recording shorter constituent parts that almost exclusively appear in this particular phrase.

To sum up:

Within the *MEAN,* group very clear differences of use can be found. While *you know what I mean* is a cluster that is Key amongst speakers in SCO, the relatively frequent chunk *I mean you* in both MAC and BNC/C is barely recorded in SCO.

While all corpora have the common feature that *mean* mostly collocates with either *"I"* or *you*, and while a large number of clusters occur in both SCO and BNC/C amongst the four highest occurring, the actual clusters are very different. SCO has mainly *what I mean* and longer clusters that incorporate this 3w cluster. In MAC, strong use is made of *I mean I* and the other top clusters are not extensions but very different clusters (*I mean you* and *but I mean*) which together highlight a very different range of the uses of *mean*.

## 10.5 The *LIKE* group

*Like* has been extensively discussed earlier (in Chapter 10.2). *Like*, in the current discussion, is no longer seen as either *redundant or a downtoner* (cf. Miller & Weinert 1995: 386). There is anecdotal evidence that *like* is extensively used as a tag amongst speakers in Liverpool and the earlier chapter discusses the corpus evidence with regards to the use of the word in depth[176]. Apart from that, there is the functional use of *like* (i.e. *I like bananas*). When we look at the single word *like*, as we

[176] See chapter 7.2

have seen, its frequency is similar to the frequency of *know* and only slightly more frequent than *mean* in SCO. In MAC, however, *like* is slightly more frequent than *mean* and both words are less frequent than *know*. It will therefore be of interest to see in what way these frequencies are or are not paralleled when it comes to comparing *LIKE* group clusters. Comparing clusters will also flag up to what extend *like* is used as a tag or with its literal meaning.

### 10.5.1 Comparing the most frequent like group clusters in SCO and MAC

*LIKE* group cluster distribution shows that the most frequent clusters are far less frequent than the 3-5 word clusters in the *KNOW* and *MEAN* groups. This indicates that *like* goes together with a larger number of other words to form clusters, none of which is nearly as predominant in its use as, for example, *You know what I mean* is.

| Cluster SCO & MAC | Freq. : SCO | per 100.000 : SCO | per 100.000 : MAC | Freq. : MAC | LL |
|---|---|---|---|---|---|
| STUFF LIKE THAT | 16 | **13.3** | **2.3** | 81 | 26.31 |
| LIKE YOU KNOW | 16 | **13.3** | **7.1** | 243 | 4.58 |
| I WAS LIKE | 15 | **12.5** | **> 1 (0.7)** | 25 | 49.57 |
| IT WAS LIKE | 14 | **11.7** | **3.7** | 129 | 11.50 |
| THINGS LIKE THAT | 12 | **10** | **10.5** | 360 | 0.07 |
| I LIKE THE | 11 | **9.2** | **7** | 237 | *0.62* |
| SOMETHING LIKE THAT | 11 | **9.2** | **15.9** | 543 | 4.35 |
| I LIKE THAT | 10 | **8.3** | **5.4** | 185 | 1.38 |
| AND STUFF LIKE | 9 | **7.5** | **1.6** | 57 | 11.90 |
| I DON'T LIKE | 9 | **7.5** | **15.8** | 564 | 11.88 |
| ANYTHING LIKE THAT | 9 | **7.5** | **3.4** | 116 | 3.96 |
| A BIT LIKE | 9 | **7.5** | **2.8** | 97 | 5.70 |
| YOU KNOW LIKE | 9 | **7.5** | **6.8** | 232 | 0.04 |
| AND STUFF LIKE THAT | 8 | **6.7** | **1.5** | 55 | 7.84 |
| I LIKE TO | 7 | **5.8** | **9.2** | 312 | 1.81 |
| I MEAN LIKE | 7 | **5.8** | **3.1** | 107 | 1.96 |
| OR SOMETHING LIKE | 6 | **5** | **7.4** | 252 | 1.16 |
| OR SOMETHING LIKE THAT | 5 | **4.2** | **7.2** | 234 | 1.60 |
| | | *% freq. higher* | *% freq. lower* | | |

Table 8: Highest frequency clusters LIKE in SCO compared to MAC clusters

Table 8 demonstrates that, firstly, *LIKE* group clusters are far more frequent in SCO than MAC (mirroring what has been said about *like* in 10.2). We, secondly, see that *LIKE* group clusters in SCO are used in a way that is very different from the *LIKE* group clusters in MAC. The only cluster that is used with the same proportional frequency is THINGS *like* THAT. However, the four highest frequency clusters in SCO are mostly marginal in MAC – in particular I WAS *like* (0.7 uses / 100k words in MAC; 12.5 uses / 100k words in SCO) and STUFF *like* THAT (2.3 uses / 100k words in MAC; 13.3 uses /100k words in SCO)[177].

By comparison, the high frequency clusters in MAC (SOMETHING *like* THAT and I DON'T *like*) are also quite frequent in SCO.[178]

Table 8 also shows that the occurrence of *like* as a function word, though frequent, does not appear to be its predominant use. Table 8 shows the 17 most frequent *LIKE* group clusters in SCO and the use of *like* appears to have three main functions. It is a discourse particle (*I mean like)* and is used as a term indicating preference three times[179]. It appears to be used to compare something, however, nine times:

[177] As the LL values indicate, apart from *like you know* this divergence is significant at a high level.

[178] It is left open to interpretation why the average UK speaker should feel twice (15.8 times in 100k words) as inclined as SCO speakers (7.5 times in 100k words) to express the negative I DON'T LIKE. Interestingly, I LIKE TO is uttered also nearly twice as often (9.2 times in 100 k words) by speakers in MAC compared to SCO (5.8 times) – this will be discussed in detail in 10.7.

[179] This is a very straightforward pattern: I LIKE + determiner indicates a noun phrase, a thing is liked; I LIKE+ TO infinitive introduces a clause – an action that is liked. This has been discussed in detail in 9.2

| LIKE to indicate *preference* | LIKE to *compare* |
|---|---|
| I LIKE THE | STUFF LIKE THAT |
| I DON'T LIKE | IT WAS LIKE |
| I LIKE TO | THINGS LIKE THAT |
| | SOMETHING LIKE THAT |
| | AND STUFF LIKE |
| | ANYTHING LIKE THAT |
| | AND STUFF LIKE THAT |
| | A BIT LIKE |
| | OR SOMETHING LIKE |

All of the latter have in common that in the cluster a point of comparison is broader: *stuff; that; something; anything* and *a bit*. *Like* is therefore used to compare one thing (or a list of things) with another, unnamed set. There remain uses of *like* that are tag-like discourse markers: *like* YOU KNOW, YOU KNOW *like* [180] & I MEAN *like*.

### 10.5.2 The most frequent LIKE group clusters

When the frequency of occurrence pattern of LIKE is compared between SCO and both MAC and BNC/C it is clearly shown that *LIKE* clusters are used in a different way by SCO speakers. Table 9(a) lists the highest frequency *like* clusters in order of frequency (rank). As before, the same clusters are in fields shaded with the same colour. This is intended to highlight the fact that SCO *like* clusters can be found not just with different frequencies but also with different rankings of frequency from those of MAC and BNC/C.

[180] How one word order is found in strong preference to the other when SCO is compared to MAC has been discussed in chapter 9.2.

Table 9(a): Highest occurring *LIKE* group clusters in SCO,MAC and BNC/C

| Cluster SCO | Freq. | per 100k | Cluster MAC | Freq. | per 100k | Cluster BNC/C | Freq. | per 100k |
|---|---|---|---|---|---|---|---|---|
| 1- STUFF LIKE THAT | 16 | 13.3 | 1 - WOULD LIKE TO | 574 | 16.8 | 1 - I DON'T LIKE | 597 | 14.8 |
| 2 - LIKE YOU KNOW | 16 | 13.3 | 2 - SOMETHING LIKE THAT | 543 | 15.9 | 2 - SOMETHING LIKE THAT | 458 | 11.4 |
| 3 - I WAS LIKE | 15 | 12.5 | 3 - I DON 'T LIKE | 540 | 15.8 | 3 - LIKE YOU KNOW | 319 | 7.9 |
| 4 - IT WAS LIKE | 14 | 11.7 | 4 - I'D LIKE TO | 495 | 14.5 | 4 - LIKE THAT AND | 279 | 6.9 |
| 5 - THINGS LIKE THAT | 12 | 10 | 5- I WOULD LIKE | 431 | 12.6 | 5 - WOULD YOU LIKE | 276 | 6.9 |
| 6 - I LIKE THE | 11 | 9.2 | 6 - THINGS LIKE THAT | 360 | 10.6 | 6 - THINGS LIKE THAT | 262 | 6.5 |
| 7 - SOMETHING LIKE THAT | 11 | 9.2 | 7 WOULD YOU LIKE | 321 | 9.4 | 7 - YOU KNOW LIKE | 234 | 5.8 |
| 8 - I LIKE THAT | 10 | 8.3 | 8 - LIKE THAT AND | 314 | 9.2 | 8 - SORT OF LIKE | 221 | 5.5 |
| 9 - AND STUFF LIKE | 9 | 7.5 | 9 - I WOULD LIKE TO | 312 | 9.2 | 9 - I'D LIKE TO | 207 | 5 |
| 10 - I DON'T LIKE | 9 | 7.5 | 10 - LIKE THAT I | 255 | 7.5 | 10 - LIKE THAT I | 206 | 5 |
| 11 - ANYTHING LIKE THAT | 9 | 7.5 | 11-OR SOMETHING LIKE | 252 | 7.4 | 11 - IT LIKE THAT | 191 | 4.7 |
| 12 - A BIT LIKE | 9 | 7.5 | 12- LIKE YOU KNOW | 243 | 7.1 | 12 - I LIKE THAT | 185 | 4.5 |
| 13 - YOU KNOW LIKE | 9 | 7.5 | 13 - I LIKE THE | 237 | 7 | 13 - OR SOMETHING LIKE | 183 | 4.5 |
| 14 - AND STUFF LIKE THAT | 8 | 6.7 | 14 - IF YOU LIKE | 235 | 6.9 | 14 - DO YOU LIKE | 178 | 4.4 |
| 15 - I LIKE TO | 7 | 5.8 | 15 - LIKE TO SEE | 234 | 6.9 | 15 - DON'T LIKE IT | 171 | 4.3 |
| 16 I MEAN LIKE | 7 | 5.8 | 16 OR SOMETHING LIKE THAT | 234 | 6.9 | 16 - OR SOMETHING LIKE THAT | 169 | 4.2 |
| 17 - OR SOMETHING LIKE | 6 | 5 | 17 - YOU KNOW LIKE | 232 | 6.8 | 17 - LIKE THAT YEAH | 164 | 4.1 |
| 18 - WAS LIKE THAT | 6 | 5 | 18 - AND THINGS LIKE | 202 | 6 | 18 - IT WAS LIKE | 164 | 4.1 |
| 19 - IT'S LIKE A | 6 | 5 | 19 - DIDN' T LIKE | 192 | 5.6 | 19 - I LIKE THE | 163 | 4.1 |
| 20- LIKE THAT AND | 6 | 5 | 20 - IT LOOKS LIKE | 188 | 5.5 | 20 - IF YOU LIKE | 157 | > 4 |
| 21 - LIKE THAT I | 6 | 5 | 21 - YOU LIKE TO | 186 | 5.5 | Table 9(a): Highest occurring *LIKE* group clusters in SCO,MAC and BNC/C | | |
| 22- DON'T LIKE IT | 6 | 5 | 22 - I LIKE THAT | 185 | 5.4 | | | |

As shown earlier, *things like that* is the one exception: though lower in frequency in BNC/C (6.5 times per 100k) than MAC or SCO (10.0 and 10.6 per 100k respectively) it ranks as the 6th most frequent 3-5w *like* cluster in BNC/C and MAC and the 5th most frequent in SCO.

Table 9(a) gives the 22 most frequent *LIKE* group clusters of all three corpora. Within those 22, SCO has 12 clusters that are not in the 22 most frequent *like* clusters of either MAC or BNC/C. Amongst these, as we saw in chapter 10.2, the combination of *like* with *stuff* is a collocation that has a high level of preference in SCO but in neither of the other two clusters. Table 9(a) also flags up the fact that the most frequent clusters in MAC and BNC/C are similar, but that *things like that* is the only high frequency cluster also found in SCO. *Something like that* and *I don't like that* are, by contrast, the most used *like* clusters in MAC and BNC/C only. It is worthy of note that these two clusters serve two very different purposes: *something like that* is a vague descriptor; *I don't like that* is a definite statement of emotion. As described in chapter 9.2, SCO speakers, though they also use *things like that* and *something like that*, seem to prefer the phrase *stuff like that* for the same purpose. Furthermore, SCO makes use of another vague descriptor: *anything like that*. This is not frequent in the other spoken corpora. SCO speakers also appear to use *like* with *know* more often to check listeners' understanding.

Table 9(b) gives a direct comparison of the ranking of the same clusters in the three corpora. The rank refers to the 22 most frequent

*LIKE* group clusters that are shared by two or more corpora. Clusters that only appear in a single corpus are not listed on this occasion.

Table 9 (b): Highest occurring *LIKE* group clusters in SCO, MAC and BNC/C in direct comparison

| Cluster SCO | Freq. | per 100k | Cluster MAC | Freq. | per 100 k | Cluster BNC/C | Freq. | per 100k |
|---|---|---|---|---|---|---|---|---|
| 2 - LIKE YOU KNOW | 16 | 13.3 | 12- LIKE YOU KNOW | 243 | 7.1 | 3 - LIKE YOU KNOW | 319 | 7.9 |
| 4 - IT WAS LIKE | 14 | 11.7 | -/- IT WAS LIKE | 129 | 3.8 | 18 - IT WAS LIKE | 164 | 4.1 |
| 5 - THINGS LIKE THAT | 12 | 10 | 6 - THINGS LIKE THAT | 360 | 10.6 | 6 - THINGS LIKE THAT | 262 | 6.5 |
| 6 - I LIKE THE | 11 | 9.2 | 13 - I LIKE THE | 237 | 7 | 19 - I LIKE THE | 163 | 4.1 |
| 7 - SOMETHING LIKE THAT | 11 | 9.2 | 2 - SOMETHING LIKE THAT | 543 | 15.9 | 2 - SOMETHING LIKE THAT | 458 | 11.4 |
| 8 - I LIKE THAT | 10 | 8.3 | 22 - I LIKE THAT | 185 | 5.4 | 12 - I LIKE THAT | 185 | 4.5 |
| 10 - I DON'T LIKE | 9 | 7.5 | 3 - I DON 'T LIKE | 540 | 15.8 | 1 - I DON'T LIKE | 597 | 14.8 |
| 13 - YOU KNOW LIKE | 9 | 7.5 | 17 - YOU KNOW LIKE | 232 | 6.8 | 7 - YOU KNOW LIKE | 234 | 5.8 |
| 17 - OR SOMETHING LIKE | 6 | 5 | 11-OR SOMETHING LIKE | 252 | 7.4 | 13 - OR SOMETHING LIKE | 183 | 4.5 |
| /-OR SOMETHING LIKE THAT | 5 | 4.1 | 16 OR SOMETHING LIKE THAT | 234 | 6.9 | 16 - OR SOMETHING LIKE THAT | 169 | 4.2 |
| 20- LIKE THAT AND | 6 | 5 | 8 - LIKE THAT AND | 314 | 9.2 | 4 - LIKE THAT AND | 279 | 6.9 |
| 21 - LIKE THAT I | 6 | 5 | 10 - LIKE THAT I | 255 | 7.5 | 10 - LIKE THAT I | 206 | 5 |
| 22- DON'T LIKE IT | 6 | 5 | -/- DON'T LIKE IT | 134 | 3.9 | 15 - DON'T LIKE IT | 171 | 4.3 |
| | | | 4 - I'D LIKE TO | 495 | 14.5 | 9 - I'D LIKE TO | 207 | 5 |
| | | | 7 WOULD YOU LIKE | 321 | 9.4 | 5 - WOULD YOU LIKE | 276 | 6.9 |
| | | | 14 - IF YOU LIKE | 235 | 6.9 | 20 - IF YOU LIKE | 157 | > 4 |

Table 9 (b): Highest occurring *LIKE* group clusters in SCO,MAC and BNC/C in direct comparison

There are further instances of *LIKE* group clusters where SCO differs both in frequency and rank when compared to MAC and BNC/C. This is well exemplified by the (2nd ranked) SCO cluster *like you know*. It is third ranked in BNC/C and 12th ranked in MAC. The frequencies per 100.000 words are close for BNC/C (7.9) and MAC (7.1) making them proportionally less than two thirds as frequent as in SCO (13.3 times per 100.000 words).

The reverse is true for the cluster *I don't like* . This is ranked 1st in BNC/C and 3rd in MAC. It is frequent in both: 14.8 times /100k words in BNC/C and 15.8 times /100k. In SCO, however, it is ranked 10th most frequent *LIKE* group cluster and is used only 7.5 times per 100.000 words.

To conclude: further to what we have described in chapter 10.2, *like* and the *LIKE* group of clusters highlight a different pattern of occurrence in SCO when compared with both MAC and BNC/C.

While *things like that* is one cluster that ranks in occurrence near-equal in all three clusters, and has about the same frequency of use in SCO and MAC, it stands out as the exception.

Table 9(a) shows that there are a large number of clusters that are highly frequent in SCO while appearing to be proportionally less frequent in MAC and BNC/C. The differences are shown to be even stronger when SCO and MAC clusters are directly compared. Table 8 clearly shows that the four most frequent *LIKE* group clusters in SCO - *stuff like that*, *like*

*you know*, *I was like* , *it was like* - stand out because of their high frequency: 13.3 – 11.7 times per 100.000 words. In MAC the same clusters stand out because of their low frequency: between 0.7 and 7.1 occurrences per 100.000 words. In other words, the *LIKE* group clusters that are in prominent use in SCO are marginal in MAC. At the same time, however, the reverse cannot be found to be true. Relatively frequent MAC clusters like *something like that* are also used frequently amongst SCO speakers. The only exception to this is the preference in MAC for the *would like* construction – this appears to be little used in the BNC/C and does not at all surface in any SCO concordance lines.

## 10.6 Cluster comparison with an extended MAC corpus

Through circumstances beyond my control, the corpus of comparison, the 3.3 million word strong Macmillan casual spoken corpus, became no longer available to me during the course of my research because of changes in the contractual relationship between Macmillan Publishers and Bloomsbury (who originated the MAC corpus). Instead, for the final set of comparisons below, I had to use the 2008 version of Macmillan English Dictionary corpus. While the corpus's composition is similar, this spoken corpus is considerably larger: a total of 8.336.253 words. Consequently, the frequencies of occurrence are considerably higher than those encountered in all previous sections. This means that Macmillan

English Dictionary (MAC:MED) is nearly twice as large as the BNC/C and nearly as large as the BoE UKSpoken corpus.

While a larger size allows a more precise overview of how regular certain words and clusters occur, a larger corpus also brings a larger variety of results which need to be included and investigated. At all events, having the latest version of the corpus available has meant that for these last comparisons the most up-to-date data are being used.

