Árboles y Rizomas Vol. VI, Nº 2 (Julio-Diciembre, 2024): 14-30
Universidad de Santiago de Chile, ISSN 0719-9805
https://doi.org/10.35588/ayr.v6i2.6758

"Erm you know that thing over there…" A General Service List of Conversational English for ELT

"Eh conoces esa cosa de allí..." Una lista de vocabulario general de inglés conversacional para la enseñanza del inglés

Kevin Frank Gerigk[1]

Abstract

This study proposes and analyses a General Service List of Conversational English based on the spoken British National Corpus 2014. Most general service lists are based on written data (Brezina & Gablasova, 2015), overlooking the inherently unique features of conversations. This paper addresses this gap by presenting a general service list of conversational British English. After compilation, the data has been sifted manually to eliminate stopwords, e.g., swear words and months. The aims of this research are threefold: 1) identifying and describing the core lexis of conversations; 2) identifying the heavy-duty vocabulary necessary for full comprehension; and 3) identifying the relation between the coverage of the core lexical and the core conversational vocabulary. This study suggests heavy-duty vocabulary and provides a description of it. In particular, the results suggest that conversational English uses a high degree of recycling of core vocabulary, which is simple in structure and operates at the phatic level of interaction. These findings are particularly interesting for EFL materials design and the teaching of conversational English, enhancing the authenticity of conversations in the EFL classroom. This paper concludes with implications for the selection of vocabulary items in EFL contexts, especially with a focus on listening activities in EFL teaching materials.

Keywords: English Language Teaching (ELT), Data-Driven Learning, Conversational British English, Corpus Linguistics, EFL.

Resumen

Este estudio propone y analiza una Lista General de Vocabulario Básico de Inglés Conversacional basada en el corpus oral British National Corpus 2014. La mayoría de las listas generales de servicios se basan en datos escritos (Brezina y Gablasova, 2015), pasando por alto las características únicas de las conversaciones. Este artículo aborda esta carencia presentando una lista general de vocabulario de inglés británico conversacional. Tras su compilación, los datos se han filtrados manualmente para eliminar stopwords (palabras vacías), por ejemplo, groserías y meses. Los objetivos de esta investigación son tres: 1) identificar y describir el núcleo léxico de las conversaciones; 2) identificar el vocabulario básico necesario para una comprensión completa; y 3) identificar la relación entre la cobertura del núcleo léxico y el núcleo de vocabulario conversacional. Este estudio sugiere la existencia de vocabulario de alta frecuencia y proporciona una descripción del mismo. En concreto, los resultados sugieren que el inglés conversacional contiene un alto grado de reutilización del vocabulario básico, de estructura simple y que opera en el nivel fático de la interacción. Estos resultados son especialmente interesantes para el diseño de materiales de enseñanza de inglés como lengua extranjera (EFL) y la enseñanza del inglés conversacional, ya que

[1] Research Associate, Lancaster University, UK. Visiting Research Fellow, Aston University, UK.
Correo: k.gerigk@aston.ac.uk ORCID: https://orcid.org/0000-0001-6187-2943

Recibido: 08-07-2024 Aceptado: 29-09-2024

aumentan la autenticidad de las conversaciones. Este artículo concluye con implicaciones para la selección de elementos de vocabulario en contextos de EFL, especialmente centrándose en las actividades de comprensión oral en los materiales de enseñanza de EFL.

Palabras claves: Enseñanza del inglés como lengua extranjera, aprendizaje basado en datos, inglés británico conversacional, corpus lingüístico, ILE.

Introduction

In the field of English language teaching (ELT), the question about what vocabulary can be considered teachable and necessary for proficient users of L2 English has been dealt with extensively over the past decades (see, for example, Adolphs & Schmitt, 2003, 2004; McCarthy & Carter, 2003; McCarthy, 2006; Carter & McCarthy, 2006; O'Keeffe et al., 2007; McCarthy & Buttery, 2023). However, much of the past and current research has focussed on written language (Basturkmen, 2001; Cheng & Warren, 2007; Timmis, 2012), which may have increased the written bias in teaching materials, curriculum design and language assessment (McCarthy & Carter, 1995), with a negative effect on what is considered conversational English (Gerigk, 2024b) and potentially the learners' interaction skills (McCarthy & McCarten, 2023). As Ur (2012) describes in her experience in French, one can learn a language for years and achieve good grades; however, when prompted to converse in the L2 one may lack sufficient conversational skills. Ruhlemann (2006) states that the inherent conversational features, such as hesitations, pausing and vagueness, are often neglected and perceived as dysfluency. Such an assumption, which is common amongst teachers and examiners, is a dangerous misconception and does not reflect naturally occurring conversations. With over 40 per cent of all communication taking place orally (Bury-Allen, 1995; as cited in Miller, 2003), it is essential to teach conversational English as a distinct and crucial variation in the L2 classroom (McCarthy, 1999; McCarthy & Carter, 2003).

This paper addresses the uniqueness of conversational British English and offers a new, comprehensive view on this genre by introducing new terminology for the categories of spoken production: 1) Conversational Vocabulary and 2) Lexical Vocabulary. Both categories are marked by their contribution to either conversational routines or to maintain inter-personal relations (1) or their contribution to meaning-making and information transfer (2). The goals of this paper are three-fold: First, to analyse how much each K-Level Band contributes to the overall word count in a corpus of conversational English to identify highly frequent and hard-working vocabulary, which allows to calculate the required vocabulary size for proficient L2 conversationalists. The second goal is to analyse the contributions of each of the categories to the individual K-Levels. This may enable teachers to identify what lexical and conversational vocabulary to include in dependence of the learners' proficiency level. The third goal is to describe and summarise the nature of the words that can be found in each of the heavy-duty vocabulary categories. This may help materials writers and teachers to discern what items to include in L2 teaching to help the learners to achieve conversational proficiency. These goals align with the following three research questions that have guided this study:

> RQ 1: What overall coverage does the teachable vocabulary have across K-Level Bands, and how much does each Band add to the overall coverage?

