THE SPOKEN CORE OF BRITISH ENGLISH: A DIACHRONIC ANALYSIS BASED ON THE BNC

MIGUEL FUSTER MÁRQUEZ AND BARRY PENNOCK SPECK
Universitat de València
barry.pennock@uv.es
miguel.fuster@uv.es

1. Introduction

This research takes as its starting point a frequency analysis of the demographic-spoken subcorpus of the British National Corpus in order to focus on two aspects of the evolution of spoken core vocabulary in British English. The first is the impact on the core of contact with other languages and, the second, the role of lexical innovation and/or replacement in the history of this core. Our analysis, which, to a certain extent, follows up on that carried out in Fuster (2007) questions the hypothesis that the spoken core is immune to foreign influence or that it is highly resistant to change.

2. The place of (core) vocabulary in linguistic research

Central to this contribution is the theoretical assumption that all speakers possess a core vocabulary that is more important to them in their daily verbal exchanges than other items in their repertoire, no matter how rich their particular vocabulary might be. In Stubbs' opinion, an outstanding feature of such basic vocabulary is that it is "known to all native speakers of the language. It is that portion of the vocabulary which speakers could simply not do without" (2002: 41). Several studies propose that basicness should be related to frequency. Lexicographers and corpus linguists would agree today with Kilgarriff's statement that the more frequently a vocabulary item is used, "...the more important it is to know it" (1997: 135). McCarthy (1990: 49) states that those learners who are equipped with knowledge of basic words have at their disposal "a survival kit [...] that they could use in virtually any situation [...] or in any situation where an absolutely precise term, the *mot juste,* might be elusive and where a core word would do".

In the literature on language contact a number of reasons have been mentioned for importing core words (see for example the exhaustive catalogue offered by Grzega 2003: 23-4). We believe that the main cause is related to the intensity of contact. While superficial language contact basically leads to the importation of so-called less central cultural items, when contact becomes more intense, it may affect core vocabulary as well as the linguistic structure (see Thomason 2001: 69-70).

Although mention of core is often made in contact, historical and comparative linguistics, whether the study of such vocabulary, or, for that matter, vocabulary in general, should be given a prominent position remains controversial. Lass (1987: 60) maintains that changes in vocabulary are largely irrelevant because "[l]exis changes easily, but the structural frames it fits into are more resistant, and tend to remain, changing only under language-internal conditions". In a similar vein, Labov (2001: 13-14) points out that "the replacement of vocabulary seems to have many characteristics of random variability". In his view, it is impossible to propose constraints on lexical change and it is hard to know "which words have a better chance of surviving and which do not" (2001: 13-4). Research in language contact has also shown conclusively that the lexicon is, without doubt, the most borrowable of all subsystems of language (Thomason 2003: 694). In his analysis of contact areas for instance, Haspelmath finds that borrowing can be quite massive. In the case of Australia, it has been found that all kinds of words, core and non-core, were "easily borrowed" (Haspelmath 2004: 209). Rankin (2003: 187) also comments on the lack of borrowing contraints in East and Southeast Asia, where basic numerals are known to have been imported from Chinese, and Campbell observes a similar phenomenon in Finnish, Turkish and Persian (2003: 271). On the contrary, "the situation of well-studied Indo-European languages such as German, French or Russian, where loanwords are easy to identify and occur only in rather circumscribed domains, may be atypical" (2003: 2). Therefore, in principle there seems to be evidence that importing of vocabulary is the rule rather than the exception in the different linguistic areas which have been studied to date.

The study of lexical change lies at the heart of work in *lexicostatistics*, initiated by Swadesh (reported in Gudschinsky 1956). Lexicostatistics deals exclusively with a select word list of vocabulary elements, for which the borrowability rate is extremely low. More recent proposals, like that of McMahon (2004), have revitalized the idea of selecting a small number of core words for the purposes of examining linguistic change before the existence of written records. One needs to refer to work done in lexicostatistics and glottochronology since the only basic word lists which still deserve the attention of comparative and historical linguists most typically come from these theoretical frameworks (see Fuster 2007).

3. The establishment of core lists: frequency and counts

From a comparative and diachronic angle, drawing up a list of core words containing stable members that show great resistance to change is, to say the least, problematic. We agree with Haspelmath that the first and foremost problem is that many researchers do not make it clear what they mean by core (2003: 2; 2004: 212), and so there may be disagreement about (1) the items which are considered as core, (2) the threshold between core and non-core word lists, and (3) what the theoretical principles which make up the basis for such lists are. We can only agree with Haspelmath that a greater refinement of this concept is urgently needed.

As stated above, the idea of *core* is not new or unknown in diachrony. The so-called Horn list (1926) (see Berndt 1984: 69 for a fuller account) is an early pioneering application of the idea of core vocabulary to etymological research in the history of the English language that was based entirely on a corpus of written American English. More recently, etymological research related to English vocabulary can be found in Bird (1987). It is to be noted that Lutz's (2002) application of core to the history of English in fact quotes earlier work based on Michael West's *General Service List* published in 1953.

We wish to single out the research carried out by Hughes (2000: 391-4), as it shows greater parallelisms with our own. His 600 word-list is made up of the most common components of the "Longman Defining Vocabulary", included in every new edition of the *Longman Dictionary of Contemporary English* (LDOCE). This vocabulary, in its turn, is based on the *British National Corpus* (BNC). Hughes' core word list therefore relies on *frequency* as a primary factor.

The importance of taking into account frequency as an index of basicness and, therefore, as an index of the relative weight of certain words in a language is that it is entirely based on the empirical observation of language use. This issue has been more amply discussed in Fuster (2007). In contrast with earlier, intuitive lists, we agree with Lee (2001) that frequency is, at the very least, an objective factor. There is a strong psychological or cognitive basis which supports the validity of frequency as more rigurous than other factors in the selection of vocabulary by speakers. For morphologists like Aronoff and Fudeman (2005: 225) it is clear that "[i]f a word is very frequent, it has been reinforced in their memories and so speakers will find it easily". Plag (2003: 49) believes that "[...] there is a strong tendency that more frequent words are more easily stored and accessed than less frequent words" (see also Haspelmath 2003: 43-4).