Table 10(a): Highest Frequency *THINK* group clusters ranked by frequency separately in SCO, MAC and BNC/C

| SCO | Freq | per 100 k | MAC:MED | Freq | per 100 k | BNC/C | Freq | per 100 k |
|---|---|---|---|---|---|---|---|---|
| 1- I DON'T THINK | 34 | 28.3 | 1- I THINK THAT | 3029 | 36.3 | 1- I DON'T THINK | 1744 | 43.4 |
| I THINK IT | 17 | 14.2 | 2- I THINK IT | 2939 | 35.3 | 2- DO YOU THINK | 742 | 18.4 |
| THINK IT WAS | 13 | 10.8 | 3- I DON T THINK | 2798 | 33.6 | 3- I THINK I | 720 | 17.9 |
| I THINK IT'S | 12 | 10 | 4- AND I THINK | 2185 | 26.2 | 4 - I THINK IT'S | 720 | 17.9 |
| I THINK IT WAS | 10 | 8.3 | 5- I THINK I | 1924 | 23.1 | 5 - I THINK IT | 610 | 14.1 |
| DO YOU THINK | 8 | 6.7 | 6- I THINK WE | 1851 | 22.2 | 6 THINK IT WAS | 357 | 8.9 |
| AND I THINK | 7 | 5.8 | 7- I THINK IT'S | 1610 | 19.3 | 7 - WELL I THINK | 345 | 8.5 |
| THINK ABOUT IT | 7 | 5.8 | 8- I THINK THE | 1419 | 17.0 | 8 -I THINK YOU | 345 | 8.5 |
| I THINK THAT'S | 7 | 5.8 | 9- BUT I THINK | 1263 | 15.2 | 9 -I THINK THAT'S | 312 | 7.8 |
| NO I DON'T THINK | 6 | 5.0 | 10- DO YOU THINK | 1169 | 14.0 | 10 - BUT I THINK | 296 | 7.3 |
| I THINK THE | 6 | 5.0 | 11- I THINK YOU | 1147 | 13.9 | 11- I SHOULD THINK | 292 | 7.3 |
| DON'T THINK SO | 5 | 4.1 | 12- I THINK THAT'S | 999 | 12.0 | 12- AND I THINK | 284 | 7.1 |
| DON'T THINK I | 5 | 4.1 | 13 - WELL I THINK | 815 | 9.9 | ***13- I THINK HE*** | 274 | 7 |
| DON'T THINK IT | 5 | 4.1 | ***14- I THINK HE*** | 727 | 8.7 | 14 - DON'T THINK SO | 268 | 6.7 |
| I THINK THIS | 5 | 4.1 | 15 - SO I THINK | 721 | 8.7 | 15 - I THINK THAT | 260 | 6.6 |
| I THINK THEY | 5 | 4.1 | 16 -THAT I THINK | 707 | 8.5 | 16- I THINK SO | 258 | 6.4 |
| I THINK YOU | 5 | 4.1 | 17 - THINK I THINK | 685 | 8.2 | 17 -THINK I THINK | 252 | 6.3 |
| | | | 18 - I THINK THEY | 680 | 8.2 | 18 - I THINK IT WAS | 243 | 6.1 |
| | | | 19 - I THINK THERE | 623 | 7.5 | 19 - I THINK THEY | 233 | 5.9 |
| | | | 20- THINK IT WAS | 453 | 5.4 | 20 - I THINK THE | 223 | 5.5 |
| | | | 21- DON'T THINK IT | 429 | 5.1 | 21 -YEAH I THINK | 209 | 5.3 |
| | | | 22- YEAH I THINK | 413 | 5.0 | I DON'T THINK SO | 206 | 5.3 |
| | | | DON T THINK SO | 334 | 4.0 | THINK ABOUT IT | 166 | 4.1 |
| | | | | | | ***I THINK IT WAS*** | ***326*** | ***3.9*** |

Table 10(a) Highest Frequency *THINK* group clusters ranked by frequency separately in SCO, MAC and BNC/C



**10**

## .7 The *THINK* Group

THINK is one of the most frequent words in spoken language. *Think* as a collocate of *I* and *think* with *discourse particles* has been been discussed in-depth earlier [181]. This section concentrates on the most frequent *THINK* group clusters that can be found in SCO and the two corpora of comparison. This comparison highlights how THINK appears in clusters with markers of negation, with connectors (like *but, and*), both the first and second persons singular, the third person plural, and markers of referral (like *it, so*). Table 10a compares the most frequently occurring 3w *THINK* group clusters ranked by frequency of each corpus.

When comparing the *THINK* group clusters of SCO, MAC:MED and BNC/C, several areas of divergence open up. Amongst the 15 highest-frequency clusters of the *THINK* group, the majority of clusters appear in all three corpora. Furthermore, the most frequently occurring clusters incorporate *I think*. SCO and BNC/C have as the most frequently occurring *I don't think* (it is third-most frequent in MAC:MED). When comparing SCO with BNC/C, certain clusters have a broadly similar proportional frequency of use (though not ranking): *I think it; I think it was; and I think; I think that's; don't think so* etc. This indicates that, at least in the direct comparison of these two corpora, there are relatively large numbers of clusters in the *THINK* group that are used with similar frequencies.

[181] See chapter 4.

In Table 10(a) we also find that clusters that diverge strongly in their rank from MAC:MED and BNC/C. These include *I think it was* (ranked 5th in SCO, ranked below 20 in MAC:MED and 18th in BNC/C) and *don't think I* or *I think this* (SCO: ranked 9; in BNC/C and MAC:MED ranked below 20)

In SCO, *don't think I* and *I think this* occur just over four times in 100k words; in MAC:MED they occur just under four times in 100k words, while BNC/C has a low 1.7 occurrences per 100k words for *I think this.* By contrast, MAC:MED has a large number of the cluster *so I think* (8.7 occ in MAC:MED, 3.5 times in BNC/C per 100k words) and BNC/C records the similar cluster *I think so* (6.4 times in BNC/C, 4.3 times per 100k words in MAC:MED). While *I think so* appears 3 times in SCO (= 2.5 times per 100k words) *so I think* is not recorded, making both clusters far more marginal in SCO than in either MAC: MED or BNC/C.

This highlights another point: both the clusters themselves and their proportional frequencies are very close in most cases where MED:MAC is compared with BNC/C. In the majority of cases, however, *THINK* group clusters are used with a proportionally different frequency in SCO when compared with the other two corpora.

Table 10(b): Highest frequency THINK group clusters in  with occurrence rank.

| SCO | Freq | per 100 k | MAC:MED | Freq | per 100 k | BNC/C | Freq | per 100 k |
|---|---|---|---|---|---|---|---|---|
| 1- I DON'T THINK | 34 | 28.3 | 3- I DON T THINK | 2798 | 33.6 | 1- I DON'T THINK | 1744 | 43.4 |
| 2 -I THINK IT | 17 | 14.2 | 2- I THINK IT | 2939 | 35.3 | 5- I THINK IT | 610 | 14.1 |
| 3 -THINK IT WAS | 13 | 10.8 | THINK IT WAS | 453 | 5.4 | THINK IT WAS | 357 | 8.9 |
| 4 - I THINK IT'S | 12 | 10 | 7- I THINK IT'S | 1610 | 19.3 | 3- I THINK IT'S | 720 | 17.9 |
| 5 - I THINK IT WAS | 10 | 8.3 | I THINK IT WAS | 326 | 3.9 | I THINK IT WAS | 243 | 6.1 |
| 6 - DO YOU THINK | 8 | 6.7 | 10 - DO YOU THINK | 1169 | 14 | 2- DO YOU THINK | 742 | 18.4 |
| 7 - AND I THINK | 7 | 5.8 | 4 AND I THINK | 2185 | 26.2 | AND I THINK | 284 | 7.1 |
| 7 - THINK ABOUT IT | 7 | 5.8 | THINK ABOUT IT | 288 | 3.4 | THINK ABOUT IT | 166 | 4.1 |
| 7 - I THINK THAT'S | 7 | 5.8 | I THINK THAT'S | 999 | 12.0 | I THINK THAT'S | 312 | 7.8 |
| 8 - NO I DON'T THINK | 6 | 5 | NO I DON'T THINK | 217 | 2.6 | NO I DON'T THINK | 64 | 1.6 |
| 8 - I THINK THE | 6 | 5 | 8 I THINK THE | 1419 | 17 | I THINK THE | 223 | 5.5 |
| 9 - DON'T THINK SO | 5 | 4.1 | -DON'T THINK SO | 334 | 4 | DON'T THINK SO | 268 | 6.7 |
| DON'T THINK I | 5 | 4.1 | DON'T THINK I | 290 | 3.4 | DON'T THINK I | 183 | 4.5 |
| I DON'T THINK I | 5 | 4.1 | I DON'T THINK I | 282 | 3.3 | I DON'T THINK I | 149 | 3.7 |
| I THINK THIS | 5 | 4.1 | I THINK THIS | 309 | 3.7 | I THINK THIS | 69 | 1.7 |
| I THINK THEY | 5 | 4.1 | I THINK THEY | 680 | 8.2 | I THINK THEY | 233 | 5.9 |
| I THINK YOU | 5 | 4.1 | 11- I THINK YOU | 1147 | 13.9 | I THINK YOU | 345 | 8.5 |
| DON'T THINK IT | 5 | 4.1 | 22-DON'T THINK IT | 429 | 5.1 | DON'T THINK IT | 139 | 3.5 |
| 9 - I DON'T THINK IT | 5 | 4.1 | I DON'T THINK IT | 267 | 3.2 | I DON'T THINK IT | 102 | 2.5 |

Table 10(b) Highest frequency THINK group clusters in  with occurrence rank. (SCO frequency ranking 1-9 as lead)

### 10.7.1 Clusters using DON'T THINK negation

In SCO, *THINK* group clusters incorporating *don't think* is fractionally more used than in the other corpora: 4.1 times in 100.000 words compared to 3.2 times per 100.00 words in MED:MAC, and 2.5 times in BNC/C, but it mainly stands out because it is amongst the 10 most frequent 3w / 4w clusters in the *THINK* group in SCO, while it is not even within the top 20 ranked clusters in MED:MAC or BNC/C.

The other exception in SCO is *no I don't think.* This appears 5 times in every 100.000 words of the whole corpus – twice as often as in the MED:MAC (2.6 times per 100.000 words) and over three times as often as in the BNC/C (1.6 times per 100.000 words). It is also the 6th most frequent 3w / 4w cluster with *think* in SCO. No other cluster, however, distinguishes the negative use of SCO from that of BNC/C and MAC:MED in the *THINK* group.

### 10.7.2 SCO distinctive use within the THINK group

This section only looks at those *THINK* group clusters where the occurrence patterns of utterances in SCO are noticeably different from the equivalent occurrence patterns in both MED:MAC and BNC/C. Although the clusters are not very different in SCO compared to the other corpora in terms of their relative ranking, they are, as Table 10(c) shows, throughout proportionally less frequent in their use:

| SCO | Freq | per 100 k | MAC:MED | Freq | per 100 k | LL |
|---|---|---|---|---|---|---|
| 2 -I THINK IT | 17 | 14.2 | 2- I THINK IT | 2939 | 35.3 | 19.02 |
| 7 - AND I THINK | 7 | 5.8 | 4 - AND I THINK | 2185 | 26.2 | 27.23 |
| 8 - I THINK THE | 6 | 5.0 | 8 - I THINK THE | 1419 | 17.0 | 13.79 |
| 9 - I THINK YOU | 5 | 4.1 | 11- I THINK YOU | 1147 | 13.9 | 10.79 |
| 4 - I THINK IT'S | 12 | 10.0 | 7- I THINK IT'S | 1610 | 19.3 | 6.31 |
| 6 - DO YOU THINK | 8 | 6.7 | 10 -DO YOU THINK | 1169 | 14 | 5.56 |

Table 10(c): SCO THINK group occurrence patterns different.

Seven of the 16 most frequent clusters that are common to all three corpora have the negation marker *don't* with *think*. This includes for all the three corpora the cluster *I don't think* as one of their most frequently used clusters.

The largest discrepancies can be found, however, in the non-negated forms. *I think it* is ranked number 2 in both corpora, yet occurs significantly less often, proportionally, in SCO. The divergence is strongest in the use of the cluster *and I think* which is ranked lower in SCO and is, significantly, more than three times as proportionally frequent in MAC.

*I think the* is an interesting case - though appearing in the same rank (8th highest *think* 3w cluster), it occurs only 5 times in SCO as opposed to 17 times in MAC:MED (5.5 times in BNC/C) in a 100.000 words. This

cluster is significantly (above the 99.9% mark) less often used, proportionally, in SCO than in MAC.

This pattern is also seen with the cluster *I think you*. Though close in their rankings, (9th in SCO, 11th in MAC:MED) and diverging only to a level that is significant just above 99.0%, it occurs only 4.1 times in SCO as opposed to 13.9 times in MAC:MED (8.5 times in BNC/C) in 100.000 words. In other words, it occurs proportionally more than three times as often in MAC:MED, and more than twice as often in BNC/C.

To summarise: the majority of 3w / 4w clusters involving *think* appear in all three clusters.

The highest occurring cluster, *I don't think*, is a rare example where the proportional frequency of use is fully in line between all three corpora (see Table 10a). Amongst the 18 listed clusters, on the whole MAC:MED and BNC/C show, overall, greater agreement with each other than with SCO. Nevertheless, the *THINK* group clusters provide a number of examples that demonstrate significant divergence between SCO and MAC. In fact, the two most significantly divergent *THINK* group clusters in SCO (*I think it & and I think*) occur together as often as the highest occurring 3w *think* cluster.

### 10.7.3 THOUGHT occurrence patterns

When looking at the simple past tense use of the verb in the *THINK* group clusters, the differences are far less pronounced.

| SCO | Freq | per 100k | MAC:MED | Freq | per 100 k | LL |
|---|---|---|---|---|---|---|
| 1- I JUST THOUGHT | 6 | 5.0 | 12 -I JUST THOUGHT | 69 | 0.82 | 0.37 |
| 2- I THOUGHT I | 5 | 4.1 | 1- I THOUGHT I | 534 | 6.4 | 0.56 |
| 3- BUT I THOUGHT | 5 | 4.1 | 11- BUT I THOUGHT | 131 | 1.6 | 0.31 |
| 4 - THOUGHT IT WAS | 3 | 2.5 | 3- THOUGHT IT WAS | 446 | 5.4 | |
| 4 - AND I THOUGHT | 3 | 2.5 | 4 - AND I THOUGHT | 418 | 5.0 | |
| 4 - I THOUGHT IT | 3 | 2.5 | 2- I THOUGHT IT | 481 | 5.8 | |
| 4 - I THOUGHT THAT | 3 | 2.5 | 7 -I THOUGHT THAT | 272 | 3.3 | |

Table 11: Direct comparison of *THOUGHT* group 3w cluster occurrence frequencies

A direct comparison of the *THOUGHT* group clusters has to be necessarily limited by the extremely low numbers found in SCO. Where it is possible, statistical testing shows that there might be differences in proportional frequencies but none of these are significant.

This demonstrates that, though there are *THINK* group clusters that markedly diverge from their use where SCO is compared with MAC, this difference does not exist amongst *THOUGHT* group clusters.

## 10.8 The *TO* Group

The word *TO* is one of the most frequent elements of spoken and written English. It freely connects with a large number of collocates. Yet, surprisingly, when looking at the work by corpus linguists, it seems to be

little discussed. Going through Biber, Conrad & Reppen (1998); Hoey (2004) Partington (2003) and Stubbs (1996) there is no special mention of *TO*[182]. Hunston & Francis (1999) show, however, in their work on *Pattern Grammar* how central *to* is for verb phrases. Indeed, when looking at *to* in corpus linguistic works for learners, Biber, Conrad & Leech (2002) and O'Keefe, McCarthy and Carter (2007), give space to the discussion of this function word.

*To* can be seen as crucial to our understanding of language and is therefore also crucial to corpus linguistics, as Hunston & Francis make clear when they look at verbs in their *Pattern Grammar*. They note that *to* is clearly problematic to categorise:

> Our description of verb patterns aims to be complete (…) This comprehensive approach has thrown up a number of problems with relation to traditional views of structure, problems which led us ultimately to conclude that traditional structural descriptions of English were neither necessary not sufficient to account for actual language use.
>
> (Hunston & Francis 1999: 160)

The majority of verbs identified by Hunston & Francis were verbs in combination with prepositional phrases and the patterns discovered using this kind of focus show the crucial role words like *to*, *from,* etc. can play. In their discussion, Hunston and Francis point out that *to* can be found as *to Infinitive* or as *to* with *Prepositional Object*. The *to* can be an *Infinitive marker* as well as a preposition (cf. Biber *et al.* 2002: 34).

[182] Sinclair (1991) highlights that seeing *of* simply as a preposition is misleading. He could have made the same claim for *to*.

There is a link to earlier definitions that concentrate on grammatical functions of *to*. One of the standard pre-corpus-linguistics works on grammar, Leech & Svartvik's *Communicative Grammar of English* demonstrates simply through the number of index entries the variety of environments where *to* can be found:

> *To*, PREPOSITION; ADJECTIVE COMPLEMENT; INDIRECT OBJECT; place; preferences; time; CONJUNCTION; *to*-INFINITIVE. *To be sure* (see sentence adverbial) (Leech & Svartvik 1975: 322f.)

This is quite a number of functions for a short word.

All of this appears to demonstrate that *to*, though short and hardly to be confused with other words, is a word that has a function that is hard to define. Looking at patterns of occurrence to see how *to* collocates and colligates therefore seems to be a justified task. Looking at the cluster lists of Spoken English (UK) provided by O'Keefe / McCarthy and Carter (2007: 65ff.) we find that *in* and *of* appear within the top 7 most frequent 2w clusters, and *to* appears within the top 20. However, within the 3w – 5w clusters, *to*, though not as frequently occurring as *of*, appears much higher, namely within the top 8 highest occurring clusters. (See Appendix IX.5 for a detailed breakdown). This gives some indication that *to*, like *of*, has a role to play in longer, stable chunks of spoken English. By comparison, in the same lists, the chunks including the highest-frequency term "I" can be seen to be ranked as becoming less frequent the longer the chunks looked at are.

This information – *to* being hard to classify, while nevertheless being an elementary part of longer clusters of spoken English – provides the background for the comparison of the occurrence pattern of *to* in SCO and a number of other Spoken English corpora.

Table 12(a) represents the 34 highest occurring 3w clusters in the *TO* group cluster. It shows clearly that the different corpora largely have separate high frequency clusters. Table 12(a) highlights that *to* chunks in a large variety of ways and each single corpus appears to show a large number of clusters that seem not to be found in any number in other corpora.

Table 12(a): Highest frequency TO group clusters SCO, MAC & BNC/C

| SCO | frq. | p100k | MAC:MED | frq. | p100k | BNCC | frq. | per 100k |
|---|---|---|---|---|---|---|---|---|
| 1 - YOU HAVE TO | 37 | 30.8 | 1 -BE ABLE TO | 3934 | 47.2 | 1- TO GO TO | 1266 | 31.5 |
| 2 - USED TO BE | 28 | 23.3 | 2 - GOING TO BE | 3194 | 38.3 | 2 - YOU WANT TO | 982 | 24.4 |
| 3 - TO GO TO | 27 | 22.5 | 3 - TO GO TO | 2806 | 33.6 | 3- YOU HAVE TO | 963 | 24.0 |
| 4 - I HAVE TO | 21 | 17.5 | 4 - TO BE ABLE | 2516 | 30.2 | 4 - BE ABLE TO | 904 | 22.5 |
| 5 - GO TO THE | 19 | 15.8 | 5 - YOU WANT TO | 2376 | 28.6 | 5- YOU'VE GOT TO | 798 | 19.8 |
| WE USED TO | 19 | 15.8 | 6- TO TRY TO | 2122 | 25.5 | TO DO IT | 754 | 18.7 |
| I USED TO | 19 | 15.8 | 7 - YOU HAVE TO | 1772 | 21.3 | TO HAVE A | 729 | 18.1 |
| I WENT TO | 18 | 15 | 8 - I WANT TO | 1744 | 20.9 | I SAID TO | 726 | 18.1 |
| YOU WANT TO | 18 | 15 | 9 - GOING TO HAVE | 1713 | 20.1 | GO TO THE | 630 | 15.7 |
| THEY USED TO | 18 | 15 | 10 - WE NEED TO | 1652 | 19.8 | I WANT TO | 603 | 15.0 |
| TO BE HONEST | 16 | 13.3 | 11- WE RE GOING | 1601 | 19.2 | DON'T WANT TO | 593 | 15.0 |
| TO DO IT | 15 | 12.5 | 12 - WE RE GOING TO | 1573 | 18.9 | HAVE TO GO | 555 | 13.8 |
| WE WENT TO | 14 | 11.7 | 13 - TO HAVE TO | 1570 | 18.9 | I USED TO | 540 | 13.4 |
| DO YOU WANT | 14 | 11.7 | 14 - IS GOING TO | 1516 | 17.9 | I'M GOING TO | 539 | 13.4 |
| HE WENT TO | 12 | 10 | 15 - NOT GOING TO | 1489 | 17.8 | I'VE GOT TO | 535 | 13.4 |
| TO LISTEN TO | 12 | 10 | TO DO IT | 1482 | 17.8 | WANT TO GO | 470 | 11.7 |
| HAVE TO GO | 12 | 10 | ARE GOING TO | 1450 | 17.4 | I HAD TO | 456 | 11.3 |
| DO YOU WANT TO | 11 | 9.2 | WE HAVE TO | 1393 | 16.7 | I HAVE TO | 439 | 10.8 |
| YOU HAD TO | 10 | 8.3 | DON T WANT | 1362 | 16.3 | TO DO WITH | 437 | 10.8 |
| TO LIVE IN | 10 | 8.3 | TO BE ABLE TO | 1269 | 15.2 | DO YOU WANT | 435 | 10.7 |
| HE USED TO | 10 | 8.3 | I M GOING | 1262 | 15.2 | I'LL HAVE TO | 431 | 10.7 |
| KNOW WHAT TO | 10 | 8.3 | TO DO THAT | 1258 | 15.2 | TO GO AND | 424 | 10.6 |
| IT USED TO | 9 | 7.5 | HAVE TO BE | 1247 | 15.0 | USED TO BE | 421 | 10.5 |
| I SAID TO | 9 | 7.5 | YOU VE GOT | 1246 | 15.0 | YOU'LL HAVE TO | 418 | 10.5 |
| HAVE TO DO | 9 | 7.5 | TO BE A | 1232 | 14.8 | GOING TO BE | 415 | 10.4 |
| HAD TO GO | 9 | 7.5 | TO HAVE A | 1228 | 14.7 | GOT TO GO | 402 | 10.0 |
| DON'T HAVE TO | 9 | 7.5 | I M GOING TO | 1216 | 14.6 | DON'T HAVE TO | 397 | 9.9 |
| I NEED TO | 9 | 7.5 | WE WANT TO | 1213 | 14.6 | HAVE TO DO | 394 | 9.8 |
| USED TO LIVE | 9 | 7.5 | DON T WANT TO | 1158 | 13.9 | YOU GOING TO | 394 | 9.8 |
| YOU GO TO | 8 | 6.4 | GO TO THE | 1157 | 13.9 | SUPPOSED TO BE | 384 | 9.4 |
| TO GET TO | 8 | 6.4 | TO DO IS | 1151 | 13.8 | TO BE A | 380 | 9.4 |
| GO TO BED | 8 | 6.4 | YOU VE GOT TO | 1135 | 13.6 | TO DO THAT | 348 | 8.7 |
| USED TO WORK | 8 | 6.4 | TO LOOK AT | 1087 | 13.0 | I WENT TO | 334 | 8.3 |
| THEY USED TO | 172 | 4.3 | | | | | | |

Table 12(a) Highest frequency TO group clusters SCO, MAC & BNC/C

### 10.8.1 Frequent TO group clusters compared in 4 corpora

It became an issue of concern that MAC:MED shows a number of clusters in frequent use which are not frequent in the other two corpora. It could have been that the (previously not used) MAC:MED is unreliable. For that reason, a first comparison is made between SCO and three larger corpora: the 4.0 million word BNC/C, the 8.3million word MAC:MED and the 9.3million word BoE (UKspoken).