RQ 2: What is the teachable core vocabulary of conversational British English and how is it composed in terms of items?
RQ 3: What is the teachable core lexical and conversational vocabulary of British English conversations?

By answering these research questions, this article offers guidelines as to the heavy-duty vocabulary which can be used by teachers as reference on what vocabulary to emphasise when they teach conversational English. The resulting General Service List of Conversational English (GSLCE) is available upon request and can be used by materials designers and English teachers as a reference guide for the selection of vocabulary for exercises that aim to represent casual conversations as well as a tool to check if the selected vocabulary falls into the heavy-duty band, which is assumed to constitute a core conversational lexicon. This may impact the L2 learners' exposure to high-frequency vocabulary, which consequently may increase noticing of common conversational features (e.g., pausing, vagueness, discourse marking).

The paper starts with a discussion of the use and applications of general service lists in ELT. Then, the dataset, the spoken British National Corpus 2014, is discussed in terms of balance and representativeness. Thereafter, this paper turns to a discussion of the concept of words and how to statistically count them reliably. Subsequently, the analysis tool (SketchEngine) is introduced with a novel proposal regarding the categorisation of spoken vocabulary. Then, the results regarding heavy-duty vocabulary bands are presented before concrete examples and their distribution thereof are introduced. Before conclusions and implications, the manuscript offers a discussion of the results and their relevance for the field of ELT, and particularly for materials design.

Vocabulary Service Lists of English

Research into vocabulary lists looks back at a long tradition, pioneered by West's 1953 General Service List (GSL). The original GSL is a list of 2,284 headwords based on a corpus of roughly 2.5m words of contemporary written English from the 1930s to the 1950s, which had relevance for L2 teaching (West, 1953). Whilst back in Michael West's days, the compilation of frequency lists was an arduous manual undertaking, the Corpus Revolution in the 1980s and 1990s (Leech, 2000) has allowed linguists to access larger volumes of data at a faster rate. For example, Nation's (2012) frequency list considered the 10,000 most frequent word families (e.g., headword and all their derivates) in the British National Corpus 1994 (BNC1994) (BNC Consortium, 2007) and the Corpus of Contemporary American English (Davies, 2008). Nation's list comprises around 10 per cent conversational British English and roughly 20 per cent conversational American English, making the list rather focussed on written English. Similarly, Brezina and Gablasova (2015) present their New General Service List, triangulating frequency lists from four different corpora. The data stem from the Lancaster-Oslo-Bergen Corpus, the BNC1994, the British English 2006 Corpus, and the English Web Corpus 2012. Whilst that project has contributed enormously to a better understanding of what constitutes the core English vocabulary in general, the fact that only the spoken part of the BNC1994 contributed a small sample of 10m words of spoken English from the early 1990s to the list can be deemed insufficient to draw conclusions about contemporary conversational English. The issue that emerges from the sources of English used in those studies is

that conversational English appears to be under-represented in the datasets, hence, no reliable conclusions can be drawn regarding the forms and functions of the core conversational vocabulary.

Establishing the value of conversational English, McCarthy and Carter (2003) can be seen as pioneers of vigorous and thorough research into conversational British English. Based on the CANCODE Corpus, they analysed the main features of conversational English and suggested a core vocabulary in qualitative and quantitative terms. Their findings suggest that conversational English takes advantage of only a very small number of vocabulary items, which are heavily recycled in conversation. Furthermore, the lexical and grammatical structures seem to be marked by semantic and conceptual simplicity not to cognitively overwhelm the interlocutors (McCarthy & Carter, 2003). This indicates a discrepancy between the words that are required in written and conversational exchanges. Furthermore, Carter and McCarthy (2006) in their book *Cambridge Grammar of English* dedicated two chapters on features that are typical for conversational English and unlikely to be found in written texts. This further emphasises the need for re-consideration of the status and importance of spoken English, necessitating the creation of new terminology and teaching approaches as well as more profound analysis of more recent data. This is particularly important to rid the ELT community of the misconception of dysfluency and sloppy language use, which are common descriptors of conversational English as contested by Ruhlemann (2006).

Data: The Spoken British National Corpus 2014

The GSLCE is based on the spoken part of the British National Corpus 2014 (spoken BNC2014). The spoken BNC2014 comprises approximately 11.1m lemmas of conversational English from 1,251 conversations and 672 volunteer participants, collected between 2012 and 2016 (Love et al., 2017). The spoken BNC2014 was chosen since it represents casual conversations in English between members of the British public. Whilst it is acknowledged here that this data set is focussed on British English only, it must also be accepted that the spoken BNC2014 is the most recent corpus that represents casual conversations. Hence, it can be assumed that the vocabulary used in the dataset may reliably represent authentic and real-life language use, as it would be encountered in a variety of settings by both native and non-native speakers.