Psycholinguistic research has been able to confirm that the most frequent words are the first that come to mind when individual speakers use their own language. Thus it follows that these items will also be the most resistant to loss, replacement or change. Aronoff and Fudeman note, for instance, that frequent vocabulary also preserves and indeed is characterized by older morphological irregularities not found in the rest of their vocabulary (2005: 225), and from a historical perspective, these morphological traits also resist sinking into oblivion.

Gradually, diachronicians have started to welcome frequency as a crucial factor in the determination of core, particularly because no other factors are equally objective. Haspelmath claims that the notion of frequency is essential in any proposal of core vocabulary since "it is well known that high-frequency items are resistant to types of language change such as analogy" (2003: 6). In their recent history of English, Brinton and Arnovick admit as a general linguistic observation that "the more frequently used the word, the more likely it is to have survived" (2006: 165).

A recurring problem arises when we start to consider the number of words which should be counted as basic since this will obviously yield different results. Traditionally, diachronicians have generally considered as basic the rather low figures of 100 or 200 items. We contend that this is a very restrictive view of core which has been challenged by research in corpus linguistics and modern lexicography. Researchers in these fields have observed that even the simplest kind of conversation in English cannot be carried out with 200 items alone. On average, lexicologists are more in favour of proposing an inventory of around 2,000 items as absolutely indispensable words (Sinclair 1991: xviii; Nation 2001: 15; Stubbs 2002: 42; McCarthy 1999: 248; and 2003: 61, note 3). A further observation of corpora shows that while there is a general consensus that up to 2,000 items is sufficient to be considered core, beyond such a figure different corpora will diverge (Kilgarriff 1997: 14). Indeed, Stubbs (2002: 42) finds that discrepancies may arise earlier, and there is only significant agreement about "the top few hundred words in different general corpora".

4. The spoken mode and the establishment of its lexical core

Practically all discussions concerning lexical calculations in diachrony carried out in the past have been based on written language, naturally because research on spoken data was impossible given the lack of adequate technology and reliable data.

However, there are sufficient reasons to think that changes in the spoken mode deserve greater attention. Stubbs (1996: 70) establishes an explicit correlation between the kind of words which can be expected to appear in the two language modes, and goes as far as to propose a correlation in terms of word origin:

spoken	**written**
everyday	*academic*
common	*specialist*
frequent	*rare*
informal	*formal*
monosyllabic	*polysyllabic*
Germanic	*Romance/Graeco-Latin*
acquired	*learned*
active	*passive*
core	*non-core*

It is undeniable that through these oppositions Stubbs (1996) sums up well-established assumptions among historians of the English language who seem to take for granted that the most important historical changes are internal, and not external, that is, due to contact. For Stubbs, the spoken language contains common vocabulary which is 'acquired' (that is, not accessed through formal education) and is Germanic. On the other hand, Stubbs states that the written mode contains more peripheral, specialised vocabulary, most of the borrowings, and is 'learned'.[1]

We are not alone in arguing that the opposition between *core* as a defining characteristic of spoken English and *non-core* as a defining characteristic of written English is not as tenable as suggested by Stubbs (1996). McCarthy &Carter (2003: 5) have held that both language modes contain core and non-core items. According to Lee, this is precisely because core lexis is "central to the language as a whole and thus not specific to any lect or register" (2001: 250). Support for McCarthy & Carter's (2003) and Lee's (2001) position can be found in the fact that the majority of the core items in our word-list of conversational English are also shared by the other sub-corpora of the BNC. The greatest difference between conversational and written English is that a larger number of words obtain higher frequencies in the written mode than in the conversational spoken mode.

5. Research on spoken vocabulary in the British National Corpus

Our study of frequency lists in spoken English is based on the lists of Leech *et al.* (2001), and its companion website.[2] Both refer to the BNC, which contains

Contemporary English vocabulary, 10% of which is devoted to spoken Contemporary English, gathered since 1991, and 90% to written Contemporary English gathered after 1960. The reasons which have led these corpus linguists to include such a comparatively small proportion of spoken English is that they found its analysis to be "a skilled and very time-consuming task" (2001: 1). Nevertheless, the size of the spoken section –10 million words– is, in their view, "also sufficiently large to be broadly representative" (2001: 1). The compilers even go as far as to suggest that "no bigger transcribed purpose-built cross-section of spoken language exists at present".

Undoubtedly, the BNC surpasses earlier corpora of the English language; and it is also more up to date. The editors provide frequency lists and interesting contrasts between the lists (2001: XI). Recently, Hoffmann stated that "the availability of large corpora such as the BNC has enabled an even more precise description of both language structure and language use"(2004: 203-4). With its 10 million words, the spoken component is, for instance, ten times larger than the whole *Brown Corpus*. It is important to note that the spoken component that we have examined in the BNC admits of a subdivision between (1) *conversational, context-governed speech* and (2) *task-oriented speech*. In fact, we have decided to discard task-oriented speech as it shows a strong resemblance to *written* English in various ways, the most important of which is that it centers on specific activities, such as lectures, business meetings, interviews, political speeches, etc., and arguably such registers are in various ways closer to the written mode. Consequently, it is also likely to exhibit greater formality than conversational English.

The conversational section is, according to the compilers, the most innovative part of the corpus. One hundred and twenty-four adults (aged fifteen or over) were selected from different places across the United Kingdom. In their selection of informants they considered as relevant the sociolinguistic variables of sex, age and class, so there is approximately an equal number of men and women; the informants, practically all adults, were grouped into six age ranges; and there is a balanced selection of social groups. Geographically, the UK was divided into three major areas, the South contributed the most with 45.61% of the informants, the North, 25.43%, the Midlands, 23.33%, and a small proportion, 5.61%, belongs to unclassified areas. For the compilers of the BNC "the importance of conversational dialogue to linguistic study is unquestionable: it is the dominant component of general language both in terms of language reception and language production".[3]

6. Lexical differences in the written and spoken subcorpora of BNC

Leech et al. (2001: XI) find that in the written section "there is an overlapping subset of only 33 words shared with the top 50 words of the spoken corpus". Thus, while the most frequent word in written English is the determiner *the*, in the case of spoken conversational English, the most frequent word is the verb *to be*. Table 1 below focuses on the 20 top lemmas[4] in *written* English alongside the top 20 items in the *demographic conversational* (henceforth conversational) and the *task-oriented* section.