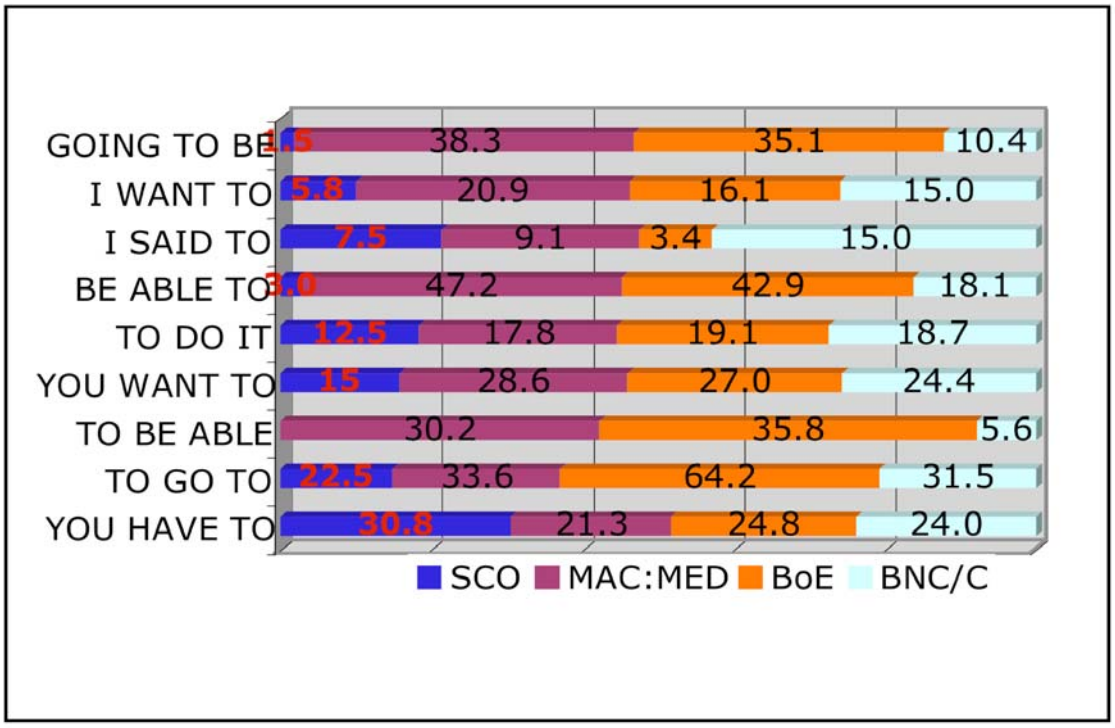



Figure1: Top *to* clusters compared in SCO, MAC:MED, BoE and BNC/C

Figure 1 demonstrates a number of salient facts for the top *to* clusters that can be found in all four corpora:

1. *To* can appear in a way that is unique to only one of the four corpora compared.

2. There are also a number of clusters where the proportional frequency is similar in a rather broad way across all corpora.

There are two further insights that are specific to SCO:

1. *To* can be found to be used substantially less frequently in SCO than in the other corpora for a number of clusters.
2. *To* is never used markedly more often in any SCO cluster than in another corpus' cluster.

*TO* group clusters appear to reveal the mystery of corpora. All of the comparators are big enough to equalise any corpus-specific context dependencies, so a large degree of agreement could have been expected. It is therefore somewhat disconcerting that 3 major corpora do not agree on clusters involving one of the most common items (see also Table 12a).

| SCO | frq | per 100k | MAC:MED | frq | per 100k | BNC/C | frq | per 100k |
|---|---|---|---|---|---|---|---|---|
| 1 - YOU HAVE TO[3] | 37 | 30.8 | 7 - YOU HAVE TO | 1772 | 21.3 | 3 - YOU HAVE TO | 963 | 24 |
| 2- USED TO BE | 28 | 23.3 | USED TO BE | 455 | 5.5 | 23 - USED TO BE | 421 | 10.5 |
| 3- TO GO TO | 27 | 22.5 | 3- TO GO TO | 2806 | 33.6 | 1- TO GO TO | 1266 | 31.5 |
| 4 - I HAVE TO | 21 | 17.5 | I HAVE TO | 691 | 8.3 | 18 -I HAVE TO | 439 | 10.8 |
| 5 - GO TO THE | 19 | 15.8 | 29 - GO TO THE | 1157 | 13.9 | 9 - GO TO THE | 630 | 15.7 |
| 6 - WE USED TO | 19 | 15.8 | WE USED TO | 375 | 4.5 | 51 - WE USED TO | 268 | 6.7 |
| 7 – I USED TO | 19 | 15.8 | 70- I USED TO | 741 | 8.9 | 13 - I USED TO | 540 | 13.4 |
| 8 - I WENT TO | 18 | 15 | I WENT TO | 346 | 4.2 | 35 - I WENT TO | 334 | 8.3 |
| 9 - YOU WANT TO | 18 | 15 | 5- YOU WANT TO | 2376 | 28.6 | 2- YOU WANT TO | 982 | 24.4 |
| 10 - THEY USED TO | 18 | 15 | THEY USED TO | 223 | 2.7 | THEY USED TO | 172 | 4.3 |
| 11- TO BE HONEST | 16 | 13.3 | TO BE HONEST | 123 | 1.5 | TO BE HONEST | 107 | 2.7 |
| 12 - TO DO IT | 15 | 12.5 | 16- TO DO IT | 1482 | 17.8 | 6 - TO DO IT | 754 | 18.7 |
| 24 - DON'T HAVE TO | 9 | 7.5 | DON'T HAVE TO | 529 | 6.3 | 27- DON'T HAVE TO | 397 | 9.9 |
| 24 -I SAID TO | 9 | 7.5 | I SAID TO | 761 | 9.1 | 8 - I SAID TO | 726 | 18.1 |
| 38 - I WANT TO | 7 | 5.8 | 8 - I WANT TO | 1744 | 20.9 | 10 - I WANT TO | 603 | 15 |
| 38 - I LIKE TO[4] | 7 | 5.8 | I LIKE TO<br>I'd like to | 117<br>732 | 1.4<br>9.0 | n/a<br>I'd like to | n/a<br>210 | n/a<br>5.0 |
| 38 - TO BE A | 7 | 5.8 | 23 - TO BE A | 1232 | 14.8 | 31 - TO BE A | 380 | 9.4 |
| 39 - WAS TALKING TO | 6 | 5.0 | WAS TALKING TO | 94 | 1.1 | WAS TALKING TO | 113 | 2.8 |
| 39 - GOING TO THE | 6 | 5.0 | GOING TO THE | >50 | n/a | GOING TO THE | 202 | 5.0 |
| GOING TO BE | 2 | 1.5 | 2- GOING TO BE | 3194 | 38.3 | 25- GOING TO BE | 415 | 10.4 |
| GOING TO HAVE | n/a | n/a | 9- GOING TO HAVE | 1713 | 20.1 | GOING TO HAVE | 266 | 6.7 |
| I M GOING TO | 5 | 4.0 | 26- I M GOING TO | 1216 | 14.6 | 14- I'M GOING TO | 539 | 13.4 |
| BE ABLE TO | 4 | 3.0 | 1- BE ABLE TO | 3934 | 47.2 | 4- BE ABLE TO | 904 | 22.5 |
| 39 -YOU HAVE TO DO | 6 | 5.0 | YOU HAVE TO DO | 63 | 0.70 | Table 12(b) Highest frequency TO group clusters directly compared | | |
| 39 - DON'T KNOW WHAT TO | 6 | 5.0 | DON'T KNOW WHAT TO | 83 | 0.90 | | | |

Table 12(b): Highest frequency TO group clusters directly compared

[3] It may appear odd this is the highest occurring *TO* group cluster only in SCO: O'Keefe *et al.* (2007: pp.66 f.) list it as the highest occurring 3w TO cluster in both the UK & US spoken parts of the *Cambridge International Corpus*.

[4] I LIKE TO (recorded in SCO and MAC:MED) and I'D LIKE TO (recorded in MAC:MED and BNC/C) suffers from the "recorders dilemma": has it actually been said one way or the other or has the transcriber edited this subconsciously? If all recorded occurrences are taken into account, SCO and BNC/C show the same proportional frequency of occurrence; MAC:MED records usage as a proportionally twice as high. Yet, with the uncertainty of what utterances were actually made in the background, I shall not discuss this particular cluster.

Table 12(b) shows that not one of the comparator corpora shows frequencies for their most prominent *TO* group clusters that are fully comparable with any of the corpora.

Looking at *to go to*, we find it is used widely in BoE – twice as often than in MAC:MED or BNC/C (three times as often as in SCO). Similarly, the less commonly used *I said to* is used more often (15 times in 100.000 words) in BNC/C than in MAC:MED (9.1 times), SCO (7.5 times) or BoE (3.4 times).[183] That *TO* group clusters allow for such a wide spread when four corpora are compared seems to indicate that *to* can be found in environments that are very specific to each corpus.

*To be able to*, *be able to* and *going to be* are all highly frequent clusters in BOE and MAC:MED only. They are far less frequent in BNC/C and marginal only in SCO. The discussion of *to* is the only word that has shown such extreme discrepancies between 2, 3 or even 4 corpora in this paper. I have no explanation for these extreme discrepancies. This, it must be noted in passing, highlights how hyper-sensitive *to* is as a mirror of language use - in extremis, the validity of corpora are called into question, as a common item like *to* should be found in a more uniform set of clusters. However, the framework of this thesis does not allow the space to explore this phenomenon in more detail.

In order to simplify and focus the research into *TO group* usage, Table 13 looks at the most frequent *to* 3w clusters and makes a pairwise

[183] Given that the BoE and the BNC are being used as standard corpora for English corpus linguistics, having a relatively common cluster appearing 5 times more frequently in the BNC/C has to be given serious consideration as to the validity of the respective statistics.

comparison with the numbers recorded for MAC:MED. The Log-Likelihood test compares occurrences in SCO compared to MAC on the basis of the full size of the respective corpora. Table 13 shows that there are a large amount of significantly divergent frequencies of occurrence with respect to *to* 3w clusters found where the two corpora are compared. These will be looked at in detail in the following subsections.

| *To* cluster | SCO freq | per 100k | MAC freq. | per 100k | LL |
|---|---|---|---|---|---|
| 1 - YOU HAVE TO | 37 | 30.8 | 1772 | 21.3 | 4.64 |
| 2- USED TO BE | 28 | 23.3 | 455 | 5.5 | 37.80 |
| 3- TO GO TO | 27 | 22.5 | 2806 | 33.6 | 4.77 |
| 4 - I HAVE TO | 21 | 17.5 | 691 | 8.3 | 9.27 |
| 5 - GO TO THE | 19 | 15.8 | 1157 | 13.9 | 0.35 |
| 6 - WE USED TO | 19 | 15.8 | 375 | 4.5 | 20.34 |
| 7 – I USED TO | 19 | 15.8 | 741 | 8.9 | 5.31 |
| 8 - I WENT TO | 18 | 15 | 346 | 4.2 | 19.94 |
| 9 - YOU WANT TO | 18 | 15 | 2376 | 28.6 | 8.94 |
| 10 - THEY USED TO | 18 | 15 | 223 | 2.7 | 31.77 |
| 11- TO BE HONEST | 16 | 13.3 | 123 | 1.5 | 40.63 |
| 12 - TO DO IT | 15 | 12.5 | 1482 | 17.8 | 1.98 |
| WE WENT TO | 14 | 11.7 | 279 | 3.35 | 14.80 |
| HE WENT TO | 12 | 11.7 | 50 | 0.6 | 42.80 |
| TO LISTEN TO | 12 | 10.0 | 383 | 4.60 | 5.68 |
| HAVE TO GO | 12 | 10.0 | 795 | 9.53 | 0.04 |
| DO YOU WANT TO | 11 | 10.0 | 590 | 7.08 | 0.70 |
| YOU HAD TO | 10 | 9.2 | 107 | 1.3 | 19.98 |
| TO LIVE IN | 10 | 8.3 | 51 | 0.6 | 32.27 |
| HE USED TO | 10 | 8.3 | 339 | 4.43 | 4.11 |
| KNOW WHAT TO | 10 | 8.3 | 164 | 1.9 | 13.36 |
| IT USED TO | 9 | 7.5 | 59 | 0.6 | 25.25 |
| I SAID TO | 9 | 7.5 | 761 | 9.38 | 0.34 |
| HAVE TO DO | 9 | 7.5 | 859 | 10.30 | 0.95 |
| HAD TO GO | 9 | 7.5 | 343 | 4.44 | 2.70 |
| DON'T HAVE TO | 9 | 7.5 | 529 | 6.3 | 0.29 |
| I WANT TO | 8 | 6.4 | 1744 | 20.9 | 15.49 |
| TO BE A | 7 | 5.8 | 1232 | 14.8 | 8.20 |
| TO DO IS | 6 | 5.0 | 1151 | 13.8 | 8.69 |
| I'M GOING TO | 5 | 4.0 | 1216 | 14.6 | 12.16 |
| BE ABLE TO | 4 | 2.5 | 3934 | 38.2 | n/a |
| GOING TO BE | 2 | 1.3 | 3194 | 47.2 | n/a |
| BE ABLE TO* | 8 | 2.5 | 7868 | 38.2 | 165.13 |
| GOING TO BE ** | 6 | 1.3 | 9582 | 47.2 | 176.19 |

*For LL, a projected doubling of the figures has been assumed.
**For LL, a projected trebling of the figures has been assumed.

Table 13: Pairwise comparison of SCO most frequent *TO Group* clusters with MAC equivalents.

#### 10.8.2.1 Comparing TO group clusters in SCO with equivalent MAC:MED and BNC/C clusters

In the previous sections, the fact that *to* appears in a large variety of chunks that are entirely context-dependent has been shown. In 10.8.2, I try to explore whether any specific usage patterns that are unique to SCO can nevertheless be detected and how far language use can form an interpretable basis for how SCO speakers are primed.

Amongst the twenty-five 3w and 4w *TO* group clusters there are only the following four that are roughly similar in their proportional frequencies[184]:

| *TO group cluster* | SCO occ. per 100k words | MAC:MED occ. per 100k words | BNC/C occ. per 100k words |
|---|---|---|---|
| YOU HAVE TO | 30.8 | 21.3 | 24.0 |
| GO TO THE | 15.8 | 13.9 | 15.7 |
| TO DO IT | 12.5 | 17.8 | 18.7 |
| DON'T HAVE TO | 7.5 | 6.3 | 9.9 |

Table 14: Long *TO* group clusters with similar frequencies in SCO, MAC:MED and BNC/C.

Yet the four clusters shown in Table 14 are the exception. And while there are some clusters where the proportional frequencies are similar for SCO and BNC/C or MAC:MED (*I used to* or *I said to* for example) the majority of clusters compared show preference by either SCO on the one side or MAC:MED *and* BNC/C on the other. Below, I will discuss these differences in detail.

[184] In Table 13 we can see that the LL figures are very low for this clusters, too.

Table 15: SCO lower proportional occurrence TO group clusters directly compared.

| SCO | freq | per 100k | MAC: MED | freq | per 100k | LL SCO:MAC | BNC/C | freq | per 100k |
|---|---|---|---|---|---|---|---|---|---|
| 3- TO GO TO | 27 | 22.5 | 3- TO GO TO | 2806 | 33.6 | 4.77 | 1- TO GO TO | 1266 | 31.5 |
| 9 - YOU WANT TO | 18 | 15 | 5- YOU WANT TO | 2376 | 28.6 | 8.94 | 2- YOU WANT TO | 982 | 24.4 |
| 38 - I WANT TO | 7 | 5.8 | 8 - I WANT TO | 1744 | 20.9 | 15.49 | 10 - I WANT TO | 603 | 15 |
| 38 - TO BE A | 7 | 5.8 | 23 - TO BE A | 1232 | 14.8 | 8.20 | 31 - TO BE A | 380 | 9.4 |
| GOING TO BE | 2 | 1.5 | 2- GOING TO BE | 3194 | 38.3 | | 25- GOING TO BE | 415 | 10.4 |
| GOING TO HAVE | n/a | n/a | 9- GOING TO HAVE | 1713 | 20.1 | | GOING TO HAVE | 266 | 6.7 |
| I'M GOING TO | 5 | 4.0 | 26- I'M GOING TO | 1216 | 14.6 | 12.16 | 14- I'M GOING TO | 539 | 13.4 |
| BE ABLE TO | 4 | 3.0 | 1- BE ABLE TO | 3934 | 47.2 | | 4- BE ABLE TO | 904 | 22.5 |

Table 15: SCO lower proportional occurrence TO group clusters directly compared

### 10.8.2.2 SCO TO group clusters less preferred

There is a group of clusters that are frequent in MAC:MED, BNC/C and also BoE but are very marginal in SCO.

Differences can be seen in the ***can do*** clusters *to be able and to be able to* as well as in the ***future*** cluster *going to be*. These appear well below five occurrences per 100.000 words in SCO, while being proportionally far more frequent in BNC/C, MAC:MED and BoE (amongst the most frequent *TO* group clusters for the latter two).

The probably largest difference amongst *TO* group clusters is associated with the phrase *be able to*. This is the highest occurring *TO* group 3w cluster in MAC:MED, and the 4th highest in BNC/C and BoE. It is however barely used at all in SCO. The contrast is stark: *be able to* occurs 3 times per 100k words in SCO. That is proportionally 7.5 times less than in BNC/C (22.5 times per 100k words) and proportionally over 15 times less often in MAC:MED (or BoE).[185] The other cluster with *be* that is found with high occurrences in the comparator corpora – *to be able* – occurs not once in SCO.

This clear underuse of *be able* phrases is striking and stands out. On a small sample as provided by SCO, no conclusive answers can be given[186].

[185] The occurrence pattern of the cluster *be able to* has been compared in 4 written and 7 spoken corpora. In the written data, it was rare in Shakespeare, and much less used by Dickens than other 19th century novelists. Amongst the spoken corpora, *be able to* occurs amongst the top five most frequent 3w clusters in **every single corpus** but not in SCO. Cf. Michael Pace-Sigge (2009): http://www.scribd.com/doc/25388239/Why-to-is-a-Weird-Word-4509 .

[186] See Table 13 for projected log-likelihood figures were SCO 2-3 times larger. On projected figures, the difference of use is stark.

Other clusters however, show valid, significantly divergent frequencies of use, where SCO and MAC are compared. *I'm going to, you want to* and *I want to* all appear with a far lower ranking in SCO than in the comparators.

*I'm going to* proportionally occurs only one-third as frequent in SCO than it does in MAC:MED or BNC/C. In a pairwise comparison (SCO - MAC:MED), the statistical test shows that this difference is significant above the 99.9% level.

*You want to* is both in its frequency and its ranking proportionally less well used amongst SCO speakers than amongst either MAC:MED or BNC/C speakers. Still, with this fixed phrase the biggest difference is between the use in SCO (15 occ. per 100k words) and MAC:MED (28.6 occ. per 100k words). The divergence between SCO and MAC is significant above the 99.0% level. The difference is more marked, however, when the speaker makes his own wishes clear. While *I want to* is less used in all three corpora than *you want to*, personal wishes are only occurring every 5.8 times per 100k words in SCO, but occur nearly three times as often in BNC/C (15 times per 100k words) and nearly four times more often in MAC:MED (20.9 times per 100k words), this means the clusters is the most significantly (above the 99.99% level) underused *TO* group cluster in SCO in comparison to either MAC or BNC/C.