Methodological Considerations

Lemmas vs Word Families

In the compilation of vocabulary lists, two overarching approaches to counting words can be identified in the literature: word families (Adolphs & Schmitt, 2003; Nation, 2012) and lemmas (Brezina & Gablasova, 2015; Webb, 2021a, 2021b). Word families usually consist of a headword, e.g. *help* (v.), and its derivates such as *helpful*, *helper*, *helped*, *helpless*, *helping* and *help* (n.). The issue with word families is two-fold. First, it assumes that the L2 learner can recognise and produce derivates only because they know or recognise the headword, which may not reflect reality (Gablasova & Brezina, 2021). Second, word families may have only limited applicability and present a skewed image of what truly high-frequency items are as they include low frequency derivates of a high frequency item (Webb, 2021a, 2021b). This may inflate a potentially low frequency item and unnecessarily skew the rank position of the headword. Hence, it has been

argued that word families lend themselves to being a good measure of the depth of receptive vocabulary knowledge in language tests (Webb, 2021a, 2021b). However, they are not the most reliable measure of the actual heavy-duty vocabulary needed for proficient conversations in the L2. Instead, Dang (2021) suggests that smaller units, e.g., lemmas, allow for a more targeted approach to vocabulary exposure and teaching for both receptive and productive skills. A lemma is a fine-grained amalgamation of a headword and its direct inflections (Brezina & Gablasova, 2015; Webb, 2021a, 2021b). This means that the lemma *good* (adj.) includes *better* and *best* but not the noun *good(s)* as would be the case for word families. For verbs, the lemma includes the basis and 3rd person forms, e.g., *run* and *runs*, but not different tenses. In the case of nouns, only singular and plural forms are shown in the lemma, e.g., *house* and *houses*. Hence, lemmas enable us to narrow down the items included in a headword and can eliminate some degree of assumption about vocabulary depth. Therefore, in this paper, *lemmas* and *words* are used synonymously.

Average Reduced Frequency vs Absolute Frequency

In the compilation of frequency lists, two statistical options seem plausible: Absolute Frequency (AF) and Average Reduced Frequencies (ARF). Whilst AF only counts the total occurrences of a word in a corpus, ARF also considers the distribution of a word across the dataset. This is important since a word may occur particularly frequently in a small share of specialised or topical conversations (e.g., a visit to the garage may elicit the word *engine* or *car* more often than they would usually occur in a conversation). Hence, the words *engine* or *car* would be assigned a much higher frequency than it is reflected in reality. When looking at dispersion in those cases, we can level down the indicated importance of words that are in general very frequent but dispersed unevenly across a few files in the corpus (Kilgariff, 1997). A solution to this was offered by Savicky and Hlavacova (2002), who incorporated a dispersion measure into the frequency statistics. ARF has been discussed and used successfully for the purpose of compiling Brezina and Gablasova's (2015) New General Service List. Following on from Brezina and Gablasova (2015), the choice was made to use ARF for this study, too. The advantages of this statistical measure are threefold: First, ARF may be able to distinguish between highly frequent and unevenly dispersed words and highly frequent and well-dispersed words across a corpus. Second, by considering the frequency and dispersion of a word, ARF helps to avoid inflation of unevenly dispersed vocabulary and eliminate false positives (Brezina, 2018). Third, the results yielded may offer a more realistic picture of what vocabulary items are used frequently in authentic, casual conversations in English.

Methodology

The spoken BNC2014 was accessed via https://www.sketchengine.eu/ (Kilgarriff et al., 2014). The Word List function in SketchEngine was used for the compilation of a frequency list, following the parameters below:

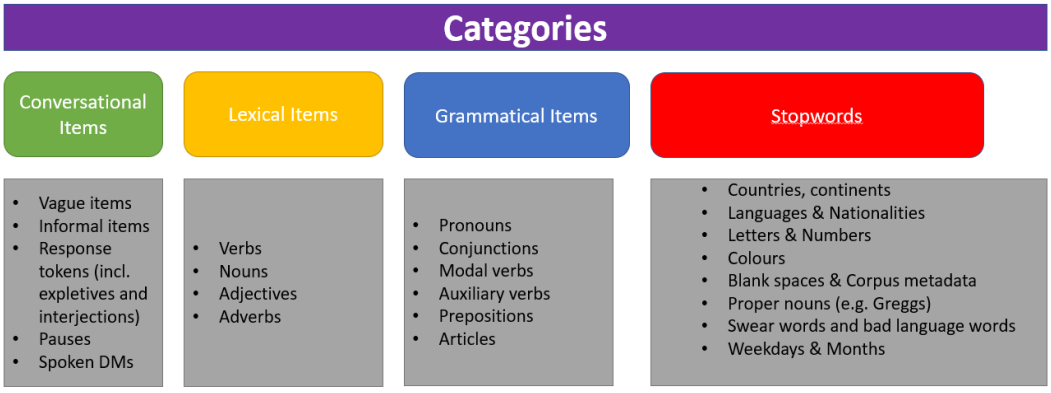
- Word Type: lemma
- POS Tag: any
- Case Sensitivity: "A=a"

To limit the turn-out of results in a sensible manner and to avoid extremely rare lemmas, the minimum occurrence of each word was set to 5.