WRITTEN SECTION		DEMOGRAPHIC SECTION		TASK-ORIENTED SECTION	
1	**the**	1	**be**	1	**be**
2	**be**	2	**I**	2	**the**
3	**of**	3	**you**	3	**and**
4	**and**	4	**it**	4	**I**
5	**a**	5	**the**	5	**you**
6	**in**	6	**not**	6	**it**
7	**to** *inf*	7	**do**	7	**a**
8	**have**	8	**have**	8	**of**
9	**to** *prep*	9	**and**	9	**to** *inf*
10	**it**	10	**a**	10	**have**
11	**for** *prep*	11	**that** *detp*	11	**in**
12	**he**	12	**to** *inf*	12	**we**
13	**I**	13	**they**	13	**that** *detp*
14	**that** *conj*	14	**yeah**	14	**do**
15	**not**	15	**he**	15	**not**
16	**with**	16	**get**	16	**they**
17	**on**	17	**oh** *int*	17	**er**
18	**by**	18	**she**	18	**that** *conj*
19	**they**	19	**what** *detp*	19	**to** *prep*
20	**she**	20	**go**	20	***erm*** *uncl*

TABLE 1: The 20 most frequent lemmas in the written, demographic and task-oriented sections of the BNC

Some items in the conversational section are not found in the written section, notably the verbs *to get* and *to go*, the pronoun *you*, the determiners *that* and *what,* and also some interjections. But important differences in terms of frequency also emerge from a comparison between the two sections of the spoken subcorpus. In order to compare the top 20 items in both lists, we have focused on their log-likelihood ratio (G), which, according to the compilers,

> show[s] how high or low is the probability that the difference observed is due to chance. This statistic can be considered to demonstrate how significantly

> characteristic or distinctive of a given variety of language a word is, when its usage in that variety is compared with its usage in another. Leech *et al.* (2001: 16)

According to Leech *et al.* (2001: 17) the higher this ratio, "the more significant is the difference between two frequency scores". Indeed, the greatest log-likelihood ratios are found among the top items. Most of those in the conversational word-list are much more frequent in that kind of variety than in the task-oriented section, e.g. *I, you, it, not, do, have, that* (determiner/pronoun), *yeah, he, get, oh, she, what* (determiner/pronoun). For certain items the ratio is extremely high, such as the personal pronoun *I*. In their research, O'Keeffe, McCarthy & Carter (2007: 33) claim that "...the high rank of *I* and *you* in the spoken data, along with discourse-marking items (e.g. *well, right...*)", seems to indicate "an overall orientation to the speaker-listener world in conversation". By contrast, some items in the task-oriented subcorpus also have a high differential ratio, as is the case of *the, of,* the filler *er,* or the conjunction *that*. When we turn our attention to the bottom of these frequency lists, other striking contrasts emerge. While some items are definitely core in 'context-governed' English, they are practically irrelevant in conversational English, and viceversa. This is the case of lemmas like *authority, development, in terms of, chairman, motion, councillor, settlement*, etc., whose frequency is lower than 10 per million in conversational speech, but higher than 100 per million in task-oriented speech. Since frequency is a key factor in our research, the conclusion to be drawn from such evidence is that these and other items in the task-oriented lists are, on the one hand, closer to the written mode, and on the other, cannot be described as basic within the conversational English sub-corpus.

McCarthy & Carter (2003: 5) point out that the written part of the corpus shows greater lexical density and variation than the spoken part. It may be observed that the task-oriented subcorpus also shows greater lexical density than the conversational subcorpus. Whereas the lemmatized task-oriented subcorpus contains over 850 lemmas which occur more than 25 times per million words, the conversational subcorpus does not reach 800 lemmas. This numerical problem has been an extremely important issue which has necessarily been reflected in our research.

It seems clear that at the lexical-semantic level there are differences between the core elements of contemporary written and spoken English. Many words which are more frequent in the written register, are practically absent from the spoken section. For instance, a number of closed class items are part of the written core but absent from our spoken word-lists: *above, according (to), among, as well as, despite, former, herself, including, latter, nor, several, such as, though* (conj.), *thus, whom, whose*. So, the obvious conclusion seems to be that not every function word is by definition a member of the core from the viewpoint of ordinary conversation.

Stubbs has noted that (2002: 42) "raw frequency lists often have odd gaps". One finds that certain discrepancies might be due to differences among the corpora themselves. For other gaps a coherent historical explanation could be offered, as long as we are willing to admit the possibility that basic items may also change, though gradually, with the passage of time. For example, though contrary to expectations, some words referring to human body parts are not listed as members of the core in the conversational section of the BNC. In our lists, nouns like *heart, mouth*, *neck* or *nose* are not among the most basic items because their frequency is lower than two per million words. Note also that while McCarthy (1999: 243) mentions three nouns which refer to seasons, namely *winter, spring* and *summer*, as part of the core in the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE), a large corpus of spoken English, none of these are members of the core in conversational English in the BNC, once again because their frequency is lower than two per million. Gaps and differences are also observable in reference to the colour spectrum. So, while *black, white, red,* and *blue* are in our list of 700 core items, others, like *green, yellow, brown* or *grey*, some of them included as core in CANCODE, are not found in our conversational list drawn from the BNC. In both corpora, the non-core items have lower frequencies, whereas *black, white* and *red* always show high frequencies. This might also prompt us to wonder if a substantial part of those items which are non-core today have ever been so. Perhaps important societal, cultural, or even structural changes may have had an effect on the core. But only a similar kind of quantitative research performed on earlier stages could answer such question. Even though we readily agree with Stubbs (2002: 42) that "the vocabulary is a structured whole, not an unordered list of words", we cannot introduce 'unjustified', non-empirical or intuitive criteria in our definition of core in order to produce structured word series. Instead, what should be done is try to give an account of the results obtained. In fact, very often we may come across an explanation for gaps in word series. For instance, McCarthy (1999: 242) argues that the reason for discrepancies in frequencies observed in the days of the week lies in the fact that "in Westernised, Christian societies, Monday is considered the start of the working week; Friday and Saturday are associated with the week's end and leisure". Also, it is not hard to come up with cognitive explanations for the differential frequencies in colour names. As we all know, some colours, say *black* and *white*, are certainly more basic than *yellow, pink* or *grey*, a fact that is backed up by extensive research on the subject.