Table 16: SCO higher proportional occurrence TO group clusters directly compared.

| SCO | freq | per 100k | MAC: MED | freq | per 100k | LL SCO:MAC | BNCC | freq | per 100k |
|---|---|---|---|---|---|---|---|---|---|
| 2- USED TO BE | 28 | 23.3 | USED TO BE | 455 | 5.5 | 37.80 | 23 - USED TO BE | 421 | 10.5 |
| 4 - I HAVE TO | 21 | 17.5 | I HAVE TO | 691 | 8.3 | 9.27 | 18 - I HAVE TO | 439 | 10.8 |
| 6 - WE USED TO | 19 | 15.8 | WE USED TO | 375 | 4.5 | 20.34 | 51 - WE USED TO | 268 | 6.7 |
| 6 – I USED TO | 19 | 15.8 | 70- I USED TO | 741 | 8.9 | 5.31 | 13 - I USED TO | 540 | 13.4 |
| 7 - THEY USED TO | 18 | 15.0 | THEY USED TO | 223 | 2.7 | 31.77 | THEY USED TO | 172 | 4.3 |
| 7 - I WENT TO | 18 | 15.0 | I WENT TO | 346 | 4.2 | 19.94 | 35 - I WENT TO | 334 | 8.3 |
| WE WENT TO | 14 | 11.7 | WE WENT TO | 279 | 3.35 | 14.80 | WE WENT TO | 273 | 6.7 |
| HE WENT TO | 12 | 11.7 | HE WENT TO | 50 | 0.6 | 42.80 | HE WENT TO | 102 | 2.9 |
| YOU HAD TO | 10 | 9.2 | YOU HAD TO | 107 | 1.3 | 19.98 | YOU HAD TO | 144 | 3.5 |
| TO LIVE IN | 10 | 8.3 | TO LIVE IN | 51 | 0.6 | 32.27 | TO LIVE IN | 44 | 1.1 |
| IT USED TO | 9 | 7.5 | IT USED TO | 59 | 0.6 | 25.25 | IT USED TO | 123 | 3.2 |

Table 16: SCO higher proportional occurrence TO group clusters directly compared

### 10.8.2.3 SCO TO group clusters more preferred

The frequently used word *to* can also be found in a large number of 3w clusters where it is clearly preferred in SCO rather than in the other corpora. Table 16 shows that, in SCO, there is a significant preference for *to* to appear in 3w clusters with either *used* or *went.*

*Used to be* is the second highest occurring *TO* group cluster in SCO. It ranks far higher than in the comparators and is proportionally significantly more frequent.

Table 16 shows that the *obligation* phrase *I have to* is ranked the fourth most used *TO* group 3w cluster in SCO with 17.3 occurrences per 100k words. That is proportionally more than twice as frequent as its occurrences in MAC:MED (8.3 times) or in BNC/C (10.8 times per 100k words – ranked 18th). Yet when compare this to the phrase *you have to* (see Table 13), this difference is even stronger, as there is no significant divergence to be found in the use of *you have to.*

The most frequent *TO* group clusters where use noticeably diverges between SCO and the other corpora are all *to* with ***past tense***, notably *used.*

*Used to be* has been discussed but *we used to*, and *I used to* are ranked 6th , furthermore, *they used to* is ranked 7th (same number of occurrences as *I went to*) most frequent 3w *TO* word clusters in SCO. None of these clusters is within the 10 most frequently occurring 3w *TO* group clusters in either MAC:MED or BNC/C (see Table 16). Comparing *I*

*used to* occurrence patterns we find it occurs 15.8 times per 100k words in SCO and 13.4 times in BNC/C but only 8.9 times per 100k words in MAC:MED. Statistically, this difference is negligible. Looking at *we used to* however, the divergence is more prominent and statistically significant: It appears 15.8 times per 100k words in SCO, 6.7 times in BNC/C and only 4.5 times per 100k words in MAC:MED. The divergence is still more prominent when we look at *they used to*: It appears 15.0 times per 100k words in SCO, 4.3 times in BNC/C and only 2.7 times per 100k words in MAC:MED. This means that SCO clusters with *used to* appear more than 3 to 4 times more often than MAC:MED clusters. Given that *I went to, he went to* and *we went to* also appear significantly more frequently per 100k words in SCO than in the other corpora. Given, to that *you had to* is significantly more frequent in SCO than MAC (while there is little difference in the use of *you have to)* there is firm evidence that *TO* group clusters in SCO have a preference for appearing in past tense 3w clusters.

All in all, *TO group* clusters in SCO show different patterns of preference with regards to nesting and semantic association with *to +able* and *to* + **past tense** constructions clusters.

Table 17(a): Highest frequency *HONEST* group 3-6w clusters by occurrence rank.

| SCO | Freq. | per 100k | MAC:MED | Freq. | per 100k | Cluster BNC/C | Freq. | per 100k |
|---|---|---|---|---|---|---|---|---|
| 1- TO BE HONEST | 16 | 13.30 | 1- TO BE HONEST | 123 | 1.50 | 1-TO BE HONEST | 107 | 2.70 |
| 2 - BE HONEST WITH | 5 | 4.00 | 2- HONEST WITH YOU | 41 | 0.50 | 2 - BE QUITE HONEST | 36 | 0.90 |
| HONEST WITH YOU | 5 | 4.00 | 3 - BE QUITE HONEST | 39 | 0.46 | 3 - BE HONEST I | 31 | 0.80 |
| TO BE HONEST WITH YOU | 5 | 4.00 | 4 - TO BE QUITE HONEST | 36 | 0.43 | 4 - HONEST WITH YOU | 29 | 0.70 |
| TO BE HONEST WITH | 5 | 4.00 | 4 - BE HONEST WITH | 36 | 0.43 | 4 - TO BE QUITE HONEST | 29 | 0.70 |
| | | | 5 - BE HONEST WITH YOU | 35 | 0.42 | 5 - TO BE HONEST I | 24 | 0.67 |
| | | | 5 - BE HONEST I | 35 | 0.42 | 6 - BE HONEST WITH YOU | 21 | 0.50 |
| | | | 6 - TO BE HONEST I | 30 | 0.36 | 6 - BE HONEST WITH | 21 | 0.50 |
| | | | 7 - TO BE HONEST WITH | 23 | 0.28 | 7 - HONEST TO GOD | 20 | 0.50 |
| | | | 8 - BE PERFECTLY HONEST | 16 | 0.19 | 8 BE PERFECTLY HONEST | 18 | 0.45 |
| | | | 9 - TO BE PERFECTLY | 16 | 0.19 | 9 - TO BE PERFECTLY | 17 | 0.42 |
| | | | 9 - TO BE PERFECTLY HONEST | 16 | 0.19 | 9 – TO BE PERFECTLY HONEST | 17 | 0.42 |
| | | | 10 - HONEST TO GOD | 15 | 0.18 | 10 - TO BE HONEST WITH | 14 | 0.35 |
| | | | 11 - TO BE HONEST WITH YOU | 7 | 0.09 | 11- BE QUITE HONEST I | 12 | 0.33 |

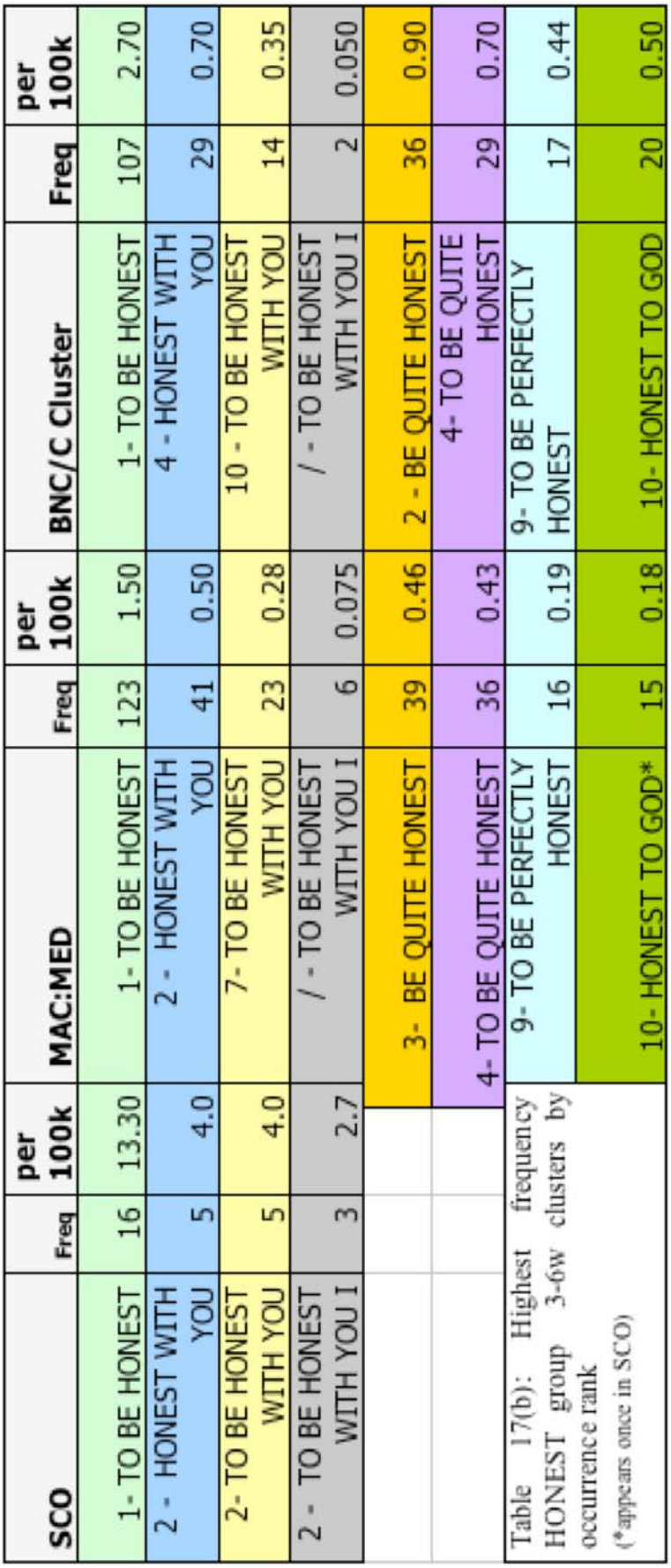
Table 17(a): Highest frequency *HONEST* group 3-6w clusters by occurrence rank

## 10.9 The *HONEST* Group

*Honest* is used with a proportionally far higher frequency in SCO than in any other corpus. When *HONEST* group clusters are compared it becomes obvious that SCO speakers use these clusters markedly more often and then mainly in one particular phrase. In the BNC/C and MAC:MED, *honest* tends to appear with prequalifiers like *quite* and *perfectly*. Yet these prequalified *honest* clusters still appear with only very low total frequencies; like all *honest* clusters in MAC:MED and BNC/C, they are rare.

Table 17(a) shows that the most frequent cluster in SCO, MAC:MED and BNC/C is the phrase *to be honest*. The one thing that is common to all three corpora is that the cluster is the most frequent by a wide margin, all other 3w clusters in the *HONEST* group occurring with far lower frequencies. *To be honest* appears 13.3 times per 100.000 words in SCO and only 2.7 times in BNC/C. This is nearly five times as frequent. Compared to MAC:MED the difference is even more striking as the 3w cluster appears only 1.5 times per 100k words here - meaning that it occurs nine times more frequently in SCO.

Table 17(b): Highest frequency HONEST group 3-6w clusters by occurrence rank.



| SCO | Freq | per 100k | MAC:MED | Freq | per 100k | BNC/C Cluster | Freq | per 100k |
|---|---|---|---|---|---|---|---|---|
| 1- TO BE HONEST | 16 | 13.30 | 1- TO BE HONEST | 123 | 1.50 | 1- TO BE HONEST | 107 | 2.70 |
| 2 - HONEST WITH YOU | 5 | 4.0 | 2 - HONEST WITH YOU | 41 | 0.50 | 4 - HONEST WITH YOU | 29 | 0.70 |
| 2- TO BE HONEST WITH YOU | 5 | 4.0 | 7- TO BE HONEST WITH YOU | 23 | 0.28 | 10 - TO BE HONEST WITH YOU | 14 | 0.35 |
| 2 - TO BE HONEST WITH YOU I | 3 | 2.7 | / - TO BE HONEST WITH YOU I | 6 | 0.075 | / - TO BE HONEST WITH YOU I | 2 | 0.050 |
| | | | 3- BE QUITE HONEST | 39 | 0.46 | 2 - BE QUITE HONEST | 36 | 0.90 |
| | | | 4- TO BE QUITE HONEST | 36 | 0.43 | 4- TO BE QUITE HONEST | 29 | 0.70 |
| | | | 9- TO BE PERFECTLY HONEST | 16 | 0.19 | 9- TO BE PERFECTLY HONEST | 17 | 0.44 |
| | | | 10- HONEST TO GOD* | 15 | 0.18 | 10- HONEST TO GOD | 20 | 0.50 |

Table 17(b): Highest frequency HONEST group 3-6w clusters by occurrence rank
(*appears once in SCO)

Table 17(b) shows the uses of *to be honest* clusters in the respective corpora. *Honest with you* is an independent cluster in both MAC:MED and BNC/C. It is used in the former 41 times and in the latter 29 times – roughly twice as often as its use in the 5w cluster *to be honest with you*. In SCO, however, *honest with you* is always part of the cluster *to be honest with you* as they occur both five times.

| ***HONEST* cluster** | **Freq. SCO** | **Freq. MAC:MED** | **Log-Likelihood** |
|---|---|---|---|
| TO BE HONEST | 16 | 123 | 40.63 |
| HONEST WITH YOU | 5 | 41 | 12.16 |
| TO BE HONEST WITH YOU | 5 | 23 | 17.00 |

Table 17(c): Areas of strongest divergence where SCO and MAC:MED *honest* clusters are compared.

The cluster *to be honest with you* is, as both the ranking and the percentages of use shows, only the most frequent *honest* group 4w cluster in SCO. *To be honest with you* appears proportionally over ten times more often (4.0 times per 100k words) in SCO than in MAC:MED (0.28 times per 100k words) or BNC/C (0.35 times per 100k words). *To be honest with you* itself mostly occurs in SCO (three times / 2.7 times per 100k words) as part of the 6w cluster *to be honest with you I* and is barely recorded in the much larger MAC:MED (six occurrences in total) or BNC/C (two occurrences in total). As Table 17(c) shows, *to be honest* and the longer cluster incorporating it 1 out of 3 times, *to be honest with you* are significantly more frequently used in SCO than in MAC. Though the difference of proportional frequencies are smaller for *honest with*

*you*, the divergence between SCO and MAC remains significant above the 99.9% level.

*Honest* is a rarely-occurring term in the comparator corpora. However, *to be honest* is ranked 11th most used 3w cluster in the *TO* group in SCO, highlighting that *honest* is used disproportionably more in the Liverpool than in the other corpora. *To be honest* and *to be honest with you (I)* are, therefore, fixed phrases that are primed for frequent use amongst SCOuse speakers while their use is rare for English speakers across the UK.

## 10.10 Conclusions on clusters

This chapter provides a more detailed comparison of the use of the highest-occurring clusters in SCO with their use in both the BNC/C and MAC, with an additional comparison with a fourth corpus, the Bank of English and with the extended MAC corpus, MAC:MED.

While the decision made as to what cluster groups to focus on was based on a direct key-cluster comparison of BNC/C and SCO, the comparisons following on from that are made between SCO and MAC as well as BNC/C.

The findings mirror some of the points that were seen when key words were directly compared. Overall, the same clusters can be found in all three (or four) corpora and the differences are in the proportional frequencies. As in earlier chapters, SCO diverges mainly in those key

terms and key clusters that are found to occur with medium-high frequency. On the whole, SCO frequencies and ranking of usage for clusters are different from those for the equivalent clusters in MAC and BNC/C. In other words, MAC (MAC:MED) and BNC/C (and, where compared, BoE) tend to be closer to each other in their proportional frequencies for the majority of clusters while they tend to diverge from SCO.

The most important findings are twofold:

1) There are a number of extended phrases (usually longer than three words) that SCO speakers appear to be primed to use with preference, while there is non-preference for other phrases. (*like* and *mean* clusters provide examples here).
2) There are cases where the colligational structure and the semantic associations of the language are clearly different in SCO when compared to the other English spoken corpora. (*thought* rather than *think* cluster distinctions).

So we can find phrases like *like you know* and *you know what I mean* as clearly identifiable as Scouse preferred choices. Conversely, the functional question *do you know* appears about only half as often in spoken Liverpool English than in the general UK spoken English corpora.

The *LIKE* group of clusters demonstrates divergence on both fronts. In SCO, *stuff like that* is a highly preferred cluster in the *LIKE* group, while it is marginal in MAC. The opposite is true for the cluster with *would like* which is used strongly in MAC but barely occurs in SCO.

Consequently, we do not speak simply of a difference of frequencies but the differences in the *nesting* of *like* is also shown.

The most noticeable differences can be found when one of the highest-occurring words in spoken English is compared, the function word *to*. The comparison involves 4 corpora as there seem to be vast discrepancies between each single corpus compared to the next. Where direct comparison of those clusters that are found throughout is possible, the contrast between SCO on the one side and BoE, BNC/C and MAC:MED on the other shows that the colligational structure and the field of semantic associations differ strongly. While the English spoken corpora all have combinations with *able to* as amongst the highest occurring clusters with *to*, the Liverpool English SCO barely records it. While BoE, BNC/C and MAC:MED all refer to future actions with clusters incorporating *going to*, this, again is rarely occurring in SCO. SCO, while also using TO less with verbs in the present tense, has a marked preference to refer back to the past with the inclusion of *used to* clusters that occur far more sparingly in the other corpora.

On the whole, the comparison of clusters shows us where there are clusters and phrases that are noticeably preferred or dispreferred by SCO when compared to other English Spoken corpora. It also can be tentatively constructed as unearthing the (lack of) confidence and self-perception of the speakers as reflected by their – subconscious – use of language.

# Chapter 11 Conclusions

In this thesis, I have looked at two main issues.

The first issue looked at is whether Liverpool English (Scouse) is an accent or a dialect, and how far *corpus linguistic* tools can be used to describe difference between variants. The second issue is strongly linked to this - to decide how far *Lexical Priming* is a valid theory that can be applied to Spoken English material.

Traditionally, dialectologists have focussed on rare words and constructions and based their decision on what to treat as a separate dialect on the degree of divergence found with regard to these words or constructions. The missing part of this argument is, at what point would you have sufficient difference to warrant treating a group of speakers as a separate Speech Community? When is it appropriate for us to speak of it as a different variety? What is the tip-over point? I make the claim that, in theory, it should also be possible to identify a dialect by behaviour of the *common* words, not the *specific* words. The corpus linguistic approach used in this thesis, therefore, has focused on common lexical

items and looked for divergence in their use. This has potentially extended the tools available to dialectology and made the notion of variety more subtle.

I have not found a consistent high degree of divergence between the Corpus of Liverpool Spoken English that I collected (SCO) and the two general UK corpora that I used. Nevertheless, there are a number of truly significant differences, for example short clusters and phrases that are more prominently used in SCO than in the comparators, but these seem insufficient in number to warrant interpreting them as evidence of Liverpool English having the status of a dialect. In other words, this entirely new way of determining what is a dialect (i.e. looking at how usage of common words diverges) has proved to be another method to confirm the traditional view of Liverpool English as an accent rather than a dialect.

I might suggest, in future research, that it would be worthwhile to take an agreed, recognised dialect and to do a key word and key cluster analysis by comparing it to one or two other recognised dialects. This should reveal two points: if two dialects are compared to a common third (a "standard") we should not just find where there are areas of divergence between the corpora but also, which non-standard features are shared between dialects[187].

[187] In this context it is interesting to note that Visberg (2010) has discovered that the Swedish translations of *stand, sit* and *lie* are closest to their use in German, whereas both Finnish and English use forms of *to be*. (e.g. for *the plate* ***is*** on the table, the literal translation would be *the plate* ***lies*** *on the table*). While this leads to different proportional frequencies of occurrence for the target words, the difference is subtle. This, tentatively, can be seen as confirming my thesis that different variants show their specific characteristics through subtle divergent use of common words.

Any future follow-up study would certainly need to work with a larger SCO corpus. This thesis presents a number of cases where I can only project statistically relevant differences, yet lack of numbers make conclusions unreliable[188]. Where reliable data is already given, however, a future study should look in how far *nesting* diverges as well. The issue of *nesting* has been touched upon several times but needs to be researched in far more detail. Ideally, a future study would also have sufficiently large Liverpool and North-West English speaker corpora available in order to find out whether current SCO findings are unique for the Liverpool area or are rooted in a wider North West of England use.

Looking at the second issue, I noted in Chapter 3 that psycholinguistic *priming* experiments are very much based on speaking and listening evidence. The hypothesis was that claims made for *lexical priming* which, so far, have been based on material based on written text, should therefore be equally valid for naturally occurring spoken language.

This is supported by the investigations reported in this thesis, as my findings can be seen as sufficiently supporting the idea of *lexical priming* even though they are insufficient to justify calling Liverpool English a separate dialect of English. While traditional dialectologists look for absolute difference (i.e. unique words), corpus linguists look for relative difference (i.e. difference in proportional frequency of use). The notion of

[188] One approach to extend the material of the SCO corpus would be to add further recorded data. However, instead of choosing the time-consuming method of doing a full transcription, a partial transcription could be made where the transcriber would only pick up on words and clusters discussed in this thesis. The obvious drawback, however, would be that the "size" of this *selective corpus* can only be an estimate in relation to what the actual size of the full-transcription corpus would be.