SketchEngine subsequently returned a frequency list of 11.1m lemmas, ranked by ARF. As suggested by O'Keeffe et al. (2007), to achieve a sound comprehension of English, which is somewhere around the B2 Level of the CEFR (Council of Europe, 2020), the first 6,000 words (K6-Level) in a frequency list are the most relevant. However, Nation (2006), suggesting that for full comprehension 98 per cent of the words should be familiar items, claims that this level of coverage could be achieved within the 9,000 most frequent words. He recognises, too, that this figure might be lower for spoken English. Accordingly, the frequency list was capped at the K9-Level. Following Brezina and Gablasova's (2015) recommendation, the first 2,000 words have been subdivided into four K-Levels (K0.5 – K2), to allow for a more detailed analysis of vocabulary contribution per K-Level. The distribution observes Zipf's Law (Brezina, 2018), where the second most frequent word occurs only half as often as the most frequent word, and the third most frequent word occurs half as much as the second most frequent item and so on. At around K2-Level, this phenomenon evens out, which indicates a heavy-duty core vocabulary around this cap. Therefore, the K2-Level is subdivided into 500-word steps, before analysis continues in 1,000-word steps. The frequency list was thoroughly sifted, and vocabulary items were categorised manually into the categories proposed in Figure 1 below.

The first category is labelled Conversational Items. Based on Carter and McCarthy's (2006) comprehensive description of spoken English, this category does not fulfil the function of information transfer but is used for bonding and emotional expressions. Furthermore, Conversational Items are strongly context-dependent and a sign of proximity between the interactants. This category comprises Vague Items (e.g., *like*), Informal Items (e.g., *gonna*), Response Tokens (including Expletives and Interjections, e.g., *mm, wow*), Filled Pauses (e.g., *erm*) and spoken Discourse Markers (e.g., *well*).

Figure 1. Categorisation of vocabulary items.



Source: Own Elaboration.

The second category refers to Lexical Items. According to Gilmore (2004) those content words convey information about a subject. Lexical Items may be equally frequent in written and spoken English; however, the nature and sophistication may differ noticeable in spoken language

(McCarthy & Carter, 2003). The Lexical Items in this category comprise Verbs (e.g., *go*), Nouns (e.g., *tree*), Adjectives (e.g., *big*) and Adverbs (e.g., *quickly*). This category provides content information, i.e., what is talked about. The third category is called Grammatical Items. These items can be considered as the glue of the utterances, rendering them coherent and relating the utterances to one another. This category comprises Pronouns (e.g., *it*), Conjunctions (e.g., *but*), Modal Verbs (e.g., *can*), Auxiliary Verbs (e.g., *be*), Prepositions (e.g., *at*) and Articles (e.g., *the*). The last category of vocabulary is called Stopwords. They are items that can be deemed inappropriate for learners, e.g., swearwords, or those words that should be covered by a structured syllabus regardless of their frequency, e.g., months. The following Stopword sub-categories have been applied to the dataset:

- Countries and Continents
- Languages and Nationalities
- Letters and Numbers
- Colours
- Blank Spaces and Corpus Metadata
- Proper Nouns
- Weekdays and Months
- Swearwords and Bad Language Words

Findings

Overall Coverage across K-Levels and their contribution for Comprehension

This section refers to RQ 1: What overall coverage does the teachable vocabulary have across K-Level Bands and how much does each Band add to the overall coverage? As can be seen in Figure 2 below, the data can be sub-divided into 3 levels of contribution: heavy-duty, or core, vocabulary (red), high-frequency vocabulary (yellow) and low-frequency vocabulary (green). The data, following Zipf's law, shows a sharp fall of frequencies between K1 and K2 Level Bands. It can be observed that the 1,000 most frequent items consist of a recycling of 9,901,655 lemmas, which contribute 93.2 per cent of all the lemmas (n_{total}=11.1m) in the corpus. Another 3 per cent are contributed by the K-2-Level Band, already indicating a sharp decline of contribution. Hence, K-2-Level items are categorised as Core Vocabulary, contributing as much as 96.2 per cent of the vocabulary that a learner is likely to encounter in conversational English.

Figure 2. Coverage of and contribution by teachable vocabulary across K-9-Level Bands.

K-Level Band	Absolute Frequency	Coverage of total (%)
K1.0	9,901,655	93.2
K2.0	311,993	3.0
K3.0	136,374	1.3
K4.0	78,629	0.7 threshold: 98.2
K5.0	48,226	0.5
K6.0	33,395	0.3
K7.0	23,699	0.2
K8.0	18,274	0.17
K9.0	13,614	0.13

Source: Own Elaboration.

Following Zipf's Law from this point, the contribution made by the K-3-Level Band is reduced by roughly half of the contribution of the previous Band (3.0 per cent vs 1.3 per cent). Whilst lower, this figure can still be seen to contribute a considerable amount of lexis to the conversation, which peaks at the K-4-Level Band, where the threshold of 98 per cent contribution is reached. Familiarity with the 4,000 most frequent items is sufficient to achieve 98.2 per cent of coverage, i.e., it may allow full comprehension of conversational English (O'Keeffe et al., 2007). The contribution of the remaining K-Level Bands continues to decrease roughly in line with my expectations according to Zipf's Law and changes again beyond K-7-Level Band. This suggests that the high frequency items can be found within the K-3 and K-7-Level Bands, contributing a total of 3 per cent to the total coverage of the corpus. It is also within that range that sufficient coverage is achieved to ensure full text comprehension (K-4-Level Band). The contribution made by the K-8 and K-9-Level Bands is less than 0.2 per cent. That is where the mark for low frequency vocabulary begins. The input and usefulness of those K-Level Bands is still high for continued proficiency. It can be argued, however, that these K-Level Bands already tap into the realm of more specialised vocabulary and may no longer constitute general language.