7. Some remarks about the nature of core conversational items

One of the salient findings in top word frequencies within corpus linguistics has been the predominance of closed-class items. In the article "What constitutes a basic vocabulary for spoken communication" published in 1999, McCarthy examines the main features of this English core vocabulary (see also O'Keeffe, McCarthy & Carter 2007: 31-57). His work is relevant to our research as the corpus he examines, CANCODE, offers very strong parallelisms with the lists obtained in BNC: practically all the items McCarthy mentions also appear in the BNC core list (see McCarthy & Carter 2003).

McCarthy addresses issues of meaning and functionality in the core items. He observes that a very large number of top items "clearly belong to the traditional province of grammar/function words, in that they are devoid of lexical content" (1999: 236). These are "articles, pronouns, auxiliary verbs, demonstratives, basic conjunctions, etc". For McCarthy & Carter (2003: 6) the category of function words contains up to 200 members. In our research, function words account for 53 of the first 100. Besides these, which are quite clearly characterised in CANCODE and the BNC, McCarthy establishes another nine categories which share the characteristics of both closed- and open-class words. These categories, "which are equally important as components of basic communication" (O'Keeffe, McCarthy & Carter 2007:37) include modal items, delexical verbs (quasi-auxiliaries, Brinton & Arnovick 2006:385), stance words and discourse markers. Another large group is made up of items which can stand in for other members of their word-class. McCarthy & Carter (2003: 6-7) label these: basic nouns, general deictics, basic adjectives, basic adverbs, and basic verbs. Although they may be deemed more lexical than grammatical, the fact that core words like *thing, do, lovely*, etc. can be used in lieu of many members of their same class is a good reason to classify them as empty of lexical content (Halliday & Hasan, 1976: 274). It is clear that these items "defy an easy fit into the traditional word classes of noun, verb, adjective, adverb or interjection" (O'Keeffe, McCarthy & Carter 2007: 46).

8. Limits of our core list for conversational English

A threshold in frequency has to be established to differentiate what is core from what is not core in present-day British English. McCarthy (2003: 46) examined the spoken subcorpus of CANCODE (part of the *Cambridge International Corpus*) which is made up of a total of 5 million words recorded between 1995 and 2000. Using a graph he shows that "round about 2000 words down in the frequency ratings" the number of occurrences per item "begins to drop more steeply" (McCarthy 1999: 235). This allows him to conclude that 2000 lexical items is quite a safe borderline which distinguishes high frequency (core items) from low frequency items (also McCarthy 2003: 61 note 3). In the CANCODE sample that McCarthy (2003) analyzed, which does not distinguish between demographic spoken English and task oriented spoken English, 1500 unlemmatized words occurred more than 100 times.

We decided to look at the frequency of lemmas rather than word forms as we wished to draw comparisons with earlier diachronic research in English which focused exclusively on lemmas. Secondly, unlike McCarthy (2003), we wished to focus on naturally occurring conversation and not any other kind of spoken variety. Our list is taken from Leech et al. (2001), which is based on the four-million word conversational subcorpus of the BNC. We believe that the BNC subcorpus is large enough to represent contemporary conversational speech. Leech et al.'s (2001) original list is made up of the 880 most frequent lemmas; the cut-off point was established by excluding lemmas which occur fewer than twice per million. We decided to omit items such as letters of the alphabet, proper names, titles, days of the week, months, currencies, countries, nationalities, religions and most interjections (Fuster 2007). This yielded a slightly smaller list of 852 words. In Kilgarriff's view items like these "though wordlike enough to be in the dictionary, were not wordlike enough to count for the purposes of the frequency list" (1997: 143).[5] An additional reason for not including them is that it is impossible to trace their evolution through the use of historical/etymological dictionaries.

Our analysis of the resulting list showed that the 700th most frequent lemma occurred 48 times per million, but the 800th most frequent lemma occurs only sixteen times per million. As we decided to work with series of 100 lemmas in line with earlier diachronic research, item 700 became our cut-off point. Less frequent items clearly show greater variability and discrepancies between the BNC and other corpora.

9. The sources of core words in conversational English

Once we arrived at our our list of 700 lemmas, we made use of two online dictionaries for the diachronic analysis of each individual item, the *Oxford English Dictionary* (OED) and the *Middle English Dictionary* (MED). The first was the main source of data while the second was used as a secondary source exclusively to offer antedatings although it had negligible effects on our final results. A caveat seems appropriate here: as we know, the information contained in such historical dictionaries is primarily based on written sources. Therefore, the results offered below need to be interpreted as indirect evidence rather than direct evidence of historical change in the spoken vocabulary of British English.

Some preliminary explanation of the ideas concerning the labels used for the source languages in this research is required. We have made use of the term *native* to refer to the lemmas belonging to the language used by the original Germanic invaders and settlers from the Continent. These are words which have not resulted from contact. The term 'native' also allows us to include core words such as *dog*, *boy* or *girl* whose origin is obscure, but with no proven links to earlier Germanic. The label Germanic, which is often found in the literature, is therefore inadequate when one wishes to refer to innovations found solely in English, not in cognate Germanic languages, like the examples above or items like the pronoun *she*, the earliest attested example of which is dated 1160 (found in the continuation of *Peterborough Chronicle*). The only drawback with the term *native*, in our view a minor one, is that, in principle, every word in English –including adoptions from other languages– eventually becomes *nativized* as soon as it is felt to be so by the speakers of that language, that is, if a word has won general acceptance and therefore its origin is no longer considered alien. For instance, a noun like *city*, adopted from French during Middle English, is undoubtedly felt by contemporary speakers of English to be as native as any other and has been considered so for many centuries.

