DOI: 10.20472/TEC.2017.004.009

AYSEL ŞAHIN KIZIL
Fırat University, Turkey

SPOKEN CORPORA AND CORPUS-INFORMED LANGUAGE PEDAGOGY: IMPLICATIONS ARISING FROM RESEARCH

Abstract:

Defined in its broadest sense as large databases illustrating actual language use, corpora have proved to be influential in enabling researchers to develop innovative perspectives not only in linguistics but also in a number of applied disciplines including speech recognition or machine translation. One discipline on which language corpora have tremendous effect is the field of language teaching. Although the research on corpus-informed language pedagogy is mostly dominated by the findings through the analysis of written corpora, it is now widely acknowledged that spoken corpora which are slower to emerge compared with written corpora could also offer great potential for language pedagogy. This study sets out to review the major findings from the research on spoken corpora and current instructional treatments with the purpose of discussing the ways of expanding spoken-corpus-informed pedagogy to language classrooms.

Keywords:

English language teaching, corpus-informed pedagogy, pedagogical implications, spoken corpora

JEL Classification: I29

1 Introduction

The use of corpora within the field of language education is not a recent application which dates back to 1970s when the early reference corpora such as Brown Corpus of American English and the LOB (Lancaster–Oslo–Bergen) Corpus of British English emerged (Chambers, Farr, & O'Riordan, 2011). Initial steps in including language corpora into language teaching practices were in the form of teaching materials mostly motivated to find out "what language facts of relevance to language learning and teaching can be derived from corpora" (Bernardini, 2004 p.16). Language learning textbooks (McCarthy, McCarten, & Standtford, 2005), dictionaries (Sinclair, 1987) and more recently grammars (Carter & McCarthy, 2006) which were produced based on various language corpora represent the early attempts in integrating actual language use into language teaching materials.

This increased interest, though considered to be underestimated (O'Keeffe & Farr, 2003), have also been justified through a number of publications comparing the language presented in the teaching materials with actual language uses through corpora. Boxer and Pickering (1995), for example, investigated language units representing speech acts in textbook dialogues and compared them with a language corpus. They have found that functions of the language units in the textbooks are underrepresented. Likewise, Carter (1998) focused on the dialogues in textbooks by comparing them with the dialogues from the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), and found that textbooks do not provide students with such basic spoken language features as discourse markers, hedges or ellipsis. Using the same corpus (i.e. CANCODE) Hughes and McCarthy (1998) analysed the use of past perfect verbs and reached the conclusion that those verbs have more complex functions in spoken language than described in the textbooks.

Apart from the studies showing the significance of corpus-informed materials, a number of investigations into the use of language corpora by individual teachers and researchers have been conducted starting from the early 1990s (Chambers et al., 2011). Research in this strand has revealed that integrating language corpora directly into the language activities could be beneficial in many ways ranging from encouraging discovery learning, raising language awareness to boosting facilitated learning (Chambers, 2005; Wichmann, Fligelstone, McEnery, & Knowles, 1997).

However, a closer look at the relevant literature demonstrates that most of the studies in this line have relied on the use of written language corpora, which results in a relative neglect of spoken corpora despite its great potential for language pedagogy. Most of the existing studies on spoken corpora have centred on language description illustrating how certain linguistic devices are utilized in spoken language either by native speakers (NS) (Biber, Johansson, Leech, Conrad, & Finegan, 2007) or second language (L2) learners (Aas, 2011; Şahin Kızıl & Kilimci, 2014). These studies doubtlessly contributed to our understanding of the differences between spoken and written language to a great extent. Yet, the field is still in need of invigorating attempts to transfer the pedagogical potential of the findings of the existing research (Caines, McCarthy, & O'Keeffe, 2016).

This paper aims at reviewing the relevant literature on spoken corpora with a focus on pedagogical implications arising from the research. Specifically, the following research question has guided the present study:

- What could be learned from current research on the use of spoken corpora in language learning and teaching?

Subsequent to the brief introduction above, this paper starts with the definition and types of spoken corpus, which is followed by a section on major findings from research into spoken corpora. It ends with elaborating on major pedagogical implications.

2 Spoken Corpora: An Overview

In its general sense, spoken corpora refer to the language databases that include recorded and transcribed versions of speaking. Baker, Hardie and McEnery (2006) define spoken corpus as "a corpus consisting entirely of transcribed speech. This could be from a range of sources: spontaneous informal conversations, radio phone-ins, meetings, debates, classroom situations etc." (p.148). In the relevant literature, two terms are used to describe the recorded speaking (i.e. speech corpora and spoken corpora), which should be distinguished from each other. Differently from spoken corpus, a speech corpus refers to the recordings of speaking, usually made in a studio and consisting unnatural language in some cases, with the purpose of investigating pronunciation and other phonetic features (Baker et al., 2006; Caines et al., 2016).

Historically, development of spoken corpora dates back to 1970s when London-Lund Corpus emerged. Over the past few decades, the projects on compilation of spoken corpora have accelerated and a number of spoken corpora have been put at the disposal of researchers. The Lancaster/IBM Spoken English Corpus (SEC, 1992), the Wellington Corpus of Spoken New Zealand English, (1998) the Cambridge and Nottingham Corpus of Discourse in English (CANCODE, 1997), the Michigan Corpus of Academic Spoken English (MICASE, 2002) and the Vienna-Oxford International Corpus of English (VOICE, version 1.0 online, 2009) are the some among many others (O'Keeffe, Mccarthy, & Keeffe, 2010).