Teachable Core Vocabulary and Contribution per category

In this section, the findings regarding RQ 2 are presented: What is the teachable core vocabulary of conversational British English and how is it composed in terms of items? As indicated in the previous section, the first 2,000 items can be deemed Core Vocabulary, offering the greatest contribution to the dataset. As can be seen in Table 1 below the greatest contribution of all K-Level Bands to the vocabulary comes from the K-0.5-Level Band with a total of 9,461,597 lemmas, which corresponds to a coverage of 89.1 per cent. Interestingly here, 31.6 per cent of the contribution is made by lexical items and a further 16.4 per cent is contributed by the category of Conversational Items. The remainder of contribution is made by the excluded categories of Grammatical Items and Stopwords.

Table 1. Results of Composition of Core Vocabulary in K-2-Level Bands

K0.5		
	AF (lemma)	%
Lexical Items	3,352,858	31.6
Conversational Items	1,743,803	16.4
Total (incl. Stopwords and Grammatical Items)	9,461,597	89.1
K1.0		
	AF (lemma)	%
Lexical Items	386,929	3.6
Conversational Items	41,178	0.4
Total (incl. Stopwords and Grammatical Items)	440,058	4.1
K1.5		
	AF (lemma)	%
Lexical Items	182,150	1.7
Conversational Items	12,146	0.1
Total (incl. Stopwords and Grammatical Items)	197,282	1.9
K2.0		
	AF (lemma)	%
Lexical Items	107,778	1.0
Conversational Items	5,652	0.005
Total (incl. Stopwords and Grammatical Items)	114,711	1.1

Source: Own Elaboration.

Whilst considerably smaller, the next sub-level, the K-1.0-Level Band contributes a total of 440,058 lemmas to the total lemma count of the corpus. This corresponds to 4.1 per cent contribution to conversational vocabulary. A total of 3.6 per cent are contributed by Lexical Items, with a mere 0.4 per cent contribution from Conversational Items. As this trend is to continue, it can be suggested that the K-0.5-Level Bands is the richest in terms of Conversational Items, with subsequent K-Level-Bands only making a marginal contribution. The same decline in contribution can be observed in the two excluded categories, which only contribute 0.1 per cent to this sub-level combined. Similarly, the K-1.5-Level Band contributes a total of 1.9 per cent of the vocabulary to the entire corpus. This is a total of 197,282 lemmas and composed of 1.7 per cent of Lexical Items and 0.1 per cent of Conversational Items. Like in the previous K-Level, the excluded categories amount to a mere 0.1 per cent of contribution made by this K-Level. The smallest contribution, which can still be deemed essential, can be found at the threshold of the K-2.0-Level Band. The total contribution to the corpus is 1.1 per cent, with 1.0 per cent from Lexical Items, 0.005 per cent from Conversational Items and 0.95 per cent from the two excluded categories. Based on the data, two overarching observations can be made: 1) The Core Vocabulary is primarily found in the K-

0.5-Level Band, and 2) Within the Core Vocabulary, Conversational Items are extremely frequent and make up a huge part of conversational vocabulary.

Composition of Core Vocabulary: What is it in concrete terms?

The following addresses RQ 3: What is the teachable core lexical and conversational vocabulary of British English conversations? Subsequently, a summary of the nature of the vocabulary in the two focus categories of Core Lexical Vocabulary and Core Conversational Vocabulary is presented. It must be borne in mind that this qualitative analysis refers to the K-2-Level only, which has been identified as the Core Level.

Core Lexical Vocabulary

The core lexical vocabulary is composed of verbs, nouns, adjectives and adverbs. In terms of the verbs that can be found on that K-Level, it can be observed that they refer to very simple cognitive processes and actions. Some of the verbs from the list are *look, get* and *want*. This suggests that core verbs carry only little to no abstraction and show a low level of complexity. Similarly, nouns have been found to refer to simple concepts or objects within the context of conversation. The most-frequent nouns are *time*, *people* and *work*. Again, these nouns are marked by their simplicity and shortness. It can also be said that these nouns refer to everyday concepts, increasing their familiarity and usefulness in conversation. Along the same lines, it can be said that adjectives, for example *big, long* and *first*, are used to realise non-complex descriptions of subjective physical appearance or quality of a concept or object. Whilst a thesaurus would offer (and an examiner might expect) more eloquent ways of describing one's surroundings, the data show that when we converse, simplicity may win. Additionally, it can be observed that adverbs follow suit: with representatives such as *now, very* and *always*, the core adverbs function as simple intensifiers or descriptive-evaluative elements. In this function, they commonly refer to frequency or intensity of an action.

Core Conversational Vocabulary

It has been suggested by the previous sub-section, that especially the K-0.5-Level Band makes a huge contribution in terms of Conversational Items. The following data presents items from that K-Level only as it is there where the richest data can be found. The Core Conversational Items include Vague Items, such as *like* and *just*. These lemmas are often employed to hedge or down-tone a statement, crucial for face-keeping. Furthermore, Informal Items, such as *cos* and *gonna* can be found in this K-Level. These items can be seen as markers of informal language and consist of a contraction of either one word or a phrase. A substantial contribution is made by Response Tokens, e.g., *yeah* and *mm*. These tokens are issues by the active and engaged listener in a conversation (Gerigk, 2024a). In addition, Vocalised Pauses, such as *erm* and *um*, can be found in this dataset as markers of cognitive processes of both speaker and listener. Although often described as signs of dysfluency (Ruhlemann, 2006), they act as turn-holders and give the speaker time to cognitively process what to say next. Similar in function are Spoken Discourse Markers, like *actually* and *well*. Although the concept of Discourse Markers is not necessarily exclusive to

spoken language, the function and forms are slightly different. Here, Spoken Discourse Markers also aide in creating time for cognitive processing what to say next; however, they also mark the emotional stance of the speaker and function as the glue for cohesive and coherent utterances. In summary, it can be said that the Core Vocabulary, both lexical and conversational, can be described as follows: The Core Vocabulary is marked by utterances that fulfil a strong phatic and discourse-organising function (i.e., the Core Conversational Vocabulary) as well as by its simplicity and brevity of words, which contributes to intelligibility (i.e., the Cover Lexical Vocabulary).