"dialect" here indeed could be argued to have become less relevant, if we retain the idea that a variant called a dialect has clear identifying characteristics. This ties in with the priming hypothesis that "everybody has an idiolect: these idiolects differ in subtle ways from person to person" (cf. Hoey 2005:181). At the same time, no one idiolect can be so significantly different from a uniform set of primings as to break the chain of communicability. It is the area between the personal idiolect (evidence gathered from a single speaker) and uniform features (collocations, colligations and semantic associations that are found to be similar in comparable similar corpora, for example material in both MAC and BNC/C) that the study undertaken here can be seen as consistent with *lexical priming*.

This leads to the crucial issue about *lexical priming* that has not been discussed yet: whether or not the differences in *primings* found are sufficiently strong to support the claims made by the theory. If we look at the socio-economic, cultural and geographical set-up of Liverpool, we find a fairly homogeneous, a fairly tight community. It follows that, if the theory is correct, members within such a community to a degree will influence (*prime*) each other and that these primings will be mutually self-reinforcing. This, furthermore, would mean that features of *Scouse* ought to be found.

Within the speech community of Liverpool English speakers, there are particular words and clusters of words that, though not unique, appear to be more strongly preferred than in MAC and BNC/C. And a

number of divergent preferences have been shown to be in medium-high frequent clusters. We have seen that clusters appear in different structural formations, for example *you know // yeah* and *yeah// you know* are two-speaker clusters in SCO, while they appear only to be found mid-utterance in MAC. This thesis shows that clusters with target words like *honest, just, like* and *well* appear proportionally more frequently in SCO than in MAC while, conversely, clusters with words like *don't, know, think* and *yeah* are proportionately less frequent[189].

All in all, however, as already noted, we have not enough evidence of *Scouse*-specific material to call *Scouse* a dialect. The question is, then: does this also mean that there is not enough evidence to support the notion of *lexical priming* in evidence in Scouse.

When we look at the evidence present for *lexical priming* in SCO, what we find is not a massive, but rather a subtle difference - and therefore different degrees of likelihood of use. With reference to John Sinclair (2004) we can say that patterns found in the English Language are based on *likelihoods* and not on *certainties*. As a consequence, we find *lexical priming* expressed in *Scouse* expressed through the greater likelihood of (or, conversely, lower preference for) key-words and key-phrases used by speakers within this speech community.

[189] Phrases in this thesis found to be significantly more frequent in SCO are : *You know what I mean, I just, Anything like that, people say, as well, it was really, just like, stuff like that, I was like, and they're like, used to be* and *to be honest*. Significantly more frequent in MAC were the following: *yeah yeah, yeah well, you know // yeah, do you know, really really really, I don't like, I don't think, and I think* and *I mean I*.

These patterns are found to be neither idiolects (single speaker occurrence) nor are they widely used in other parts of the UK. Their prominent use on Merseyside can be seen as evidence of patterns within a speech community that have become self-re-enforcing. Users have presumably become primed by the constant usage of the speakers they engage with on a daily basis. Within any community, there are variables that are different by degree - and each member of this community needs to know them to fully fit into this community. Looking at these degrees of variation could be argued to show the patterns of priming of such a particular community.

In the course of this thesis, it has been found that significant differences in the frequencies of *collocations* can be, yet are not necessarily, strong indicators to where we would find clusters that diverge strongly between corpora. What we have found with SCO is that the differences are mostly found to be significant for medium-high frequency clusters of the target words investigated. The statistical testing undertaken provides sufficiently strong evidence that speakers of this community are reflecting characteristic use of the English language that is consistent with the claims of the *Lexical Priming Theory*.

Specific usage that is in a strong associative bind with what has been sufficiently often (and/or sufficiently strongly) experienced and subsequently and successfully employed by users is congruent with the *Lexical Priming Theory*. When comparing natural occurring language of a

select speech community with natural occurring language representing an average found across the United Kingdom, we can find that the collocations of one word, the colligations of one word, and the semantic associations of a word is, to a degree, more preferred in one of the two. Where we have found patterns of such a preference or non-preference in the SCO corpus that are divergent from the patterns found in a general corpus (like the MAC or BNC/C), these patterns found amongst respective groups of speakers are congruent with *Lexical Priming Theory* hypotheses.

# Bibliography


Aijmer, Karin (1985). Just. In: Bäckman, Sven & Kjellmer, Göran (eds.), Papers on language and literature presented to Alvar Ellegård and Erik Frykman. Gothenburg: Gothenburg Studies in English 60. pp. 1–10.

Aijmer, Karin (2002). *English Discourse Particles: Evidence from a corpus*. John Benjamins: Amsterdam / Philadelphia.

Aijmer, Karin & Simon-Vandenberge, Anne-Marie (2004). A model and a methodology for the study of pragmatic markers: the semantic field of expectation. In: *Journal of Pragmatics* 36. pp. 1781–1805.

Aijmer, Karin & Stenstroem, Anna-Brita (2005). Approaches to interaction. In: *Journal of Pragmatics* 37. pp. 1743–1751.

Aitchinson, Jean (1989). *The Articulate Mammal*. Routledge, London.

Angell, James Rowland (1896). Review of *Outlines of Psychology* by Oswald Külpe. In: *The Philosophical Review*, Vol. 5, No. 4, pp. 417-421.

Ash, Sharon (2002). Social class. In: *The Handbook of Language Variation and Change.* Chambers, J. K., Trudgill, P. & Schilling-Estes, N. (eds.). Blackwell Publishing: Oxford. pp. 402-422

Ashcraft, M.H. (1976). Priming and property dominance effects in semantic memory. In: *Memory & Cognition,* Vol. 4, No. 5. pp. 490-500.

Atkins, S; Rundell, M. and Sato, H. (2003). The contribution of framenet to practical lexicography. In: *International Journal of Lexicography,* Vol. 16 No. 3. London, OUP. pp. 333-357

Austin, JL (1962) [2001]. *How to do things with words*. Oxford, OUP.

Bache, Carl (1978). The Order of Pre-modifying Adjectives in Present-Day English. Odense: Odense University Press.

Bahns, Jens (1993). Lexical collocations: a contrastive view. In: *ELT Journal* Volume 47/1 January. pp. 56-63

Baker, Paul (2006). *Using Corpora in Discourse Analysis.* London, Continuum.

Bauer, Laurie & Bauer, Winifred (2001). Adjective Boosters in the English of Young New Zealanders. In: *Tres. New Zealand English Journal* 14. pp. 7-17.
Also in: Journal of English Linguistics (2002) 30. pp. 244-257.

Bernstein, Basil (1971). *Class, Codes and Control.* Volume 1. London: Routledge & Kegan Paul.

Belchem, John (2000). 'An accent exceedingly rare' : Scouse and the inflexion of class. In: *Merseypride. Essays in Liverpool exceptionalism.* Liverpool: Liverpool University Press. pp. 31-64.

Lou Bernard: PALC97: Practical applications of language corpora (University of Lodz, April 12-14 1997) http://users.ox.ac.uk/~lou/reports/9801mcenery.htm Accessed 07/11/2007

Biber, D; Conrad, S; Reppen, R: (1998). *Corpus Linguistics*. Cambridge: CUP.

Biber, Douglas, Johansson, S.; Leech, G.; Conrad, S. & Finegan, E. (2000). *Longman Grammar of Spoken and Written English*.

Biber, Douglas; Conrad, S. & Leech, G. (2002). *Student Grammar of Spoken and Written English.* Harlow, Essex: Longman.

Biber, Douglas 2009: A Corpus-Driven Approach to Formulaic Language in English: Multi-Word Patterns in Speech and Writing. Presentation, Corpus Linguistics 2009 Conference, University of Liverpool.

Blakemore, D. (2002). Relevance and linguistic meaning: the semantics and pragmatics of discourse markers. Cambridge: CUP

Bracher, Mark (1993). *Lacan, Discourse and Social Change*. *A psychoanalytic cultural criticism.* Ithaca: Cornell University Press

Brazil, David (1995). *A Grammar of Speech.* Oxford: OUP

Brinton, L.C. (2003). *I mean*: the rise of a pragmatic marker. Paper presented at GURT 2003.

Brown, Gordon (MP) (1989). *Where there is greed. Margaret Thatcher and the Betrayal of Britain's Future.* Edinburgh: Mainstream Publishing.

Brown, Hedy (1985). *People, Groups and Society*. Milton Keynes: Open University Press.

Brown, R.W.; Black, A.H. & Horowitz, A.E. (1955). Phonetic symbolism in natural languages. In: *Journal of Abnormal and Social Psychology 50*, pp. 388–393.

Brunner, J.S. (1955). *A Study of Thinking*. New York: Wiley.

Bursill-Hall, G.L. (1960). The linguistic theories of J. R. Firth. In: *Thought from the Learned Societies of Canada*, Toronto. pp. 237-50

Burt, Ronald S. (2001). Attachment, Decay, and Social Network. *Journal of Organizational Behavior*, Vol. 22, No. 6. (Sep., 2001), pp. 619-643.

Bybee, Joan L.; Perkins, Revere D.; & Pagliuca, William (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world.* Chicago: University of Chicago Press

Cacoullos, Rena Torres (2002). *Le*: from pronoun to intensifier. In: *Linguistics* 40 – 2. pp. 285 – 318.

Cameron, Deborah (2001). *Working with Spoken Discourse.* London: Sage.

Carter, Ronald (2004). Grammar and Spoken English. In: *Applying English Grammar.* C. Coffin; A. Hewings & K. O'Halloran (Eds.) London: Arnold/OUP. pp. 25-39

Carter, Ronald & McCarthy, Michael (1998). *Vocabulary and Language Teaching*. Harlow: Pearson Education (form. Longman).

Carter, Ronald & McCarthy, Michael (1995). Grammar and the Spoken language. In: *Applied Linguistics*, Vol. 16, No. 2. pp. 141-158.

Carter, Ronald & McCarthy, Michael (1997a). *Exploring Spoken English*. Cambridge: Cambridge University Press.

Carter, Ronald & McCarthy, Michael (1997b). Written and Spoken vocabulary. In: Schmitt, N. & McCarthy, M. (eds.): *Vocabulary.* Cambridge: Cambridge University Press. pp. 20-39

Castner, Joanna E. (2007). Semantic and affective priming as a function of stimulation of the subthalamic nucleus in Parkinson's disease. In: *Brain* 130, 1395-1407

Cavalli-Sforza, Luigi Luca (2001). *Genes, Peoples and Languages.* London: Penguin.

Chaffe, Wallace L (1982). Integration and involvement in speaking, writing and oral literature. In: Tannen, D. (Ed.): *Spoken and Written Language: Exploring Orality and Literacy*. New Jersey: Ablex Publishing Corp. pp.35-53.

Chapman, S. (2006). *Thinking about Language*. *Theories of English.* Houndmills: Palgrave Macmillan.

Cheng, P. C-H & Ananya, R. (2006). A temporal signal reveals chunk structure in the writing of word phrases. In: *Proceedings of the Twenty Eighth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum. http://www.cogs.susx.ac.uk/users/peterch/papers/ChengCogSci06.pdf (last accessed 2/7/2010).

Cheng, Winnie & Warren, Martin (2008). //-> ONE country two SYStems//: The discourse intonation patterns of word associations In: Ädel, A & Reppen, R (eds.): *Corpora and Discourse. The challenges of different settings*. Amsterdam/ Philadelphia: John Benjamins. pp. 135-153.

Christakis, N.A. & Fowler, J.H. (2008). The collective dynamics of smoking in a large social network. In: *The New England Journal of Medicine.* May 2008. Vol. 358. pp. 2249-58.

Cleland, A.A. & Pickering, M.J. (2003). The use of lexical and syntactical information in language production: Evidence from the priming of noun-phrase structure. In: *Journal of Memory and Language* 49. pp.214-230. http://www.psy.ed.ac.uk/people/martinp/pdf/cleland-pickering-jml03.pdf (last accessed 2/7/2010)

Coates, Jennifer (1986). *Women, Men and Language*. London: Longman.

Coffin, C; Hewings, A. & O'Halloran, K. (2004). *Applying English Grammar. Functional and Corpus Approaches.* London: Arnold Publishers (Hodder Headline).

Collins, Allan, M. & Quillian, Ross M. (1969). Retrieval Time from Semantic Memory. In: *Journal of Verbal Learning and Verbal Behaviour* 8. pp.240-248.

Collins, A. M., & Quillian, M. R. (1970). Facilitating retrieval from semantic memory: The effect of repeating part of an inference. *Acta Psychologica, 33,* 304-314.

Collins, A. M., & Quillian, M. R. (1972a). Experiments on semantic memory and language comprehension. In L. W. Gregg (Ed.), *Cognition in learning and memory.* New York: Wiley.

Collins, A. M., & Quillian, M. R. (1972b). How to make a language user. In E. Tulving & W. Donaldson (Eds.), *Organisation of memory.* New York: Academic Press. pp. 319-322.

Collins, A.M. & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. In: *Psychological Review.* Vol. 82, No. 6, pp.407-428

Collins Cobuild – English Grammar. J. SINCLAIR [Editor-in-Chief] (1990). Collins COBUILD English Grammar. London: Collins.

Cullen, Andrew: Scouse. A Comedy of Terrors (unpublished script). Commissioned by and first performed at the Everyman Theatre, Liverpool, 12th of February 1997. Directed by Peter Rowe.

de Beaugrande, R.: Text linguistics at the millennium: Corpus data and missing links. http://www.proz.com/translation-articles/articles/50/1/Corpus-Linguistics:-Meaning-in-Context/print/50 . Accessed 07/11/2007

de Groot, A.M.B., Thomassen, A.J.W.M., & Hudson, P.T.W. (1982). Associative facilitation of word recognition as measured from a neutral prime. In: *Memory and Cognition*, 10, pp. 358 - 370.

de Groot, A.M.B. (1989) Representational aspects of word imageability and word frequency as assessed through word association. In: *Journal of Experimental Psychology: Learning, Memory and Cognition,* 15, pp. 824 - 845.

de Mornay Davies, Paul (1998). Automatic semantic priming: The contribution of lexical- and semantic-level processes. In: *European Journal of Cognitive Psychology. 10* (4), pp. 389 – 412.

Desmet, T. & Declercq, M. (2006). Cross-linguistic priming of syntactic hierarchical configuration information. In: *Journal of Memory and Language*. Vol. 54, Issue 4. pp. 610-632.

Dickerson, Paul (2000). `But I'm different to them': Constructing contrasts between self and others in talk-in-interaction. In: *British Journal of Social Psychology* 39, pp. 381-398

Duguid, Alison (2009). Newspaper discourse – starting with salience: a diachronic comparison (forthcoming)

Ellis, N; Frey, E and Jalkanen, I. (2006a). The Psycholinguistic Reality of Collocation and Semantic Prosody. Presentation. Exploring the Lexis-Grammar Interface. Hanover, Germany 5-7th October

Ellis, N; Simpson-Vlach, R and Maynard, C. (2006b). The Processing of Formulas in Native and L2 speakers Psycholinguistic and Corpus Determinants. Presentation. Exploring the Lexis-Grammar Interface Hanover, Germany 5-7th October

Ellis, N; Frey, E (2007). The Psycholinguistic Reality of Collocation and Semantic Prosody (2): Affective Priming. Paper submitted for the Proceedings of the Symposium on Formulaic Language. University of Wisconsin-Milwaukee, March 18-21st.

Ellis, N; Frey, E and Jalkanen, I.: The Psycholinguistic Reality of Collocation and Semantic Prosody (1): Lexical Access. In: U. Römer & R. Schulze, (Eds.)
Exploring the Lexis-Grammar Interface. Studies in Corpus Linguistics. Amsterdam: John Benjamins. (in press).

Ellis, Stanley (1974): The Survey of English Dialects and Social History. http://www.jstor.org/pss/40178388. Accessed 10/3/2010.

Fail, Lia: Corpus Linguistics: Meaning in Context. http://www.proz.com/translation-articles/articles/50/1/Corpus-Linguistics%3A-Meaning-in-Context. Published on 09/20/2004. Accessed 07/11/2007.

Fairon, C; Singler, J.V. (2006). I'm like, "Hey it works!": Using GlossaNet to find attestations of the quotative (be) like in English –Language Newspapers In: *Language and Computers*, Vol. 55, No. 1, pp. 325-336.

Fasulo, A & Zucchermaglio, C. (2002). My selves and I: identity marker in work meeting talk. In: *Journal of Pragmatics* 34, pp. 1119 – 1144.

Fernandez, Paula Rubio (2007). Suppression in metaphor interpretation: differences between meaning selection and meaning construction. In: *Journal of Semantics* 24, June 16, pp. 345–371

Firth, J.R. (1957). *Papers in Linguistics 1934 – 1951*. London: Oxford University Press.

Firth, J.R. (1964). *The Tongues of Men & Speech*. London: Oxford University Press.

Firth, J.R. (1968). *Selected Papers 1952 – 1959*. London: Longmans, Green & Co.

Fowler, J.H. & Christakis, N.A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis of the Framingham Heart Study social network. In: *British Journal of Medicine.* Vol. 337, pp. 23-38

Fox Tree, J.E. & Schrock, J. C. (2002). Basic meanings of *You know* and *I mean*. In: *Journal of Pragmatics* 34, pp 727–747.

Francis, G; Sinclair, J. (1994). "I bet he drinks Carling Black Label" In: *Applied Linguistics.* Vol. 5 No. 2, pp. 190-200.

Fraser, B. (1999). What are discourse markers? In: *Journal of Pragmatics* 31, pp. 931-952.

Gagné, C.L. (2001). Relation and Lexical Priming During the Interpretation of Noun-Noun Combinations. In: *Journal of Experimental Psychology.* Vol. 27, No. 1, pp. 236-54.

Garretson, Gregory (2007). Book review *Lexical Priming* In: *International Journal of Corpus Linguistics* 12:3, pp.445–452.

Gaskins, Richard (2005). Network dynamics in saga and society In: *Scandinavian Studies.* Lawrence: Vol. 77, Iss. 2. http://findarticles.com/p/articles/mi_hb275/is_2_77/ai_n29196060/ (accessed 2/7/2010)

Gobet, F.; Lane, PCR.; Croker, S.; Cheng Peter; Jones, G.; Oliver, I. & Pine, J.M.: (2001). Chunking mechanisms in human learning. In: *Trends in Cognitive Sciences.* Vol.5 No.6 June, pp. 236-243.

Goede, Dieuwke de (2006). Verbs in Spoken Sentence Processing: Unravelling the Activation Pattern of the Matrix Verb Pattern of the Matrix Verb. PhD Thesis, University of Groningen: http://irs.ub.rug.nl/ppn/298832666 2006. Accessed 10/2007

Granovetter, Mark S. (1973). The Strength of Weak Ties. In: *American Journal of Sociology* 78, pp.1360-80.

Greenbaum, S. 1988. Good English and the Grammarian. London: Longman.

Greenbaum, Sidney & Whitcut, Janet ([1988] 1991). *Longman Guide to English Usage*. London: Longman.

Grey, C. and Sturdy, A. (2007). Friendship and Organizational Analysis: Toward a Research Agenda. In: *Journal of Management Inquiry*, 16; pp. 157-172

Gries, Stefan Th. (2005). Syntactic Priming. A Corpus-based Approach. In: *Journal of Psycholinguistic Research.* Vol. 34, No. 4, pp. 365-399.

Gries, Stefan Th. (2009). *Quantitative Corpus Linguistics with R*. New York & London: Routledge

Gumperz, J. (1982). *Discourse Strategies*. London: CUP

Habib, Reza (2001). On the relation between conceptual priming, neural priming and novelty assessment. In: Scandinavian Journal of Psychology. Vol 42. pp. 187-195.

Halliday, M.A.K. (1959). *Language of the Chinese "Secret History of the Mongols"* London: Longman.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M.A.K. (1991). Corpus Studies and probabilistic grammar. In: Aijmer, Karin & Altenberg, Bengt: *English Corpus Linguistics.* London: Longman, pp. 8-29.

Halliday, M.A.K. (2004). The Spoken Language Corpus: a foundation for grammatical theory. In: *Language and Computers.* Vol. 49, no. 1, pp. 11-38.

Hamer, Andrew (1995). Public Lecture. Held in the Baltic Fleet Pub, The Strand, Liverpool.

Hamer, Andrew (1995. Revised 2009). Principal Distinguishing Features of the Scouse Accent. Lecture Hand-out. 'Sound to Speech' Module. University of Liverpool.

Hamer, Andrew (2007). English on the Isle of Man. In: Britain, David: *Language in the British Isles*. Cambridge: CUP, pp.171-175.

Hanks, W. F. (1990). Referential practise. Language, and lived space amongst the Maya. Chicago: University of Chicago Press.

Healy, A. I. & Miller, G. A. (1970). The verb as determinant of sentence meaning. In: *Psychonomic Science, 20,* 372.

Hernamdez, A, Fennema-Notestine, C., Urdell, C., Bates, E. (2001). Lexical and sentential priming in competition: Implications for two-stage theories of lexical access. In: *Applied Psycholinguistics* 22, pp. 191-215.