With regard to the foreign sources included in our tables, ON stands for Old Norse. In this particular case we can be more precise because it is indeed from that donor and stage that practically all Scandinavian words in the core come. As diachronicians know, the case of French is more problematic as different varieties and periods have, according to the literature, exerted their influence on English. It has been repeatedly indicated that *Anglo-Norman* is the most important donor variety to judge from the number of items currently in the core; but language historians will readily admit that there are a substantial number of items for which it is not always possible to establish a distinction between varieties in relation to their contrasting impact during the Middle English period (see also Rothwell 1998). Moreover, since French has continued to exert its influence on the core during the Modern period, the general label *French* seemed more appropriate than any other. An almost identical problem concerns the term-*Latin*, which we have adopted as a general label, although it is beyond discussion that different varieties of the Latin language have been the source of English words at different times. For some words though, the label *Latin/French* has been found convenient as the OED cannot distinguish clearly one or the other as single donors. Indeed distinguishing between these two sources of Romance words, Latin and French, in medieval times has often proved notoriously challenging (Pyles, 1971: 318; Brinton & Arnovich 2006: 239-40). In relation to the words we examined in our research, we have found that, in contrast to the OED, the MED is less inclined to distinguish between these two sources. Part of the problem in the selection of a donor language relates to the historical development of meanings. If we acknowledge a certain identity of form, we often see that French has actually reinforced the presence of items adopted earlier from Latin. This is well illustrated by the evolution of the noun *place*: the OED gives post-classical Latin *platea* as its source in Old English, but Anglo-Norman and Old French are mentioned as later donors which strengthen this particular Latin word's continuity in the language throughout Middle English. Let us now examine statistically the origin of core words in spoken conversational Contemporary English.

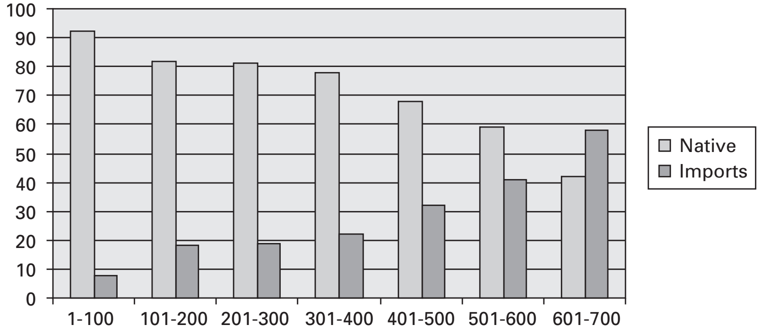

FIGURE 1: Native vs imported lemmas per 100 lemmas

Figure 1 shows the proportion of native and imported items in the most frequent 700 lemmas divided into series of one hundred items. We may observe that foreign elements are certainly present in this conversational core, but not evenly in all series. The native element is predominant in the top 100 items, with a percentage of imports as low as 8%, but then diminishes gradually until we reach items from 600 to 700, where the number of imports surpasses that of native lemmas.

Table 2 above shows the origin of such words in greater detail. Again we have subdivided the list in groups of 100 lemmas in descending order of frequency. As expected, the lists show on the whole that the native element is the most important in conversational English (71.7%), but the non-native and non-Germanic element is also substantial (28.3%). If words of Norse origin are added to the native element, we might be tempted to conclude that the 'Germanic' core amounts to 75.3%, but this is somewhat misleading, as we shall explain. The greatest contribution to the core from foreign sources is due to the incorporation of French

Lemmas in descending order	**Native**	**ON**	**French**	**Latin**	**French/ Latin**	**Other**
1-100	92	5	3	—	—	—
101-200	82	6	9	2	1	—
201-300	81	1	14	3	—	1
301-400	78	4	12	5	1	—
401-500	68	5	15	9	2	1
501-600	59	1	31	3	4	2
601-700	42	3	45	5	5	—
TOTAL ***%***	**502** **71,7**	**25** **3,57**	**129** **18,42**	**27** **3,85**	**13** **1,85**	**4** **0,57**

TABLE 2: Origin of core words in conversational Contemporary English (based on the demographic subcorpus of BNC)

and Latin words whose statistical importance is clear. In principle, if we added up all the French and Latin words, the adoptions from both Romance languages amounts to 24.1%. It is also quite patent that only French and Latin should be seriously counted as donors to the spoken conversational core, with a clear numerical predominance of French over Latin. The influence of other languages is extremely low in the conversational core, contributing only 0.57% of the total.

But such figures alone do not explain how or to what extent the vocabulary of English has been renewed through importation from Romance languages. Indeed, the vocabulary of English cannot be viewed as consisting of separate layers of inherited material and adoptions from foreign languages, since foreign elements may also be detected in new native coinages. Romance roots are recognisable in at least 31 words in the native group. This is the case of common words like *just, difficult*, or *really*, which do not have a Germanic ancestry. The very existence of this group of words also provides an argument for discarding the term 'Anglo-Saxon' to refer to the whole 'native' element. Also included within the native stock are eleven words whose etymology is obscure or unkown. This means that items which can safely be called native as they have been formed in English with entire Germanic morphology or inherited from an ealier ancestor, that do not contain Romance elements, decreases to 65.7%. On the other hand, if we count 'native' lemmas which contain romance morphology as part of the Romance element, its presence would rise to 28.5%. All in all, then, French and Latin are the source of close to a third of all core words in spoken conversational Contemporary English. Unfortunately, to our knowledge there is no similar word list with which our findings may be contrasted. The only widely used core lists of Modern English almost exclusively have written English as their basis. This research and that carried out by Fuster (2007) based on the top 1000 items in the entire BNC cast reasonable doubts on the assertion often made that over 80% of the top 1000 words in Modern English vocabulary is of Germanic origin (see, for example, Brinton and Arnovich 2006: 166).