Discussion

This section discusses the implications for ELT teaching, in particular listening and speaking, of the above findings. It must be acknowledged that the dataset, however, is relatively small (11.1m lemmas) and the variation of English is limited to British English and everyday informal conversation amongst friends and family only. Hence, results may slightly differ when considering other variations. This, however, does not devaluate the currency of the results. This study has shown that there may, indeed, be a mismatch between suggestions based on written English that cannot easily be transferred upon spoken English. This is particularly obvious in K-Level coverage (O'Keeffe et al., 2007; Nation, 2012). As shown in the findings section, a 98 per cent coverage of the corpus, which has been suggested as the threshold for full comprehension (Nation, 2012) can be achieved within the K-4-Level Band in spoken English. This figure lies considerably below what Nation (2012) and O'Keeffe et al. (2007) suggest, namely K-9 and K-6-Level vocabulary, respectively. Furthermore, it can be seen that the heavy-duty vocabulary resides within the K-2-Level Band with a contribution of 96.2 per cent to the total lemma count. These findings are rather similar to Adolphs and Schmitt's (2004), which suggest that the K-2 Level Bands contributes up to roughly 94 per cent of the total vocabulary in spoken discourse. Beyond this K-Level Band, any add-on is minimal. These findings suggest that speaking and listening exercises in ELT may not benefit from complex vocabulary. Instead, the teacher may want to consider focussing on a limited set of vocabulary and the incorporation of conversational routines and their functions as realised by the Conversational Items. This may facilitate learning interaction in the L2, with the learner feeling less overwhelmed when starting to speak in the L2 and perceived fluency may be reached sooner (which may be particularly true for low level learners). One might argue that the overall aim of language learning is expansion of the vocabulary and grammatical knowledge. Whilst I agree with such a position, it must not be overlooked that spoken English is fairly simplistic in terms of both grammar and lexis. However, Conversational Items which are essential for communicative routines, bonding and sharing the same immediate context may be crucial for the development of communicative competence in the L2. This, in turn, drives the main purpose of oral communication and Communicative Language Teaching, namely, to get one's message across, i.e., to be understood and to understand (Richards & Rodgers, 2014).

This simplicity can also be observed in the contribution of each K-Level to the overall corpus and the fine-tuned staging of the first 2,000 words on the frequency list. The largest contribution to the corpus is made by recycling the items from the K-0.5-Level Band (89.1 per cent). When having a closer look at the items in question here, one can see that a large proportion of those items refer to simple Lexical Items and Conversational Items. This allocates a lot of importance to vocabulary in this Band and supports Ruhlemann's (2006) argument, in which he

calls for new terminology for spoken English. In conjunction with this finding, the nature of the Core Vocabulary is interesting and holds important implications for L2 teaching. This paper proposes four categories to distinguish the types of vocabulary in this frequency list: Conversational Items, Lexical Items, Grammatical Items and Stopwords. The sheer number of Conversational Items and their decreasing contribution to the total lemma count in ascending K-Levels indicates that they are an integral part of conversation and essential to be known by the L2 learner. If the learners were unaware of those forms and their functions in conversation, they would lose access to approximately one fifth of what is uttered. Furthermore, if 98 per cent familiarity with the vocabulary is the goal, the lost 16.4 per cent from excluding Conversational Items such as Response Tokens would be replaced by less-frequent vocabulary, hence mis-representing the language that is used and needs to be learnt. This may lead to the L2 learner being overwhelmed and less-well prepared for oral communication. For example, without awareness of conversational strategies like Filled Pauses, the learners might believe that immediate responses are required of them, or they may lose their turn to the interlocuter since no turn-holding device indicates their willingness to continue. Equally, the absence of Response Tokens may lead to the listener being perceived as cold or disinterested (Gilmore, 2004). Since these features are often perceived as dysfluency or sloppy language use (Ruhlemann, 2006), increased awareness about the naturalness of those features is crucial. This is why I suggest the umbrella term Conversational Items alongside established terms such as Lexis and Grammar. This new terminology has been borne out of the sheer necessity to rid ourselves as language teachers and our learners from unnecessary pressure and unrealistic expectations in conversational L2 language production and comprehension. This transpires in the description of the Lexical Items category. As indicated above, spoken language is less complex and is marked by simple concepts (McCarthy & Carter, 2003). Instead, nouns and verbs in conversations describe simple processes and refer to simple concepts, often with reference to their immediate context. This may be the result of reduced planning time in spontaneous interaction and needs to be addressed in L2 teaching. Complex linguistic structures may even overwhelm native speakers and, hence, should not be expected from learners. In addition to this, adjectives and adverbs seem to occur in simple forms, too. In short, it can be said that the Core Lexical Items stand out by their non-complexity and seem to aim at getting one's idea across without overwhelming neither speaker nor listener in the conversation (Brunfaut & Revesz, 2015). This is an important strategy for face-keeping and avoiding conversation breakdown or exhaustion.