Hoey, Michael (1991). *Patterns of Lexis in Text.* Oxford: Oxford University Press.

Hoey, Michael (1993). A common signal in discourse: How the word reason is used in texts. In: Sinclair, J.M., Hoey, M.& Fox, G.: *Techniques of description. Spoken and written discourse. A festschrift for Malcolm Coulthard.* London: Routledge, pp. 67-82.

Hoey, Michael (1995). The Lexical Nature of Intertextuality: A Preliminary Study. In: Warvik, B., Tanskanen, S-K, and Hiltunen, R. (eds). *Organization in Discourse. Proceedings from the Turku Conference 1995*. Anglicana Turkuensia 14. pp. 73-94.

Hoey, Michael (1997). From Concordance to Text Structure: New Uses for Computer Corpora. In: Melia, J. & Lewandoska, B. (eds) *Proceedings of PALC 97*. Lodz: Lodz University Press.

Hoey, Michael (2001). *Textual interaction*. *An introduction to written discourse analysis.* London: Routledge.

Hoey, Michael (2001). A world beyond collocation: new perspectives on vocabulary teaching. In: Lewis, Michael: *Teaching Collocations.* Hove: Language Teaching Publications, pp. 224-243.

Hoey, Michael (2003a). Textual Colligation: A special kind of Lexical Priming.
http://www.lexical priming.org (Accessed 09/09/2007)

Hoey, Michael (2003b). Lexical Priming and the Properties of Text.
http://www.monabaker.com/tsresources/Lexical primingandthePropertiesofText.htm (Accessed 09/05/2006)

Hoey, Michael (2003c). Why Grammar is beyond Belief.
In: University of Liège. *Belgian Association of Anglicists in Higher Education*: Belgian essays on language and literature pp. 183-196.

Hoey, Michael (2004). The Textual Priming of Lexis. In: Bernardini, S & Stewart, D: *Corpora and Language Learners*. Amsterdam: John Benjamins Publishing Co., pp. 21-41.

Hoey, Michael (2005). *Lexical Priming.* A new theory of words and language. London: Routledge.

Hoey, Michael (2008a). Lexical priming and literary creativity. In: Hoey, M; Mahlberg, M; Stubbs, M & Teubert, W.: *Text, Discourse and Corpora*. London: Continuum, pp. 7-30.

Hoey, Michael (2008b). Grammatical creativity: a corpus perspective. In: Hoey, M; Mahlberg, M; Stubbs, M & Teubert, W.: *Text, Discourse and Corpora*. London: Continuum, pp. 31-56.

Hornby, A.S. (1954). *A Guide to Patterns and Usage in English*. London: OUP.

Honeybone, P. (2007). New-Dialect Formation in 19th Century Liverpool: A Brief History of Scouse. In: Grant, A., Grey, C. & Watson, K. (eds) *The Mersey Sound: Liverpool's Language, People and Places*. Liverpool: Open House Press.

Hudson, R.A. (1980). *Sociolinguistics.* Cambridge: CUP.

Hunston, Susan & Francis, Gill. (1999). *Pattern Grammar.* A corpus-driven approach to the lexical grammar of English. Amsterdam: John Benjamins Publishing Co.

Hunston, Susan (2001). Colligation, lexis, pattern and text. In: Scott & Thompson (eds.) *Patterns in Text: in honour of Michael Hoey.* Amsterdam: John Benjamins Publishing Co., pp. 13-33.

Hunston, Susan (2002). *Corpora in Applied Linguistics.* Cambridge: CUP.

Hunston, Susan (2007). Semantic prosody revisited. In: *International Journal of Corpus Linguistics* 12:2 . pp 249–268.

Hymes, Dell (1971). *On communicative competence*. Philadelphia: University of Pennsylvania Press.

Israel, Michael (2002). Literally Speaking. In: *Journal of Pragmatics* 34, pp. 423–432.

Ito, Rika & Tagliamonte, Sali (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. In: *Language in Society* 32, pp. 257-279.

Jacques, Brian (1981). *Brian Jacques meets Paddy Kelly.* Stories from the BBC Radio Merseyside Series. Liverpool: Raven Books (Anvil Press).

Juncker, A.H. (1993). The discourse-marker *well*: A relevance-theoretical account. In: *Journal of Pragmatics* 19, pp. 435-453.

Kashima, E. & Kashima, Y. (1998). Culture and Language: The Case of Cultural Dimensions and Personal Pronoun Use. In: *Journal of Cross-Cultural Psychology* 29, pp. 461-488.

Klein, B., Cosmides, L., Tooby, J. & Chance, S. (2001). Priming Exceptions: A test of the scope hypothesis in naturalistic trait judgements. In: *Social Cognition*, Vol. 19, No. 4, pp. 443-468.

König, Ekkehard and Siemund, Peter (2000). The Development of Complex Reflexives and Intensifiers in English. In: *Diachronica* XVII:1, pp. 39–84.

Knowles, Gerald O. (1973). Scouse: The Urban Dialect of Liverpool. PhD Thesis, unpublished. University of Leeds

Labov, W. 1966(a). *The social stratification of English in New York City.* Washington DC: Center for Applied Linguistics.

Labov, W. (1966b). The Effect of Social Mobility on Linguistic Behavior. In: *Sociological Inquiry*. Vol. 36, Issue 2 04/1966, pp. 186-203.

Labov, W. (1972). *Language in the Inner City*. Oxford: Basil Blackwell.

Lamb, Sydney (2000) *langbrain*. Language and Brain: Neurocognitive Linguistics. (2000-2006 Rice University). Last accessed 09/10. http://www.ruf.rice.edu/~lngbrain/

Langendoen, Terence F. (1971). A review of Selected Papers of J. R. Firth, 1952-59 by F. L. Palmer; J. R. Firth. In: *Language*, Vol. 47, No. 1. (Mar. 1971), pp. 180-181.

Ledoux, Kerry, Camblin, C, Swaab, and Gordon, P.C. (2006). Reading Words in Discourse: The Modulation of Lexical Priming Effects by Message- Level Context. In: *Behavioral and Cognitive Neuroscience Reviews*, Vol. 5, No. 3, pp. 107-127.

Lee, Sarah & Ziegeler, Debra (2006). Analysing a semantic corpus study across English dialects: Searching for paradigmatic parallels. In: Language and Computers, Vol. 56, No. 1. pp. 121-139. http://www.citeulike.org/article/459697 last accessed 12/08/09.

Leech, G. & Svartvik, J. (1975). *A communicative Grammar of English.* London: Longman.

Lewis, Michael (2001). Language in the lexical approach. In: Michael Lewis: *Teaching Collocations*. Hove: Language Teaching Publications, pp. 126-154.

Loftus, E. F. (1973). Activation of semantic memory. In: *American Journal of Psychology,* 86, pp. 331-337.

Loewenberg, Ina (1982). Labels and hedges: the metalinguistic turn. In: *Language and Style* 15 (3), pp. 193–207.

Longacre, Robert E. (1970). Sentence structure as a statement Calculus In: *Language*, Vol. 46, No. 4. (Dec., 1970), pp. 783-815.

Love, Tracy; Swinney, D; Walenski, M & Zurif, E. (2008). How left inferior frontal cortex participates in syntactic processing: Evidence from aphasia. In: *Brain and Language 107*, pp. 203-219.

Louw, B. (1993). Irony in the Text or Insincerity in the Writer?
The Diagnostic Potential of Semantic Prosodies. In: Baker, M.; Francis, G.; and Tognini-Bonelli, E. (eds) Text and Technology.
Amsterdam: Benjamins. pp. 157–76.

Macaulay, R. (2002). You know it depends. In: *Journal of Pragmatics* 34. pp. 749–767.

Macaulay, R. K. S. (2005). Talk that Counts. Age, Gender, and Social Class Differences in Discourse. New York: Oxford University Press.

Mair, Christian (2006). Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. In: *Language and Computers*, Vol. 55, No. 1, pp.355-376.

Mair, Christian (2009). Corpus Linguistics meets sociolinguistics: the role of corpus evidence in the study of sociolinguistic variation and change. In: Renouf, A. & Kehoe, A.: *Corpus Linguistics: Refinements and Reassessments*. Amsterdam & New York: Rodopi, pp. 7-32.

Mahlberg, Michaela (2005). *English General Nouns. A corpus theoretical approach*. Amsterdam / Philadelphia: John Benjamins.

Mahlberg, Michaela (2006). *but it will take time …*points of view on a lexical grammar of English In: *Language and Computers*, Vol. 55, No. 1, pp. 377-390.

Mahlberg, Michaela (2008). Corpus stylistics: bridging the gap between linguistic and literary studies. In: Hoey, M; Mahlberg, M; Stubbs, M & Teubert, W.: *Text, Discourse and Corpora*. London: Continuum, pp. 219-246.

Marcinkeviciene, Ruta: Parallel corpora and bilingual lexicography. http://donelaitis.vdu.lt./publikacijos/vilniausk.htm (last accessed 08/09).

Marconi, Diego (1997). *Lexical Competence*. Cambridge, Mass.: The MIT Press.

McCarthy, Michael (1993). Spoken discourse markers in written text. In: Sinclair, Hoey & Fox: *Techniques of Description.* Spoken and written discourse. London: Routledge, pp.170-182.

McCarthy M. & Carter, R. (2004). "There's millions of them'': hyperbole in everyday conversation. In: Journal of Pragmatics 36. pp. 149-184.

McEnery, Anthony and Xiao, Zhonghua (2005). HELP or HELP to: What Do Corpora Have to Say? English Studies, 86:2, pp. 161 — 187.

Melinger, A. & Dobel, C. (2005). Lexically-driven syntactic priming. In: *Cognition* 98, Issue 1, pp.B11-B20.

Meyer, Charles F. (2004). Can you really study language variation in linguistic corpora? In: *American Speech* 79.4 pp. 339-355.

Meyer, Charles F. (2005). Response to Newmeyer's 'Grammar is grammar and usage is usage' In: *Language.* Vol. 81, 228 Number 1, pp. 226-228.

Meyer, D.E. & Schvanefeldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. In: *Journal of Experimental Psychology*. Vol. 90; No. 2, pp. 227-234.

Meyer, D.E. & Schvanefeldt, R.W. (1976). Meaning, memory structure and mental processes. In: *Science*. Vol. 192. 2nd April, pp. 27-33.

Meyer, D.E. & Schvanefeldt, R.W. (1984). Discussion of Meyer, D.E. & Schvanefeldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations (1971) In: *Citation Classic* 47, November 19.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review.* 63, pp. 81–97.

Miller, J & Weinert, R.(1995). The function of LIKE in dialogue. In: *Journal of Pragmatics* 23, pp. 365-393.

Millar, Neil (2009). The processing demands of learner collocational deviations: Evidence from self-paced reading. ICAME 30 presentation. Slides available at:
http://www.lancs.ac.uk/postgrad/millarn/Files/ICAME30_slides.pdf
(last accessed 28/07/09)

Milroy, L. (1980). *Language and Social Networks*. Oxford: Blackwell.

Morley, Barry: (2006). WebCorp: A tool for online linguistic information retrieval and analysis. In: *Language and Computers*, Vol. 55, No. 1, pp. 283-296.

Mukherjee, J. & Hoffmann, S. (2006). Describing verb-complementational profiles of New Englishes. In: *English World-Wide,* Vol. 27, No. 2, pp. 147-171

Murphy, G.L.: (1988). Comprehending Complex Problems. In: *Cognitive Science* 12, pp. 529-562.

Neely, J.H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. In: *Memory & Cognition*, Vol. 4, No. 5, pp. 648-654.

Neely, J.H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. In: *Journal of Experimental Psychology*. General, 106, 226 - 254.

Nelson, M. (2000). The Methodological Background: British Traditions of Text Analysis, Correlative Register Analysis and Corpus Linguistics. http://users.utu.fi/micnel/thesis/Chapter5%20.html. Accessed 07/11/2007

Novick, J., Kim, A., Trueswell, J. (2003). Studying the grammatical aspects of word recognition: Lexical Priming, Parsing and Syntactic Ambiguity Resolution. In: *Journal of Psycholinguistic Research*. Vol. 32, No. 1, pp. 57-75.

O'Donnell, Matthew Brook (2009). The *Adjusted Frequency* List: Evaluating a method for producing cluster-sensitive frequency counts. Presentation, AACL Edmonton, Canada – 10 October 2009.

O'Keefe, A., McCarthy, M., Carter, R. (2007). *From Corpus to Classroom.* Cambridge: Cambridge University Press.

ONS – Office of National Statistics. Various – http://www.ONS.gov.uk (last accessed 07/09).

Orwell, George (1944). Propaganda and Demotic Speech. In: *George Orwell: Complete Writings* (London, 1980), Secker & Warburg.

Otten, Marte & Van Berkum, Jos J. A.(2008) Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? In: *Discourse Processes*, 45:6, pp. 464-496. http://dx.doi.org/10.1080/01638530802356463 last accessed 29/01/2009

Pace-Sigge, Michael (2002). /P/ /t/ /k/ Stop Sounds Lenition in Liverpool English. Unpub. MA Thesis. University of Liverpool

Paradis, Carita (1997). *Degree modifiers of adjectives in spoken British English*. Lund Studies in English 92. Lund: Lund University Press.

Paradis, Carita (1998). Well weird. Degree modifiers of adjectives revisited: The nineties. www.vxu.se/hum/publ/cpa/well_weird.pdf Accessed at 11.06.10

Paradis, Carita (2003). Between epistemic modality and degree: the case of *really.* In: *Modality in contemporary English.* Eds. R. Facchinetti, F. Palmer & M. Krug. Berlin: Mouton de Gruyter, pp. 197-220.
@ http://www.vxu.se/hum/publ/cpa/Really.pdf .Accessed 24.10.06

Partington, Alan (1993). Corpus evidence of language change: The case of the intensifier. In: Baker, M. Francis, G. & Tognini Bonelli, E. (eds.), *Text and technology: In honour of John Sinclair*, Amsterdam & Philadelphia: John Benjamins, 177–92.

Partington, Alan (1998). *Patterns and Meanings*. Using Corpora for English Language Research and Teaching. Amsterdam: John Benjamins Publishing Co.

Partington, Alan (2003). *The Linguistics of Political Argument*. The spin-doctor and the wolf-pack at the White House. London: Routledge.

Patrick, Peter L. (2002). The Speech Community. In: Chambers, Trudgill & Schilling-Estes (eds.), *Handbook of language variation and change.* Oxford: Blackwell.

Paul, S.T. & Kellas, G. (2004). A time course view of sentence priming effects. In: *Journal of Psycholinguistic Research.* Vol. 33, No. 5, Sept, pp. 383-405

Pawley, Andrew & Hodgetts Syder, Frances ([1983] 1996) Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, J & Schmidt, R.W.: *Language and Communication.* London: Longman Ltd., pp.192-228.

Peters, Ann M. (1983). *The Units of Language Acquisition.* Cambridge: CUP.

Philip, Gill: Book review of Hoey, Michael: *Lexical Priming*. To appear in Language Awareness Journal. http://amsacta.cib.unibo.it 2007 (accessed 03/03/2008)

Posner, M. L, & Snyder, C. R. .R. (1975a). Facilitation and inhibition in the processing of signals. In Rabbitt, P. M. A. & S. Domic (Eds.), *Attention and performance V.* New York: Academic Press.

Posner, M. L, & Snyder, C. R. R. (1975b). Attention and cognitive control. In: . L. Solso (Ed.), *Information processing and cognition: The Loyola symposium.* Hillsdale, NJ.: Erlbaum.

Prucha, Jan (1972). Psycholinguistics and Sociolinguistics – Separated or Integrated. In: *Linguistics* Vol.:89. issue: 89, pp. 9-23.

Quillian, Ross M. (1961). The elements of human meaning: a design for and understanding machine. In: *Communications of the ACM*. Volume 4, Issue 9 (September 1961) Page: 406

Quillian, R.M. (1962). A revised design for an understanding machine. *Mechanical Translation, 7,* pp. 17-29.

Quillian, R.M. (1966). *Semantic memory.* Unpublished doctoral dissertation, Carnegie Institute of Technology, (Reprinted in part in M. Minsky [Ed.], *Semantic information processing.* Cambridge, Mass.: M.I.T. Press, 1968.)

Quillian, R.M**.** (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12,* pp. 410-430.

Quillian, R.M. (1969). The Teachable Language Comprehender: A Simulation Program and Theory of Language. In: *Computational Linguistics.* Vol. 12, No. 8, August, pp. 459-476.

Quirk, Randolph; Sidney Greenbaum; Geoffrey Leech & Jan Svartvik (1985). *A comprehensive Grammar of the English Language.* Harlow, Essex: Longman.

Rayson, Paul: *Log-likelihood calculator*. (Last accessed on 1/10/10) http://ucrel.lancs.ac.uk/llwizard.html

Rayson, P., Berridge, D. and Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium, March 10-12, 2004*, Presses universitaires de Louvain, pp. 926 - 936.

Ratcliff, R. & McKoon, G. (1988). A retrieval theory of priming in memory. In: *Psychological Review*, 95, pp. 385 - 408.

Renouf, A & Sinclair, J.M. (1991). Collocational frameworks in English. In: Aijmer, Karin and Altenberg, Bengt: *English Corpus Linguistics.* London: Longman, pp.128-143.

Rundell, Michael (Editor-in-Chief); *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

Salamoura, A. & Williams, J. (2006). Lexical activation of cross-language syntactic priming. In: *Bilinguism: Language and Cognition* 9 (3), pp.299-307.

Saller, Harald (2004). Zugriff auf Wissen, Zugang zum Sinn. Anmerkungen zu Texten, Kommentaren und sematischen Netzen. In: *PhiN-Beiheft* 2/2004, pp.66-82.

Saraceni, Mario (2002). Review of Patterns of Text: In Honour of Michael Hoey In: *LINGUIST List* 13.230 http://www.linguistlist.org/issues/13/13-230.html (last accessed 08/09)

Schiffrin, Deborah (1987). *Discourse Markers*. Cambridge: Cambridge University Press.

Schmied, Josef (2006). New Ways of analysing ESL on the WWW with WebCorp and WebPhraseCount. In: *Language and Computers*, Vol. 55, No. 1, pp. 309-324.

Schourup, Lawrence C. (1985). *Common Discourse Particles in English Conversation*. New York & London: Garland Publishers.

Schourup, Lawrence (1999). Discourse Markers. In: *Lingua* 107, pp 227-265.

Schourup, Lawrence (2001). Rethinking *well. Journal of Pragmatics* 33, pp. 1025-1060.

Scott, Michael (since 1996). WordSmith. Lexical analysis software for the PC. Published by Oxford University Press (now at version 5.0). http://www.lexically.net/wordsmith/index.html (last accessed 02/10)

Seidl, J & McMordie, W. (1988). *English Idioms*. Oxford: Oxford University Press.

Sharoff, S. (2006). How to handle lexical semantics in SFL: a corpus study of purposes for using size adjectives. In: Hunston, S. & Thompson, G. (eds.) *Systemic Linguistics and Corpus*. London: Equinox, pp. 184-205. http://corpus.leeds.ac.uk/serge/publications/size-adjectives-equinox.pdf (last accessed 08/09)

Sherman, S.J.; Crawford, M.T.; Hamilton D.L. & Garcia-Marques, L. (2003). Social Inference and Social Memory: The Interplay between Systems. In: Hogg, M. & Cooper, J. (eds.): *The SAGE handbook of social psychology*. London: SAGE Publications Ltd. pp. 45-68.

Sinclair, J. McH & Coulthard, R.M (1975). *Towards an Analysis of Discourse. The English used by teachers and pupils.* Oxford: Oxford University Press.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, John (1997). The Lexical Item. In: Weigand. E. (ed). *Contrastive Lexical Semantics*. Amsterdam: Benjamins, pp. 1-24.

Sinclair, John (Editor-in-Chief) et al. (1998a). *Collins Cobuild Grammar Patterns*. Vol. 1: *Verbs.* London: Collins.
(1998b). Vol. 2: *Nouns and Adjectives.* London: Collins.

Sinclair, John. (2000). Lexical grammar. In: Naujoji Metologija. Vol. 24, pp. 191 – 203.
http://donelaitis.vdu.lt/publikacijos/sinclair.pdf (Accessed 13/07/09)

Sinclair, John (2004). *Trust the text. Language, corpus and discourse.* London: Routledge.

Strangert, Eva. (2004). On modelling of conversational speech. In*: Proceedings, FONETIK 2004*, Dept. of Linguistics, Stockholm University.

Streeck, Juergen (2002). Grammars, Words and Embodied Meanings: On the Uses and Evolution of *So* and *Like*. In: *Journal of Communication*, Volume 52 Issue 3, pp.581-596.

Stewart, Dominic (2010). *Semantic Prosody*. A Critical Evaluation. New York / Abingdon: Routledge.

Stubbs, Michael (1983). *Discourse Analysis.* The Sociolinguistic Analysis of Natural Language. Oxford: Basil Blackwell.