A remark should also be made about the contribution of Old Norse, which, as seen in Table 2, is conisiderably less than that of French and Latin in the core. Although it has been suggested that Scandinavian is the source of many common core items in ordinary spoken English, our research contradicts this assumption. Although the relevance of Old Norse is undeniable, many supposedly core items from that donor quoted in textbooks have not been found in this list. For example, the nouns *neck*, *skin*, *sister, sky*; adjectives like *ugly*, *weak* or *ill*, the verbs *to guess, to cast, to smile,* and many others. We may conjecture that the contribution of Old Norse vocabulary, though of unquestionable importance to the conversational core in Contemporary English, has been most probably overrated, while that of Romance languages may have been downplayed. One possible reason for the low percentage of Old Norse words is that core items in frequency lists tend to be either functional or have a more abstract meaning, whereas many Scandinavian imports like those often quoted in textbooks show more concrete meanings.

10. The Rate of change: renewal of English vocabulary across time.

By definition, the core is in principle the layer that changes the least in the lexicon, but diachronicians agree that over a long period even the core, like any other aspect of language, is subject to change. We consider that more work should be carried out on the possible relationship between the importation of words, as the logical outcome of contact, and the rate of renewals and retentions in lexis. Our contribution attempts to put together both aspects in a historical perspective, considering the rate of retention and/or change these lemmas have gone through to date. Tables 3 and 4 show the evolution of each set of one hundred words in descending frequency order and the stages in which new core words have been incorporated. We have ventured to go as far back as the so-called Germanic stage, not beyond. Principles of a more hypothetical sort are required to discuss whether certain words are inherited from IE in an unbroken genealogical line, or have resulted partly or entirely from contact during the pre-Germanic period. One of the greatest difficulties in this kind of research is the lack of data regarding specific word-formation processes during periods where no written records exist, so any results would be quite speculative. Moreover, when etymological work has been carried out in order to determine whether certain terms go as far back as the Germanic or the Indo-European stages, it has been limited to the detection of roots, not entire lexical items (see for example Bird 1987). We considered that basing our approach on the examination of 'roots' to determine a word's ancestry was inadequate because, on the one hand, it did not address the issue of possible later word formation processes or the role of contact, which we surmised were likely to be present in very ancient native forms. Again, we have avoided any distinction made on the basis of whether words should be labelled Proto-Germanic, West Germanic or Anglo-Frisian, all of these representing acknowledged evolutionary stages within Germanic prior to the emergence of English. Thus the term Germanic as used here includes lexical items that can be traced back to a Germanic source in any (sub)period earlier than Old English, and items or roots which in other kinds of research might be traced as far back as Indo-European:

	0-100	-200	-300	-400	-500	-600	-700
Gmc(includes IE):	69	49	43	45	43	27	21
OE (until 1150):	13	16	13	7	12	10	7
ME (until 1499):	17	26	29	31	31	45	60
Mod (until 1900):	1	9	11	17	14	18	12

TABLE 3: Retention and change in the core by period

Table 3 shows the rates of retention and change to demonstrate that the core has undergone historical changes which cannot be accounted for entirely through contact. The layer pertaining to the Germanic stage, the oldest in this analysis, makes up 42.4%, whereas 57.6% of this core has been renewed in historical times. If Germanic is interpreted as those lexical items which were inherited from the ancestor(s) of Old English, the figures are definitely smaller than have been suggested to date. Some lexical renewals seem to have occurred during OE, but most of these took place during the ME period and, interestingly, there is also some observable continuity in ModE. The fastest rate of change for each one-hundred-word set in descending order takes place during the Middle English period. Note also that some items which correspond to the Germanic period are, in fact, adoptions from Latin, consequently renewals, such as the nouns *pound* and *box,* or the verbs *to turn* and *to stop.* More important, however, is the fact that not every innovation in ME is the result of copying. Table 4 and Figure 2 allow a much more detailed analysis. The columns corresponding to Germanic and OE have been transferred unaltered from Table 3, but when we reach the twelfth century the distribution can be properly carried out century by century. The twelfth century should be interpreted as corresponding to 50 years, from 1150 until 1200, the earliest stage in Middle English, since there is a wide, though not unanimous, consensus that the period up to 1150 corresponds to Old English. In brackets, we include terms for which no recorded date has been provided in the OED.

By centuries	**Gmc**	**OE**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0-100	69	13	7	6	2	2	—	—	—	1	—
101-200	49	16	5	11	9	1	4	1	1	2	(1)
201-300	43	13	4	16	8	1	6	1	2	—	(2)
301- 400	45	7	6	11	10	4	12	1	1	3	—
401-500	43	12	4	15	8	4	8	2	1	2	1
501-600	27	10	1	23	13	8	6	5	2	1	1(3)
601- 700	21	7	2	25	24	6	5	4	1	1	(1)
Total per period	297	78	29	107	74	26	41	14	8	10	9
Percentage %	42,4	11,1	4,1	15,28	10,5	3,7	5,8	2	1,1	1,4	1,2

TABLE 4: Innovations in the core from Germanic up to today

The inherited Germanic element amounts to 42.4% of the core, so more than half has been renewed from OE until today. Whereas 53% of the core was present in the Old English period, the rest has changed from the start of the Middle English period to date. The OE period, with a share of around 14.5%, shows a comparatively low rate of innovation, which practically equals the rate of renewals during the Modern English period.