In essence, this paper sides with the suggestions and claims made in previous research (see McCarthy & Carter, 2003; Carter & McCarthy, 2006; Ruhlemann, 2006). The study presented here has shown that spoken conversational vocabulary follows its own internal and external structures. Whilst it is still the English language, the conversations stand out by items that realise conversation routines and phaticity. Hence, teachers and stakeholders in L2 educational programmes may benefit from a clearer understanding of conversational patters and functions, which have been introduced in this paper. This may encourage and promote the development and implementation of new teaching technologies, which allow the L2 learner to access conversational English from naturally occurring conversations instead of scripted dialogues or monologues as they are still common in mainstream textbooks (Gerigk, 2024b). One approach to exposing L2 learners to authentic conversational English could be the introduction of applied corpus linguistics into the L2 classroom (Timmis, 2015; O'Keeffe, 2021; Arellano & Gerigk, 2023) or the creation of corpus-informed teaching materials through material writers and the teachers themselves. This may shift the focus

away from 2012oral language production that resembles a well-composed text, and instead put the focus on communication and interaction. Therefore, I suggest that spoken language needs to be presented, theorised and taught in a way that recognises its unique features and allows for a focus on interaction, necessitating changes in teaching curricula to avoid unnecessary complexity and confusion with similarities to its written sibling (McCarthy & McCarten, 2023). This may allow teachers and stakeholders, including the learners, to focus more explicitly on conversational patterns and meaning-making through interactional conventions. Being aware of this aspect of English may equip learners with a more realistic expectation as to what to expect when travelling to an English-speaking country and having conversations with proficient speakers there, whilst it simultaneously reduces the burden of studying a highly artificial variation of the English language, which Roemer (2006; 2023) calls 'school English'. This paper has offered insights into what authentic conversational English entails in terms of vocabulary. This data may help to refocus practices in the field of L2 English teaching, assigning a more practical focus to the field of conversations, which may lead to more confidence and confluency in the L2 learners.

Conclusion

This paper has presented and analysed the General Service List of Conversational English. The motivation behind this study resulted from a lack of research in the area of teaching conversation from an interactional viewpoint. Whilst a considerable amount of research considers mixed-mode and written English language (e.g., Brezina & Gablasova's (2015) New General Service List), a General Service List of Conversational English had not yet been composed. This paper also consolidated previous suggestions in terms of word form and has produced a strong argument in favour of lemmas (as suggested by Dang, 2021). Following on from the insightful results produced by Brezina and Gablasova (2015), this study has adopted their statistical considerations and further supported the argument of a combined measure of frequency and dispersion, using ARF, for the ranking of the results of the corpus query. The result is the 9,000-word GSLCE. The GSLCE is available for research and teaching purposes upon request. The novelty of this contribution is that the word list has been sub-divided into vocabulary categories to inform language educators and stakeholders as well as policy-makers in the field of ELT. The categories Conversational Items and Lexical Items received particular attention in data analysis, showing that their main contribution lies in highly frequent and little complex vocabulary, which appears to be recycled a lot. In contrast with written English, its spoken sibling stands out by simplicity and collaboration between the interlocutors. Furthermore, it has been established that, siding with O'Keeffe et al. (2007) and Nation's (2012) suggestion to use 98 per cent of coverage for full comprehension, the 4,000-most-frequent items in that list are sufficient to gain this level of comprehension. This comes with several implications for ELT: 1) when teaching spoken language, the focus should lie on communication strategies and simple lexis. 2) Contrived, somewhat artificial examples of language may not present the features necessary for learners to acquire and may neglect them of the opportunity to notice important aspects of oral communication such as response tokens. 3) Due to the uniqueness of spoken language, a separate curriculum and syllabus for ELT teaching is recommended, which is more closely aligned with reality. In this vein, corpus linguistics and corpus-informed or corpus-based materials such as BNClab (Brezina et al., 2018) may prove to be an invaluable teaching resource.

It must be acknowledged that this study is based on British English alone and particularly in terms of word forms, may not capture the whole picture of spoken Englishes (e.g., American English or non-native English varieties) that exists. Hence, I suggest and strongly encourage further research in this field that involves more corpus data from different variations of English (e.g., including the CANCODE, Wellington Corpus, Santa Barbara Corpus, Limerick Corpus of Irish English).

This paper concludes by strongly encouraging research into other genres of spoken English, e.g., monologues, lectures and service encounters, which may be able to prove invaluable for specific fields of ELT such as English for Specific or Academic Purposes. A triangulation of other forms of oral communication, including more Englishes, may lead to a more generalisable picture of the core vocabulary of the English language used by a multitude of stakeholders globally. Furthermore, an expansion of source corpora may allow the creation of genre specific sub-lists for distinct speech acts (e.g., explanations), conversational contexts (e.g., classroom group discussions) and genres (e.g., presentations).

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438.

Adolphs, S., & Schmitt, N. (2004). Vocabulary coverage according to spoken discourse context. *Vocabulary in a second language*, *10*, 39-49.