Stubbs, Michael (1995). Collocations and Cultural Connotations of Common Words. In: Linguistics and Education 7, 379-390

Stubbs, Michael (1996). *Text and Corpus Analysis.* Computer-Assisted Analysis of Language and Culture. Oxford: Basil Blackwell.

Stubbs, Michael (2001a). On inference theories and code theories: Corpus evidence for semantic schemas. In: *Text* 21(3) pp. 437-465.

Stubbs, Michael (2001b). *Words and Phrases*. Oxford: Basil Blackwell.

Stubbs, Michael (2006). Corpus Analysis: the state of the art and three types of unanswered questions. In: Thompson, G. & Hunston, S.: System and Corpus. London and Oakville: Equinox, pp.15-36.

Stubbs, Michael (2008a). Quantitative data on multi-word sequences in English: the case of word *world*. In: Hoey, M; Mahlberg, M; Stubbs, M & Teubert, W.: *Text, Discourse and Corpora*. London: Continuum, pp. 163-190.

Stubbs, Michael (2008b). On texts, corpora and models of language. In: Hoey, M; Mahlberg, M; Stubbs, M & Teubert, W.: *Text, Discourse and Corpora*. London: Continuum, pp. 127-162.

Tagliamonte, S. (2005). So who? Like how? Just what? Discourse markers in the conversations of Young Canadians. In: Journal of Pragmatics 37, pp.1896–1915.

Tannen, Deborah (1982). The Oral/Literate continuum in discourse. In: Tannen, D. (Ed.): *Spoken and written language: Exploring Orality and Literacy*. New Jersey: AblexPublishing Corp., pp. 1-16.

Tao, Hongyin (2003). Turn initiators in spoken English: A corpus-based approach to interaction and grammar. In: Pepi Leistyna & Charles F. Meyer (Eds): Language and Computers, Volume 46. (entitled *Corpus Analysis: Language Structure and Language Use*), pp. 187-207.

Thomas, Beth (1988). Differences of sex and sects: linguistic variation and social networks in a Welsh mining village. In: Coates, J. & Cameron, D.(eds) *Women in their speech communities*. New York: Longman, pp. 51-60.

Thompson, Geoff (1996). *Introducing Functional Grammar.* London: Edward Arnold.

Thompson, Geoff & Susan (2001) Patterns of cohesion in spoken text. In: Scott & Thompson (eds.) *Patterns in Text: In Honour of Michael Hoey.* Amsterdam: John Benjamins Publishing Co. pp.55-81.

E. B. Titchener (1922). A Note on Wundt's Doctrine of Creative Synthesis In: *The American Journal of Psychology*, Vol. 33, No. 3, pp. 351-360.

Traxler, M., Foss, D., Seely, R., Kaup, B. & Morris, R.K. (2000). Priming in sentence processing: Intralexical spreading activation, schemas and situation models. In: *Journal of Psycholinguistic Research*. Vol. 29, No.6, pp. 581-595.

Trofimovich, Pavel (2005). Spoken-word processing in native and second languages: An investigation of auditory word priming. In: *Applied Psycholinguistics* 26, pp. 479-504.

Trudgill, Peter (1974). *The Social Differentiation of English*. Cambridge: CUP.

Trudgill, Peter (2000). *The Dialects of England*. Oxford: Blackwell.

Tsiamita, Fanie. 2009. Polysemy and lexical priming: The case of drive. In: Römer, Ute and Schulze, Rainer (eds.) *Exploring the Lexis–Grammar Interface.* Amsterdam: John Benjamins, pp. 247–264.

Tsujimura, Natsuko (2001). Degree words and scalar structure in Japanese In: *Elsevier Lingua* 111, pp. 29-52.

Ullman, M. & Hartshorne, J. (2006) Study Of Language Use In Children Suggests Sex Influences How Brain Processes Words. In: *Medical News Today* 11/06
http://www.medicalnewstoday.com/medicalnews.php?newsid=57633
(Last Accessed 13/10/2007)

Van Dijk, Teun A. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

Walter, Elisabeth (Senior Commissioning Editor) (2008). *Cambridge Advanced Learner's Dictionary, Third edition (CALD3).* Cambridge: CUP.

Watson, Kevin (2007). Is Scouse getting Scouser? Exploring phonological change in contemporary Liverpool English. In: Grant, A., Grey, C. & Watson, K. (eds.) *The Mersey Sound: Liverpool's Language, People and Places.* Liverpool: Open House Press.

Watts, R.J. (1989). Taking the Pitcher to the 'Well': Native speakers' perception of the use of Discourse markers in Conversation. In: *Journal of Pragmatics* 13, pp. 203-237.

Well, J.C. (1982) *Accents of English. An Introduction*. Cambridge: CUP.

Whitney, Paul (1998). *The Psychology of Language*. Boston / New York: Houghton Mifflin Company.

Wikberg, Kay (2007). A review of Michael Hoey's *Lexical Priming: A New Theory of Words and Language* (2005). In: *JLS* 36, pp. 89–101.

Williams, J.N. (1996). Is automatic priming semantic? In: *European Journal of Cognitive Psychology*, *8*, pp. 113-161.

Williams, Geoff (2006). A review of Michael Hoey's *Lexical Priming: A New Theory of Words and Language* (2005). In: *Oxford Review*. 27 August 2006. pp. 327-335

Winter, Eugene (1982). *Towards a Contextual Grammar of English.* London: Allen Unwin Ltd.

Whitsitt, Sam (2005) A critique of the concept of semantic prosody In: *International Journal of Corpus Linguistics* 10:3. pp.283-305.

Wolfram Alpha. Computational Knowledge Engine. At: http://www.wolframalpha.com/ (last accessed 10/09).

Wolfram, Walt (1978). Contrastive Linguistics and Social Lectology In: *Language Learning.* Vol. 28, No. 1. http://www.eric.ed.gov/PDFS/ED111202.pdf (accessed 2/07/2010).

Wong, May L-Y (2009). Expressions of gratitude in ICE-HK. Presentation, Corpus Linguistics 2009 (Liverpool).

Wray, Alison (2002a). *Formulaic Language and the Lexicon*. Cambridge University Press. Cambridge.

Wray, Alison (2002b). Formulaic Language in Computer-supported Communication: Theory Meets Reality. In: *Language Awareness,* Volume 11, Issue 2 October 2002 , pp. 114 - 131 http://www.*multilingual-matters.net/*la/011/0114/la0110114.pdf (accessed 1.11.2008)

Xioa, R. & McEnery, T. (2006). Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. In: *Applied Linguistics* 27/1, pp. 103–129.

Yurchak, Alexei (2005). *Everything was forever, until it was no more. The last Soviet generation.* Princeton, NJ.: Princeton University Press.

Zimmermann, H.H. (1972). Zur Konzeption der automatischen Lemmatisierung von Texten. In: *SFB 100 "Elektronische Sprachforschung": Aspekte d. automatischen Lemmatisierung.* Bericht 10-72. Linguistische Arbeiten 12, pp. 4-10.

# Appendices

## Appendix I (chapter 2.1.2)

Appendix I.1 Code of Informants & their socio-economic background.

| NAME /Category | occupation (now) | Occupational Class previous - now | living in | est. annual income ('03) | ONS Class | fam. Background where known |
|---|---|---|---|---|---|---|
| 1.Alf | pensioner | manual w | urban - low cost | < 10 k | L13 | single income casual - father |
| 2 Diane | pensioner | housewife | urban - low cost | < 10 k | L 13 | |
| 3 Lisa | nurse | | urban - low cost | < 20 k | L4 | |
| 4. Steve M | baggage handler | manual w | urban - low cost | < 20 k | L12 | |
| 5. Alistair | sixth-former | | urban - medium cost | n/a | L15 | father retired, mother housekeeper: state pension |
| 6. Mr C. | teacher (primary) | | n/a | ~ 20 k | L3 | |
| 7. Yasmin | @ school | | urban - low cost | n/a | <16 | |
| 8. Lillie | @ school | | urban - low cost | n/a | <16 | living in foster care |
| 9. Sophie | @ school | | urban - low cost | n/a | <16 | living in foster care |
| 10. Lauren | @ school | | urban - low cost | n/a | <16 | |
| 11. Daryl | @ school | | urban - low cost | n/a | <16 | |
| 12. Sarah | @ school | | urban - low cost | n/a | <16 | |
| 13. Joan | @ school | | urban - low cost | n/a | <16 | |
| 14. Chris A | @ school | | urban - low cost | n/a | <16 | |
| 15. Karole | museum attendant | | n/a | < 20 k | L13 | |
| 16. Paul | museum attendant | | n/a | < 20 k | L13 | |
| 17. Steve R. | museum attendant | | n/a | < 20 k | L13 | |
| 18. John | museum attendant | manual w | n/a | < 20 k | L13 | |
| 19. Pauline | museum attendant | | n/a | < 20 k | L13 | |
| 20. Alan | museum attendant | | n/a | < 20 k | L13 | |
| 21. Brian | museum attendant | manual w | n/a | < 20 k | L13 | |
| 22. Mick | student | | urban - low cost | < 10 k | L15 | single parent |
| 23. Simon | student | | n/a | < 10 k | L15 | single parent |
| 24. Joe | librarian | | n/a | < 10 k | L11 | |
| 25. Jane | student | | n/a | < 10 k | L15 | |
| 26. James | student | | urban - low cost | < 10 k | L15 | |
| 27.Dave W. | bookkeeper | | urban - low cost | < 20 k | L11 | father gas fitter / mother lower supervisory |
| 28. Dean | student / bank clerk | | n/a | < 10 k | L11 | |

| NAME /Category | occupation (now) | Occupational Class previous - now | living in | est. annual income ('03) | ONS Class | fam. Background where known |
|---|---|---|---|---|---|---|
| 29. Zoe | @ school | | urban - low cost | n/a | <16 | |
| 30. 1 (m) | n/a | | n/a | n/a | | |
| 31. 2 (f)/ Liz | homemaker | | n/a | < 10 k | | |
| 32. Ellie | sixth-former | | suburban - comfortable | n/a | L15 | father BBC presenter |
| 33. Lisa(2) | homemaker | | urban - low cost | < 10 k | L12 | |
| 34. Lacy | pensioner | | urban - low cost | < 10 k | L12 | |
| 35. Jan (f.) | homemaker | | urban - low cost | < 10 k | L13 | father council worker - retired. Mother homemaker |
| 36. Tony | animator | | urban - low cost | < 20 k | L 9 | father Mersey pilot skipper |
| 37. Melissa | @ school | | urban - low cost | n/a | < 16 | daughter of Steve M & Jan |
| 38.3(f) /Lorraine | homemaker | | n/a | < 10 k | | |
| 39. Billy | museum attendant | | n/a | < 20 k | L13 | |
| 40. Chris | museum attendant | | n/a | < 20 k | L13 | |
| 42. Claire | museum attendant | | n/a | < 20 k | L13 | |
| 43. Elaine | museum attendant | | n/a | < 20 k | L13 | |
| 44 Greg | museum attendant | | n/a | < 20 k | L13 | |
| 45. John D | museum attendant | | n/a | < 20 k | L13 | |
| 46. Peter | museum attendant | | n/a | < 20 k | L13 | |
| 47. Sheila | museum attendant | | n/a | < 20 k | L13 | |
| 48. Tammy | self-employed artist | | n/a | < 10 k | L9 | |
| 49. Tom | pensioner | manual w | n/a | < 10 k | L13 | parents born in late 1800s |
| 50. Tim | student | | n/a | < 10 k | L15 | |

**Appendix I.2**: Facts about the SCO Corpus informants. Line-by line break-down to be found on CD-ROM.

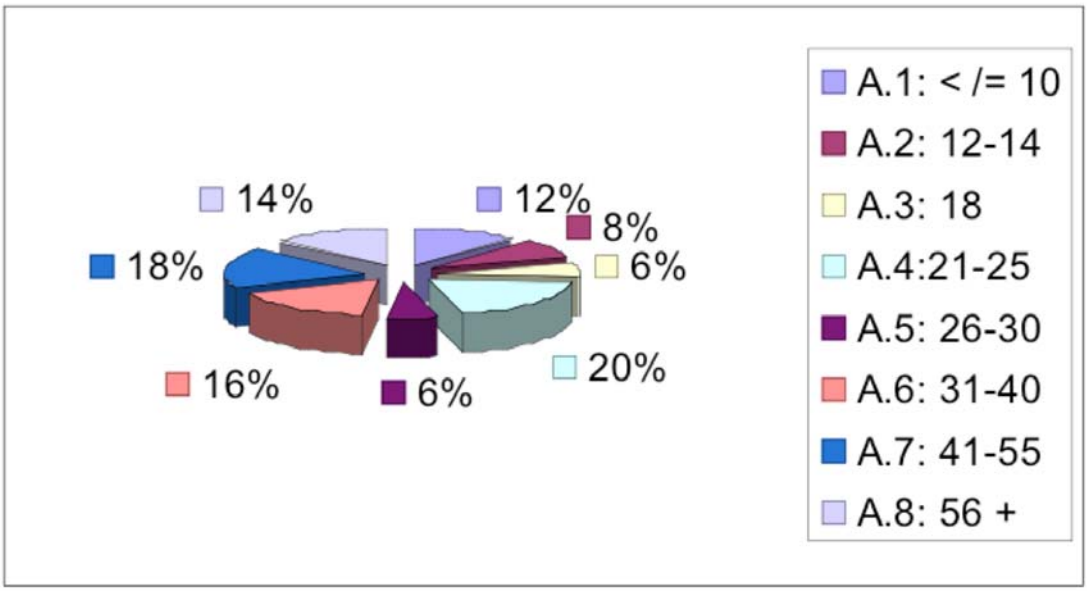



Figure 1: Age distribution in SCO

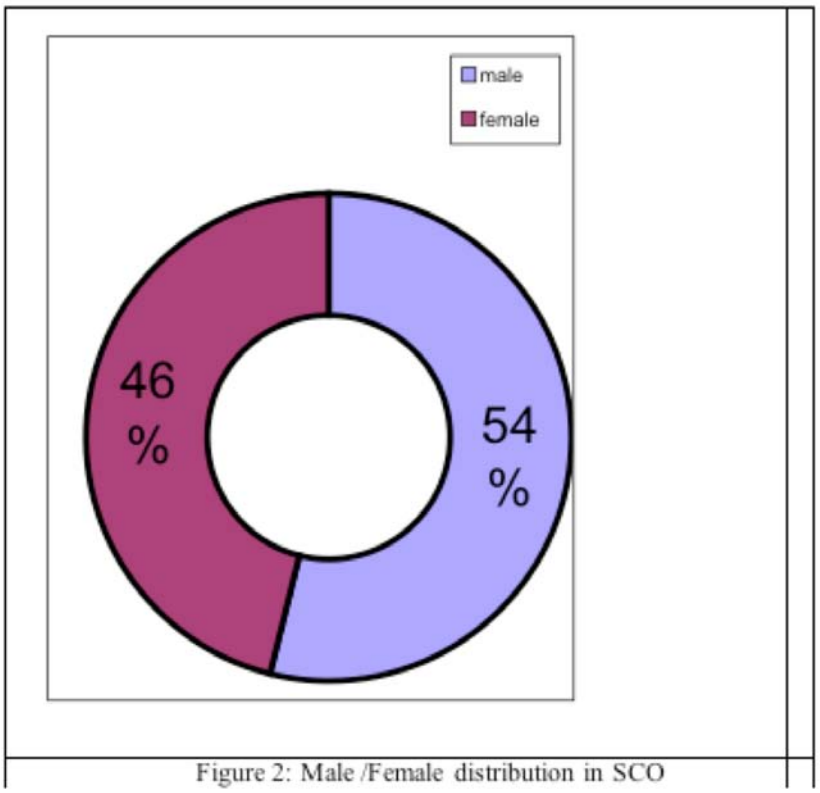



Figure 2: Male /Female distribution in SCO

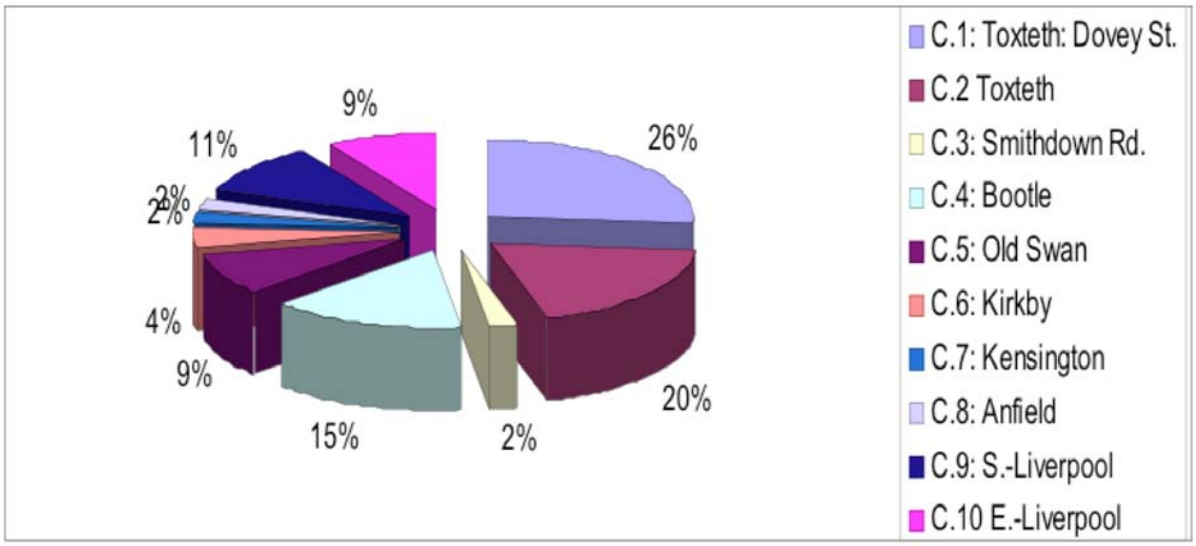



Figure 3: Informant habitation profile

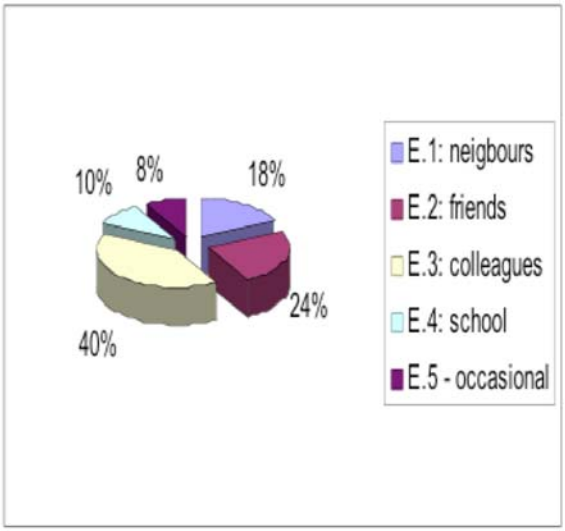



Figure 4: information where/ under which circumstances recording was undertaken

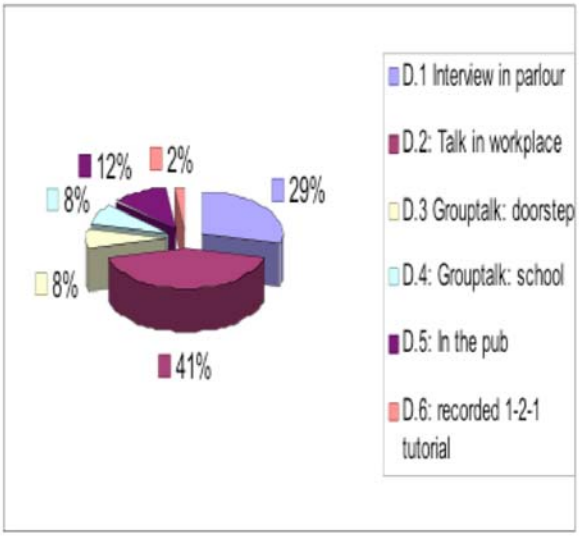



Figure 5: Relation between interviewee and interviewer1

# Appendix II (chapter 2.4)

## *(1) What is a* keyword *programme and what is it for?*

This is a program for identifying the "key" words in one or more texts. Key words are those whose frequency is unusually high in comparison with some norm.

Key-words provide a useful way to characterise a text or a genre. Potential applications include: language teaching, forensic linguistics, stylistics, content analysis, text retrieval.

The program compares two pre-existing word-lists, which must have been created using the *WordList tool*. One of these is assumed to be a large word-list which will act as a reference file. The other is the word-list based on one text which you want to study.

The aim is to find out which words characterise the text you're most interested in, which is automatically assumed to be the smaller of the two texts chosen. The larger will provide background data for reference comparison.

Key-words and links between them can be plotted, made into a database, and grouped according to their associates.

## *(2) How Key Words are Calculated*

The "key words" are calculated by comparing the frequency of each word in the wordlist of the text you're interested in with the frequency of the same word in the reference wordlist. All words which appear in the smaller list are considered, unless they are in a stop list.