If we are to trust the dates of first recorded instances of words in the OED and MED, the fastest rate of change takes place during the Middle English period, starting from the twelfth century. The thirteenth century witnesses a notable acceleration in the rate of innovations as reflected in the conversational core today. In fact, this century alone sees the renewal of a greater number of items than those effected during the entire Old English period. Together with the fourteenth century, both the second half of the twelfth century and the thirteenth century represent by far the highest rate of core renewal in the whole history of the English language from the Germanic period until the twenty-first century, with no less than

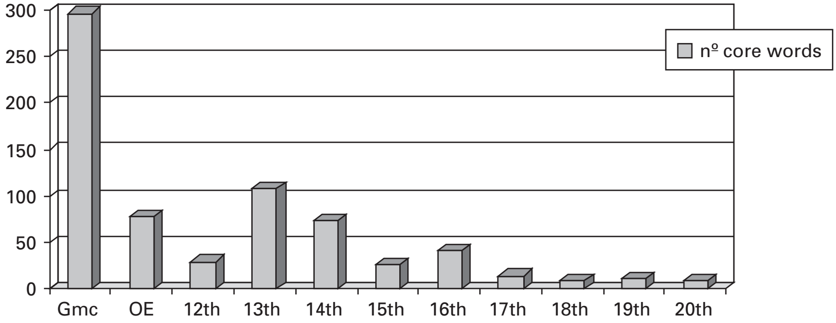

FIGURE 2: Innovations in the core from Germanic up to today

25.7% of all renewals. It is no accident that this coincides with the period during which the greatest number of adoptions, core and non-core, are recorded in the whole English vocabulary. Jespersen had already noted "that the linguistic influence did not begin immediately after the conquest and that it was strongest in the years 1251-1400, to which nearly half of the borrowings belong" (1967: 87). Of course, Jespersen's study was based on dictionary searches, and not on corpus linguistics, since this was not available in his time. Yet, imports alone cannot explain all the innovations in core vocabulary.

If attention is paid to the rate of change while taking into account the order of frequency, it is also obvious that the top one hundred words show the highest degree of permanence as 69% of these words go as far back as the Germanic stage. After the first hundred items, there is a steady decrease in the number of Germanic words.

It may be useful to establish a rapid comparison with the results obtained from a similar analysis effected on the *whole* BNC as seen in Table 5 (Fuster 2007: 715).

CENTS	OE	12TH	13TH	14TH	15TH	16TH	17TH	18TH	19TH	20TH
Total	413	22	138	222	73	84	36	9	5	1

TABLE 5: Statistics from the 1000 core words in the whole BNC corpus (90% written, 10% spoken)

We observe once again that the rate of change accelerates during the thirteenth and fourteenth centuries; and that after the early Modern English subperiod, it slows down again. The obvious difference is that there is a greater number of items renewed during the fourteenth, fifteenth and sixteenth centuries in the whole corpus than in the core list corresponding to conversational English. In all likelihood this is because when one considers the BNC as a whole, both written English and task-oriented speech have a greater weight. The degree of retention, looking at the Germanic subperiod and Old English together, makes up 53.5%, whereas in the core lists obtained from the whole corpus it is 41.3%. In any case, there is no doubt that both written and ordinary conversational English have changed in important ways mostly during the Middle English period. Incidentally, these results also confirm Haspelmath's observation that the rate of lexical replacement in language change is not necessarily constant (2003: 2).

11. Conclusions and further research

We have carried out a corpus-based diachronic analysis of the conversational core vocabulary of contemporary British English entirely based on the BNC in order to show statistics that highlight two important features which are often mentioned by historians of the English language. The first one refers to the importance of the native element in contrast with the historical role of contact and the second, to the resistance of core vocabulary to historical (internal) changes and/or contact. Such analysis is not devoid of theoretical problems of various kinds: first and foremost the decision to adhere to frequency as the only truly objective and empirical factor in the determination of core. We are well aware of the dangers of slavishly accepting word lists based on frequency. Yet, we agree with Haspelmath (2003: 2; 2004: 212), that frequency is the single most objective factor we have in the establishment of a core list, and it should certainly supersede intuitive lists diachronicians and comparative linguists have relied on in the past.

The statistics offered here, based on 700 lemmas, show that though native elements are numerically outstanding, even the most basic kind of English, such as that represented by spoken conversational interactions, cannot exist without a substantial number of French, Latin and even a few Scandinavian words or word elements. Consequently, it can no longer be accepted that this core has remained impervious to foreign, particularly non Germanic, or Romance influence. Although Contemporary English qualifies as a highly cosmopolitan language, to be fair, the role of donor languages other than Latin and French, is practically insignificant in their contribution to the core. The second aspect examined here was that of renewal. Here statistics showed that more than half of the current core has suffered replacement within historical times. Lexical renewal was particularly significant during the so-called Middle English period. Although this was expected, we have been able to show that vocabulary replacement was not only effected via importation but that word formation processes and the inclusion of items of obscure origin have also contributed to this renewal.

Further research should attempt to provide explanations for changes in the core. For instance, there are interesting lines of enquiry which seek to shed new light on old questions in historical and areal linguistics such as whether there is a neutral universal core which is truly 'culture free', or whether even the core vocabulary for each language or period is subject to cultural influence. Wierzbicka (2006) has claimed that many of the most frequent items have functions which owe their existence to what they call the "Anglo cultural script". For Goddard (2002) extremely frequent verbs in Contemporary English such as *forget, decide, understand,* etc., adjectives like *wrong, stupid,* etc., or nouns like *idea, sense,* and *reason* are part of the "Anglo cultural script", they are not neutral. If these authors are right, the development of such an Anglo cultural script would necessarily imply that different cultural scripts existed in earlier periods and would be reflected in the core vocabularies corresponding to each stage in language development.

Notes

[1]. See Fuster and Martí (2000) and Fuster (2007) for a more exhaustive examination of these and the other oppositions mentioned by Stubbs.

[2]. This is in the public domain, freely accessed from http://www.comp.lancs.ac.uk/ucrel/bncfreq/.

[3]. See http://www.natcorp.ox.ac.uk/docs/userManual/design.xml.ID=spodes. A fuller description of the contents of the spoken component is found at this site.

[4]. We shall be considering here the less technical term "word" and the more technical "lemma" as interchangeable. Corpus linguists and lexicographers often refer to lemmas in this sense of "word" as dictionary "headwords".

[5]. These same items have also been discarded from the frequency list in the preparation of the top 3000 words of the defining vocabulary in LDOCE3. It will be recalled that Hughes' diachronic research, with which interesting comparisons may be established, is based on LDOCE3 (2000: 391-4).

Works cited

ARONOFF, Mark and Kirsten FUDEMAN. 2005. *What is Morphology?* Malden Oxford/Victoria: Blackwell Publishing.

BERNDT, Rolf. 1984. *A History of the English Language*. Leipzig: VEB Verlag Enzyklopädie.