Arellano, R., & Gerigk, K. (2023). Insights from corpus linguistics: Using keywords-in-context to teach and assess English learning. *English Australia Journal*, *39*(2), 66-73. https://search.informit.org/doi/10.3316/informit.383511093653714

Basturkmen, H. (2001). Descriptions of spoken language for higher level learners: The example of questioning. *ELT Journal*, *55*(1), 4-13. https://doi.org/10.1093/elt/55.1.4

BNC Consortium (2007). *The British National Corpus. XML Edition*. Literary and Linguistic Data Service. http://hdl.handle.net/20.500.14106/2554

Brezina, V. (2018). *Statistics for Corpus Linguistics—A practical guide*. Cambridge University Press.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1–22. https://doi.org/10.1093/applin/amt018

Brezina, V., Gablasova, D., & Reichelt, S. (2018). *BNClab* [Electronic resource]. http://corpora.lancs.ac.uk/bnclab/search

Brunfaut, T., & Revesz, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, *49*(1), 141–168. https://doi.org/10.1002/tesq.168

Carter, R., & Mccarthy, M. (2006). *Cambridge grammar of English—A comprehensive guide—Spoken and written English grammar and usage*. Cambridge University Press.

Cheng, W., & Warren, M. (2007). Checking understandings: Comparing textbooks and a corpus of spoken English in Hong Kong. *Language Awareness*, *16*(3). https://doi.org/10.2167/la455.0

Council of Europe. (2020). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Council of Europe Publishing.

Dang, T. N. Y. (2021). Critical commentary: Selecting lexical units in wordlists for EFL learners. *Studies in Second Language Acquisition*, *43*, 954–957. https://doi.org/10.1017/S0272263121000681

Davies, M. (2008) *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Gablasova, D., & Brezina, V. (2021). Critical commentary: Words that matter in L2 research and pedagogy—A corpus-linguistics perspective. *Studies in Second Language Acquisition*, *43*, 958–961. https://doi.org/10.1017/S027226312100070X

Gerigk, K. F. (2024a). A Corpus Pragmatics Approach to Active Listenership in Conversations in British English: Contextualised Function-To-Form. *Corpus Pragmatics*. https://doi.org/10.1007/s41701-024-00173-2

Gerigk, K. F. (2024b). *Active listenership in EFL coursebooks in Germany and Austria: a context-driven corpus pragmatics enquiry.* [Doctoral Thesis, Lancaster University]. Lancaster University. https://doi.org/10.17635/lancaster/thesis/2308

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal*, *58*(4), 363–374. https://doi.org/10.1093/elt/58.4.363

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, *10*(2), 135–155.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9

Leech, G. (2000). Grammars of spoken English: new outcomes of corpus-oriented research. *Language Learning, 50*(4), 675–724. https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xpapers/Leech_spoken_grammar.pdf

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The spoken BNC2014—Designing and building a spoken corpus of everyday conversation. *International Journal of Corpus Linguistics*, *22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

McCarthy, M., & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal, 49*(3), 207–218.

McCarthy, M. (1999). Is there a basic spoken vocabulary: Technology and common sense. *The Journal of TESOL France*, 29–36

McCarthy, M., & Carter, R. (2003). What constitutes a basic spoken vocabulary. *Research Notes*, *3*(13), 5-7.

McCarthy, M. (2006). Discourse. In R. Carter & D. Nunan (Eds.), *The Cambridge Guide to Teaching English to Speakers of Other Languages* (pp. 48–55). Cambridge University Press.

McCarthy, M., & Buttery, P. (2023). Lexis in spoken discourse. In M. Handford & J. P. Gee (Eds.), *The Routledge Handbook of Discourse Analysis* (pp. 391–410). Routledge.

McCarthy, M., & McCarten, J. (2023). Speaking and listening: Two sides of the same coin. In K. Harrington & P. Ronan (Eds.), *Demystifying corpus linguistics for English language teaching* (pp. 59–77). Springer.

Miller, L. (2003). Developing listening skills with authentic materials. *ESL Magazine*, *6*(2), 16–18.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82.

Nation, P. (2012). *Learning vocabulary in another language*. Cambridge University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom—Language use and language teaching*. Cambridge University Press.

O'Keeffe, A. (2021). Data-driven learning, theories of learning and second language acquisition. In G. Mark & P. Perez-Parades (Eds.), *Beyond Concordance lines—Corpora in language eduaction* (pp. 35–55). John Benjamins Publishing Company.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge University Press.

Roemer, U. (2006). Looking at looking: Functions and contexts of progressives in spoken English and 'school' English. In A. Renouf & A. Kehoe (Eds.), *The changing face of corpus linguistics* (pp. 231–242). Brill.

Roemer, U. (2023). Usage-based approaches to second language acquisition vis-à-vis data-driven learning. *TESOL Quarterly*, *0*(0), 1-11. https://www.researchgate.net/publication/375571518_Usage-Based_Approaches_to_Second_Language_Acquisition_Vis-a-Vis_Data-Driven_Learning

Ruhlemann, C. (2006). Coming to terms with conversational grammar—'dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics*, *11*(4), 385–409.

Savicky, P., & Hlavacova, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Timmis, I. (2012). Spoken language research and ELT: where are we now? *ELT Journal*, *66*(4), 514-522. https://doi.org/10.1093/elt/ccs042

Timmis, I. (2015). *Corpus Linguistics for ELT*. Routledge.

Ur, P. (2012). *A course in English Language Teaching*. Cambridge University Press.

Webb, S. (2021a). Critical commentary: The lemma dilemma—How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition*, *43*, 941–949. https://doi.org/10.1017/S0272263121000784

Webb, S. (2021b). Critical commentary: Word families and lemmas, not a real dilemma—Investigating lexical units. *Studies in Second Language Acquisition*, *43*, 973–984. https://doi.org/10.1017/S0272263121000760

West, H. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longman.