If the occurs say, 5% of the time in the small wordlist and 6% of the time in the reference corpus, it will not turn out to be "key", though it may well be the most frequent word. If the text concerns the anatomy of spiders, it may well turn out that the names of the researchers, and the items spider, leg, eight, etc. may be more frequent than they would otherwise be in your reference corpus (unless your reference corpus only concerns spiders!)

To compute the "key-ness" of an item, the program therefore computes

- its frequency in the small wordlist
- the number of running words in the small wordlist
- its frequency in the reference corpus

- the number of running words in the reference corpus and cross-tabulates these.

Statistical tests include:

- the classic chi-square test of significance with Yates correction for a 2 X 2 table
- Ted Dunning's Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus.

A word will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlist.

Unusually infrequent key-words are called "negative key-words" and appear at the very end of your listing, in a different colour. Note that negative key-words will be omitted automatically from a keywords database and a plot.

Words which do not occur at all in the reference corpus are treated as if they occurred 5.0e-324 times (0.0000000 and loads more zeroes before a 5) in such a case. This number is so small as not to affect the calculation materially while not crashing the computer's processor.

(Mike Scott: WordSmith 4.0.: 2003)

## Appendix III

See: Michael Pace-Sigge: “A sociolinguistic justification for using a spoken Liverpool Corpus”
11th Warwick Postgraduate Conference; Wednesday, June 18, 2008.
URL of presentation (last accessed 21/09/2010):
http://www.scribd.com/doc/25428566/Sco-Socio-FIN

## Appendix IV (Chapter 3.2.1)

Palmer outlined his ... synthetic approach to the traditional parsing of sentences, terming this alternative *mechanism grammar* (or, later, *pattern-grammar*). In a development of his earlier London work on *ergonics* and substitution tables (see Howatt 1984: 236-9; Smith 1998a), and referring to materials already published for the Grammar and Structure Line of Approach of the Standard Course (1924d, 1925g), Palmer attempted to show how construction-patterns can be taught as a basis for (spoken and written) production, accompanying theoretical explanation and sample exercises with a patented This approach was later returned to in 1932t and in collaborative research with Hornby (1934aa), joining up at that point with

collocational considerations to lead ultimately to a classification of the most significant *sentence patterns* for learners of English as a foreign language (this achievement being realized, in particular, in Hornby et al. 1942 and Hornby 1954).
(Richard E. Smith:1999. P.121)

## Appendix V (Chapter 6.5.3.3)

*People like* appears divergent in its proportional frequency of use as well as its nesting.

1 Have you heard - eh - 322. people like 323. Liken Steven Gerrard to ..
2 Move communities 543. people like - th - thrown together -
3 Marilyn Manson is my mate (pause) We need more people like him (pause)
4 I don't know 59. L S'pose ... people like it don't they ... but ...
5 hard (inaud) (pause) G I suppose people like the languages You know ..
6 Like a ward - environment is 556. I mean 557. Some people like that I'd couldn't work there...

Concordance 2: PEOPLE LIKE in SCO

Line 1 in Concordance 2 is a false start. However, lines 4, 5 and 6 show the most commonly occurring form of *people like* in SCO employs *like* to indicate a preference. Line 3 (*we need more people like him*) is a comparison using a pronoun (*him*) while line 2 sees *like* as a filler that is employed to clarify a point: *we did then like .. move communities .. people like...* Here, *people like* is used to reformulate the formal term *communities*.

The highest occurring 3w clusters incorporating *people like* in MAC are *people like that* (*people like* + back-referent) and *people like to* (*people like* + verb phrase). These clusters appear in contexts as in *I like people like that and people like to tell stories* which means that the structure is similar to the one found in SCO. We must take into account, however, the difference of proportional use between these two forms in MAC: MAC appears to record the even more definite pattern

*people like + name / pronoun* (i.e. *We need people like Nader / people like you)* as the he most common usage of *people like*. It appears in 163 out of 364 occurrences of *people like* in MAC – 44.8% of all uses[190] while it only occurs once in the 6 occurrences of *people like* (*Marilyn Manson [...] We need more people like him*) in SCO – 16.7%

# Appendix VI (chapter 8.1)

## *VI.1 - Concordance (SCO) (P - positive / N – negative)*

Which is - you know Really very good - I think everyone who enrolled - Came P
2 She can (Pause) 225. Speak like 226. Ten (Pause) different languages 227. C Very good 228. She'd be 229. No good for - you studies P
She's a bit older than me 44. ((pause)) 45. And she - 46. She's very good 47. Isn't she 48. My sister 49. When talkin'about P
5 173. You know what I mean 174. So - 175. (pause) 176. But ehmm- 177. She's very good at that P
it was only like 220. Windowcleaner 221. You know 222. I mean 223. Wasn't very good 224. And he went to - 225. Aussi 226. And ... ehh... 227. He told N
he didn't have a -253. very good house254. just one to rent N
Yes -see We know our We know our technology (response) B (inaud) Not very good Is it I was expecting Nightvision (laughs) The bedrooms N
11 ly have to do that much work 95. L So you do re ... research 96. Ah that is very good P
12 e's a musician Me i thought she was your girlfriend M She is (...) J That's very good isn't (inaud) play that P
13 ause) Mi I thought you were watching it Neill M (laughs) (...) Mi It's Not very good either Is it N
We watched George I think it was George Stevenson (inaud) And that was very good Ta (inaud) Mi Yeah P

[190] In MAC, 40 occurrences of these are *people like myself* (18 occ.) and *people like me (22occ.*).

## ***VI.2 - Concordance MAC (neg)***

1 rvellous things for me rcking machines . I ' m afraid it ' s very , that ' s not very good ladies but no no , no no , oh no , oh they were very , very ve

2 rliament Street . Er f of them , the shape of them isn ' t very conven it ' s not very good for ar a ut of it . What we ' ve got here is a very conventional

3 o very going round to the various churches three that er , them teabags are not very good are they ? That what ? Them tea w a bloke at Waddington he ' s v
4 very , very good with videos Is he ? Yeah , real it ' s been er I know it ' s not been very good or nice for you Oh how you doin s and Spencers ' vouchers o
5 s very good It ' s been alright in there lo . That ' s . So , and it ' s just not very good , it ' s part of the tax code c we had beef last week which was

6 e very good at ! Do you Well like being bos hop no problem I know , but there not very good there , about three times as mu The Days was quite good this wee
7 a very good driver does he ? now , are you ackie said these cooker hoods are not very good well they ' re alright , but I got the two bars mm yeah , they

## ***VI.3 - Concordance (MAC) positive***

1 ry ! they and er they ' ve been around ? I know a bloke at Waddington he ' s very , very good with videos Is he ? Yeah thorough . Mhm . And her children are
3 ry , very slowly Here are , give us it d white . Mm . It was really good ! A very , very good night it was . That ' s aylight is improving now isn ' t it ?

5 ry , very well done ! Extremely Oh ! we Mm . But , th he is you know , he is very , very good about it , he would know ng but I didn ' t like the accent you
7 ry , very heavy turning over they ' ve e Antiques Roadshow . Ah yes ! It ' s very , very good ! Anything you like ther inking chair ! Can ' t remember now .
10 f that is very , a good design , reasonably good va ing ? Go ahead . Mhm . Oh very good ! Very , very good that is !

## Appendix VII (chapter 8.2)

### *The example of REALLY - Multiple repetition can lead to false statistics*

The comparator MAC, however, records multiple repetitions of *really*. And this looks even more drastic when we have a look at the five- and six-word clusters with *really:*

| N | Cluster | Freq. |
|---|---|---|
| 1 | REALLY REALLY REALLY REALLY REALLY REALLY | 72 |
| 2 | IT'S REALLY REALLY REALLY REALLY REALLY | 30 |
| 3 | REALLY REALLY REALLY REALLY REALLY LOOKING | 25 |
| 4 | I'M REALLY REALLY REALLY REALLY REALLY | 25 |
| 5 | REALLY REALLY REALLY REALLY REALLY VITAL | 25 |
| 6 | I ' M REALLY REALLY REALLY REALLY | 20 |
| 7 | IT ' S REALLY REALLY REALLY REALLY | 20 |
| 8 | REALLY REALLY REALLY REALLY VITAL DOES | 16 |
| 9 | DEAR I ' M REALLY REALLY REALLY | 12 |
| 10 | REALLY REALLY REALLY REALLY LOOKING FORWARD | 12 |

Table 1: The 10 highest occurring REALLY 5w and 6w clusters in MAC

Table 1 is the result that *WordSmith* presents when the top 5-6 word clusters are calculated. The exceptionally long lines of really single-word repetition seem to be confined to MAC. In fact, when the concordances are compared in detail, it turns out that there is one single speaker that uses *really* seven times as a single-word repetition and that repeats the same utterance. This means cluster one, listed as occurring 72 times above, has been recorded only twice - and is idiosyncratic to only one speaker.

## Appendix VIII (chapter 9.2)

| Cluster | SCO Freq. | % | MAC Freq. | % | LL | BNC/C Freq. | % |
|---|---|---|---|---|---|---|---|
| THINGS LIKE THAT | 12 | 1.3 | 365 | 1.6 | 0.82 | 268 | 1.2 |
| AND THINGS LIKE | 4 | 0.4 | 210 | 0.9 | n/a | 144 | 0.7 |
| STUFF LIKE THAT | 16 | 1.7 | 81 | 0.4 | 22.30 | 67 | 0.3 |
| AND STUFF LIKE | 9 | 0.9 | 59 | 0.3 | 9.37 | 53 | 0.2 |
| A BIT LIKE | 9 | 0.9 | 96 | 0.4 | 4.18 | 126 | 0.6 |
| BIT LIKE THAT | 2 | 0.2 | 28 | 0.1 | n/a | 46 | 0.2 |
| SOMETHING LIKE THAT | 11 | 1.2 | 555 | 2.3 | 8.08 | 458 | 2.1 |
| OR SOMETHING LIKE | 6 | 0.6 | 254 | 1.1 | 2.44 | 183 | 0.8 |
| ANYTHING LIKE THAT | 9 | 0.9 | 117 | 0.6 | 2.50 | 87 | 0.4 |

Table 4/b Pairwise comparison and LL of the most frequent SCO 3-4w like clusters in SCO and their MAC equivalents and BNC/C frequencies.

# Appendix IX (chapter 10)

## *Appendix IX.1*

| Cluster | Freq. | % | key word |
|---|---|---|---|
| I DON'T KNOW | 97 | 0.08 | **know** |
| YOU KNOW WHAT | 62 | 0.05 | know |
| A LOT OF | 60 | 0.05 | **a * of** |
| WHAT I MEAN | 55 | 0.05 | **mean** |
| KNOW WHAT I | 49 | 0.04 | know |
| YOU KNOW WHAT I | 47 | 0.04 | know |
| KNOW WHAT I MEAN | 46 | 0.04 | know/mean |
| YOU KNOW WHAT I MEAN | 45 | 0.04 | know/mean |
| YOU HAVE TO | 34 | 0.03 | **have /to** |
| I DON'T THINK | 31 | 0.03 | **think** |
| WHAT DO YOU | 31 | 0.03 | do |
| IT WAS A | 29 | 0.02 | was |
| USED TO BE | 27 | 0.02 | **to** |
| WHEN I WAS | 27 | 0.02 | was |
| YOU KNOW THE | 25 | 0.02 | know |
| YOU KNOW I | 23 | 0.02 | know |
| A BIT OF | 22 | 0.02 | **a * of** |
| A COUPLE OF | 21 | 0.02 | **a * of** |
| WHAT DID YOU | 21 | 0.02 | did/you |
| DO YOU WANT | 20 | 0.02 | do /you |
| YOU KNOW YOU | 20 | 0.02 | know |
| I HAVE TO | 19 | 0.02 | **I * to** |
| I USED TO | 19 | 0.02 | **I * to** |
| ALL THE TIME | 18 | 0.02 | the/time |
| THERE IS A | 18 | 0.02 | there |
| WE USED TO | 18 | 0.02 | used |
| AND IT WAS | 17 | 0.01 | was |
| DO YOU DO | 17 | 0.01 | do/you |
| DON'T KNOW WHAT | 17 | 0.01 | know |
| I THINK IT | 17 | 0.01 | think |
| I WENT TO | 17 | 0.01 | to |
| THEY USED TO | 17 | 0.01 | to |
| YOU WANT TO | 17 | 0.01 | to |
| AND HE SAID | 16 | 0.01 | **said** |
| DO YOU KNOW | 16 | 0.01 | know |
| GO TO THE | 16 | 0.01 | to |
| ONE OF THEM | 16 | 0.01 | of |
| STUFF LIKE THAT | 16 | 0.01 | **stuff/like** |
| TO BE HONEST | 16 | 0.01 | **honest** |
| WHAT ARE YOU | 16 | 0.01 | you |
| DID YOU DO | 15 | 0.01 | you |
| I WAS LIKE | 15 | 0.01 | like |
| ONE OF THE | 15 | 0.01 | the |
| WAS IN THE | 15 | 0.01 | the |
| WHAT DID YOU DO | 15 | 0.01 | you |
| BIT OF A | 14 | 0.01 | bit of |
| HAVE YOU GOT | 14 | 0.01 | you |
| LIKE YOU KNOW | 14 | 0.01 | like |
| OUT OF THE | 14 | 0.01 | the |
| TO DO IT | 14 | 0.01 | to |
| AND I WENT | 13 | 0.01 | and |
| AND THAT WAS | 13 | 0.01 | and |
| IT WAS LIKE | 13 | 0.01 | like |
| THAT WAS A | 13 | 0.01 | was |
| THAT'S WHAT I | 13 | 0.01 | that |
| THINK IT WAS | 13 | 0.01 | think |
| TO GO TO | 13 | 0.01 | to |
| WE WENT TO | 13 | 0.01 | to |
| YOU KNOW WHEN | 13 | 0.01 | know |

full version of **Table 1** SCO top 50 most frequent 3-5w clusters

## Appendix IX.2

| SCO CLUSTER | Freq | per 100k | **BNCC Cluster** | **Freq.** | per 100k |
|---|---|---|---|---|---|
| A LOT OF | 61 | 48.5 | A LOT OF | 1758 | 43.3 |
| A BIT OF | 26 | 21.7 | A BIT OF | 1040 | 25.2 |
| A COUPLE OF | 21 | 17.5 | A COUPLE OF | 754 | 8.3 |
| ONE OF THEM | 17 | 14.2 | ONE OF THE | 677 | 18.4 |
| BIT OF A | 16 | 13.3 | ONE OF THEM | 588 | 14.6 |
| ONE OF THE | 16 | 13.3 | THE END OF | 578 | 14.3 |
| OUT OF THE | 14 | 11.6 | ONE OF THOSE | 443 | 11 |
| A BIT OF A | 13 | 10.4 | SORT OF THING | 401 | 10 |
| SORT OF THING | 10 | 8 | CUP OF TEA | 388 | 9.6 |
| LOT OF MONEY | 9 | 7.5 | BIT OF A | 353 | 8.7 |
| A LOT OF MONEY | 8 | 6.7 | A CUP OF | 340 | 8.5 |
| IN FRONT OF | 8 | 6.7 | ONE OF THESE | 320 | 8 |
| LOT OF PEOPLE | 8 | 6.7 | A BIT OF A | 289 | 7.2 |
| ONE OF THESE | 8 | 6.7 | THE END OF THE | 289 | 7.2 |
| ONE OF THOSE | 8 | 6.7 | THAT SORT OF | 275 | 6.8 |
| BECAUSE OF THE | 8 | 6.7 | SOME OF THE | 273 | 6.8 |
| OF THEM ARE | 7 | 6 | THE REST OF | 271 | 6.8 |
| OF COURSE YOU | 7 | 6 | A CUP OF TEA | 264 | 6.6 |
| THE WHOLE OF | 6 | 5 | SOME OF THEM | 261 | 6.6 |
| COUPLE OF WEEKS | 6 | 5 | GET RID OF | 248 | 6.4 |
| OF THOSE THINGS | 6 | 5 | THE BACK OF | 241 | 6 |
| SOME OF THE | 6 | 5 | AT THE END OF | 240 | 6 |
| A LOT OF PEOPLE | 6 | 5 | IN FRONT OF | 232 | |
| THE REST OF | 6 | 5 | A LOAD OF | 222 | |
| WHAT KIND OF | 6 | 5 | SORT OF LIKE | 215 | |
| LIKE SORT OF | 6 | 5 | A SORT OF | 201 | |
| A CUP OF | 5 | 4 | OUT OF IT | 197 | |
| | | | THE MIDDLE OF | 190 | |

Table 1(b): SCO- BNC/C OF-cluster (3-5w) comparison

## Appendix IX.3

| **Key cluster (3-5w)** | **Freq.** | **RC. %** | **Keyness** |
|---|---|---|---|
| YOU KNOW WHAT I MEAN | 328.00 | 0.04 | -64.07 |
| YOU KNOW WHAT I | 375.00 | 0.04 | -60.81 |
| KNOW WHAT I MEAN | 372.00 | 0.04 | -58.64 |
| WHAT I MEAN | 619.00 | 0.05 | -45.68 |
| KNOW WHAT I | 515.00 | 0.04 | -44.99 |
| STUFF LIKE THAT | 69.00 | 0.01 | -35.45 |
| I WAS LIKE | 62.00 | 0.01 | -34.22 |

*Table 2(b)* : *3-5 word SCO cluster Keyness when compared to BNC/C clusters.*

## ***Appendix IX.4***

**Table 2: Top 20 three-word chunks**

| | item | frequency |
|---|---|---|
| 1 | I don't know | 5,308 |
| 2 | a lot of | 2,872 |
| 3 | I mean I | 2,186 |
| 4 | I don't think | 2,174 |
| 5 | do you think | 1,511 |
| 6 | do you want | 1,426 |
| 7 | one of the | 1,332 |
| 8 | you have to | 1,300 |
| 9 | it was a | 1,273 |
| 10 | you know I | 1,231 |

| | item | frequency |
|---|---|---|
| 11 | you want to | 1,230 |
| 12 | you know what | 1,212 |
| 13 | do you know | 1,203 |
| 14 | a bit of | 1,201 |
| 15 | I think it's | 1,189 |
| 16 | but I mean | 1,163 |
| 17 | and it was | 1,148 |
| 18 | a couple of | 1,136 |
| 19 | you know the | 1,079 |
| 20 | what do you | 1,065 |

**Table 3: Top 20 four-word chunks**

| | item | frequency |
|---|---|---|
| 1 | you know what I | 680 |
| 2 | know what I mean | 674 |
| 3 | I don't know what | 513 |
| 4 | the end of the | 512 |
| 5 | at the end of | 508 |
| 6 | do you want to | 483 |
| 7 | a bit of a | 457 |
| 8 | do you know what | 393 |
| 9 | I don't know if | 390 |
| 10 | I think it was | 372 |

| | item | frequency |
|---|---|---|
| 11 | a lot of people | 350 |
| 12 | thank you very much | 343 |
| 13 | I don't know whether | 335 |
| 14 | and things like that | 329 |
| 15 | or something like that | 328 |
| 16 | what do you think | 312 |
| 17 | I thought it was | 303 |
| 18 | I don't want to | 296 |
| 19 | that sort of thing | 294 |
| 20 | you know I mean | 294 |

**Table 4: Top 20 five-word chunks**

| | item | frequency |
|---|---|---|
| 1 | you know what I mean | 639 |
| 2 | at the end of the | 332 |
| 3 | do you know what I | 258 |
| 4 | the end of the day | 235 |
| 5 | do you want me to | 177 |
| 6 | in the middle of the | 102 |
| 7 | I mean I don't know | 94 |
| 8 | this that and the other | 88 |
| 9 | I know what you mean | 84 |
| 10 | all the rest of it | 76 |

| | item | frequency |
|---|---|---|
| 11 | and all that sort of | 74 |
| 12 | I was going to say | 71 |
| 13 | and all the rest of | 68 |
| 14 | and that sort of thing | 68 |
| 15 | I don't know what it | 63 |
| 16 | all that sort of thing | 61 |
| 17 | do you want to go | 61 |
| 18 | to be honest with you | 59 |
| 19 | an hour and a half | 56 |
| 20 | it's a bit of a | 56 |

Above: The most frequent (3w to 5w) clusters from the CIC corpus (5 million spoken ) in O'Keefe *et al.* (p.66: 2007)

***Appendix IX.5***

| Size of chunk | 3 word | 4 word | 5 word | 6 word |
|---|---|---|---|---|
| ranks in top 20 for: **OF cluster** | 2nd; 7th; 14th; 18th | 4th; 5th; 7th; 11th; 19th | 2nd; 10th; 11th; 13th; 14th; 16th; 19th | 2nd; 3rd; 4th; 6th;8th; 9th; 10th; 2th; 14th; 15th; |
| ranks in top 20 for: **TO** | 8th; 11th | 6th; 18th | 5th; 17th | 11th; 13th; 17th |

Both OF & TO are prominently used words. However:

- OF appears the more frequent the longer the cluster
- TO usage decreases in frequency the longer the cluster
- > TO appears to prefer occurrence in short chunks of English

(Adapted from O'Keefe / McCarthy and Carter (2007: 65ff.)

# INDEX

## *Q*



## *R*



## *S*



## *T*



## *V*



## *W*



## *Y*