BIRD, N. 1987. "Words, Lemmas and Frequency Lists: Old Problems and New Challenges". (Parts 1& 2). *Al-manakh,* 6: 42-50.

BRINTON, Laurel J. and Leslie K. ARNOVICK. 2006. *The English language: A Linguistic History*. Oxford/New York: Oxford U. P.

CAMPBELL, Lyle. 2003. "How to Show Languages are Related: Methods for Distant Genetic Relationships". In Joseph, B. D. and R. D. Janda. (eds.) *The Handbook of Historical Linguistics*. Malden: Blackwell Publishing: 262-282.

FUSTER, Miguel. 2007. "Renewal of Core English Vocabulary: A Study Based on the BNC". *English Studies,* 88 (6): 699-723.

FUSTER, Miguel and María Mar MARTÍ. 2000. "Contact and Basic English Vocabulary". *Studies in English Language and Linguistics,* 2: 97-115.

GODDARD, Cliff. 2002. "The search for the shared semantic core of all languages". In Goddard, C. and A. Wierzbicka. (eds.) *Meaning and Universal Grammar - Theory and Empirical Findings*. Vol I. Amsterdam: John Benjamins: 5-40.

GODDARD, Cliff and Anna WIERZBICKA (eds.) 2002. *Meaning and Universal Grammar- Theory and Empirical Findings*. Vol I. Amsterdam: John Benjamins.

GRZEGA, Joachim. 2003. "Borrowing as a Word-finding Process in Cognitive Historical Onomasiology". *Onomasiology Online,* 4: 22-42.

GUDSCHINSKY, Sarah C. 1956. "The ABC's of Lexicostatistics (Glottochronology)". *Word,* 12: 175-220.

HALLIDAY, M.A.K. and Rukaya HASAN. 1976. *Cohesion in English*. London: Longman.

HASPELMATH, Martin. 2003. "Loanword Typology: Steps towards a Systematic cross-linguistic study of lexical borrowability". [http://email.eva.mpg.de/~haspelmt/LWT-text.pdf].

—. 2004. "How hopeless is genealogical linguistics, and how advanced is areal linguistics?". *Studies in Language,* 28 (1): 209-223.

HOFFMANN, Sebastian. 2004. "Are Low-Frequency Complex Prepositions Grammaticalized? On the Limits of Corpus Data and the Importance of Intuition". In Lindquist, H. & C. Mair. (eds.) *Corpus Approaches to Grammaticalization in English*. Amsterdam/ Philadelphia: Benjamins: 171-210

HORN, Ernest. 1926. *A Basic Writing Vocabulary*. Iowa City: University of Iowa.

HUGHES, Geoffrey. 2000. *A History of English Words*. Oxford: Blackwell.

JESPERSEN, Otto. 1967. *Growth and Structure of the English Language*. Oxford: Blackwell.

KILGARIFF, Adam. 1997. "Putting Frequencies in the Dictionary". *International Journal of Lexicography,* 10 (2): 135-155.

LABOV, William. 2001. *Principles of Linguistic Change: Social Factors*. Malden Massachusetts: Blackwell.

LASS, Roger. 1987. *The Shape of English: Structure and History*. London: J.M. Dent & Sons Ltd.

LEE, David. 2001. "Defining Core Vocabulary and Tracking its Distribution across Spoken and Written Genres: Evidence of a Gradience of Variation from the British National Corpus". *Journal of English Linguistics,* 29 (3): 250-278.

LEECH, Geoffrey, Paul RAYSON and Andrew WILSON. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Essex: Longman. [Also companion website http://www.comp.lancs.ac.uk/ucrel/bncfreq/].

LUTZ, Angelika. 2002. "When did English begin?". In Fanego, T., B. Méndez-Naya, and E. Seoane. (eds.) *Sounds, Words, Texts and Change*. Amsterdam and Philadelphia: John Benjamins: 145-171.

MCCARTHY, Michael. 1990. *Vocabulary*. Oxford: Oxford U. P.

—. 1999. "What Constitutes a Basic Vocabulary for Spoken Communication?". *Studies in English Language and Linguistics,* 1: 233-250.

—. 2003. "Talking Back: 'Small' Interactional Response Tokens in Everyday Conversation". *Research on Language and Social Interaction,* 36 (1): 33-63.

— and Ronald CARTER. 2003. "What Constitutes a Basic Spoken Vocabulary?".*Research Notes,* 13: 5-8 [http://www.cambridgeesol.org/rs_notes/rs_nts13.pdf].

MCMAHON, April. 2004. "Language, Time and Human Histories". In Brisard, F. S. D'Hondt and T. Mortelmans. (eds.) *Language and Revolution/Language and Time*. Universiteit Antwerpen: Antwerp Papers in Linguistics,106: 155-171.

NATION, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge U. P.

O'KEEFFE, Anne, Michael MCCARTHY and Ronald CARTER. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge U. P.

PLAG, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge U. P.

PYLES, Thomas. 1971. *The origins and development of the English language*. New York: Harcourt Brace Jovanovich.

RANKIN, Robert L. 2003. "The Comparative Method". In Joseph, B. D. and R. D. Janda. (eds.) *The Handbook of Historical Linguistics*. Malden: Blackwell Publishing: 183-213.

ROTHWELL, William. 1998. "Anglo-Norman at the (Green)Grocer's". *French Studies,* 52 (1): 1-16.

SINCLAIR, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford U. P.

STUBBS, Michael. 1996. *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*.Oxford: Blackwell.

—. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

THOMASON, Sarah G. 2001. *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.

—. 2003. "Contact as a source of language change". In Joseph, B. D. and R. D. Janda (eds.) *The Handbook of Historical Linguistics*. Malden: Blackwell Publishing: 687-712.

WEST, Michael. 1953. *A General Service List of English Words*. London: Longman.

WIERZBICKA, A. 2006. "Anglo scripts against 'putting pressure' on the other people and their own linguistic manifestations". In Goddard, C. (ed.) *Ethnopragmatics: Understanding Discourse in Cultural Context*. Berlin: Mouton de Gruyter: 31-63.

Received: 3 June 2008
Revised version: 30 September 2008