Depending on the goals of compilation, spoken corpus could be categorized into two groups as large-scale corpora and specialized or domain-specific corpora (Caines et al., 2016). Large-scale corpora were initially developed as a component of much larger written corpora. This is because collecting spoken data and transcribing them requires dedication of much more time and academic labor compared with written data. One example of this kind of large-scale corpora is British National Corpus (BNC), which is made up of 100 million words of data with the spoken component representing on 10% of the whole corpus. The source of these 10 million words of spoken language is the conversations by native speakers of English chosen considering a demographic balance (Fitzpatrick, 2007). Another example of large-scale corpora is International Corpus of English (ICE). As its name suggests, ICE is among the earlier corpus projects at an international level comprising spoken language by participants from 18 different countries where English functions either as the native language or official language. ICE includes a total of 1 million words with 90% of which illustrating face-to-face informal interactions (Caines et al., 2016). The largest spoken corpus that is worth to mention is Corpus of Contemporary American English (COCA) whose spoken component includes 85 million words from conversations in TV and radio programs. Although COCA is the largest available spoken corpus, lack of spontaneous face-to-face conversation remains a limitation in terms of exploiting all the potentials a corpus could provide (Caines et al., 2016).

Regarding the language of focus, most of the spoken corpus projects are dominated by English language, yet, there are emerging projects of corpus compilation involving other languages such as German, Mandarin Chinese, French, Turkish, among others (Caines et al., 2016; Ruhi, 2011).

Alongside the aforementioned large-scale corpora, researchers also devoted time to compile spoken corpora that could serve for specialized purposes. Particular research need and particular context of use are the distinguishing aspects of these domain-specific corpora. They generally contain data around 1 million words (Caines et al., 2016). Two examples of this type are Michigan Corpus of Spoken Academic English (MICASE) and its British counterpart being British Academic Spoken English Corpus (BASE). A rising trend in this type is to collect learner corpora. Although most of the learner corpora projects include written language, there are some corpus projects regarding the spoken English by L2 learners. Notable is Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin, 2012) which put together the interlanguage data from 16 different countries including German, Norway, Turkey etc. (Kilimci, 2014).

Research based on both large-scale and specialized corpora has provided fruitful results in understanding specific properties of spoken language, which, in turn, has the potential in informing language-teaching practices. Following section elaborates on the major findings regarding this type of research.

3 Major Findings of Research on Spoken Corpora

Closer scrutiny of the studies analyzing spoken English through spoken corpora reveals that research findings could be handled under three broad categories. The first one is the findings regarding the lexical frequency of spoken English. Leech, Rayson and Wilson (2001), for example, compared the written and spoken English through the use of BNC with the purpose of determining which language items are significantly frequent in each register. As a result, they provided a thorough list of lexis of spoken English including quite frequent verbs (e.g. mean, know and think etc.), discourse markers (e.g. well, actually etc.) and hedging devices (e.g. sort of, a bit, kind of etc.). Their analysis has shed some light on the distinctive features of spoken interaction. In the same line, Carter and McCarthy (2006) investigated the lexical chunks of up to five words long in spoken English. They made use of BNC as the database of their study. The list presented by the researchers at the end of the analysis shows similarities to that of Leech et al., (2001) as they have found out that spoken English is dominated by the use of such frequent words as know and mean and hedging devices. More recently, Buttery and McCarthy, (2012) analysed the distinctive lexicon of spoken interaction by comparing a list of 2000 high frequency words in written English with a high frequency list of spoken English. The lists were obtained from the relevant subsections of BNC. The results have demonstrated that while written and spoken lexis overlaps at the rate of 65%, 35% of the language items under investigations are found to be unique for spoken interaction. The researchers reported that the most frequent language units are the devices used for interpersonal strategies (i.e. hedging and politeness) considered to be significant for a successful face-to-face interaction. The notion of lexical frequency was also researched through specialized spoken corpora (e.g. interlanguage corpora) as well (Aas, 2011; De Cock, 2004). The results obtained through the analysis of NS speech in comparison with learner spoken performance indicate that while the NS speech is characterized by discourse markers, interactive words, hedges, vagueness and politeness, learners' speech seems to lack most of the lexical items in these categories, which bears significant pedagogical implications regarding teaching speaking skills to EFL learners (De Cock, 2004; Shirato & Stapleton, 2007; Şahin Kızıl & Kilimci, 2014).

The second category through which spoken corpus research findings could be handled is the grammar of spoken English. A number of studies conducted on the large-scale spoken corpora have brought some grammatical features of spoken English to light while acknowledging the common ground between spoken and written grammar. One of the most influential studies in this direction is reported by Biber et al., (2007) who analysed the grammatical properties of spoken English through the Longman Spoken and Written English Corpus (Biber et al., 2007). The researchers have identified idiosyncratic properties of spoken English including prevalence of non-clausal units or various types of ellipsis among many others. Another leading study is reported by Hughes and McCarthy, (1998) who found that the spoken English has its own grammatical properties. They, therefore, suggest that there should be a different approach to teaching the grammar of speaking.

Finally, spoken corpus research has contributed to the fields of discourse and pragmatics. Carter and McCarthy (2006) investigated the vague language in spoken interaction and proposed the patterns of sharing context and knowledge among the interlocutors. Findings of their study are considered to be the evidence of ubiquity of vague language in conversation, which, in turn, could make a base for the necessity of introducing discourse oriented instruction into language teaching. Another pragmatic feature that come to the fore by the analysis of spoken corpora is the turn-construction. Studies in this vein, though limited in number, show that there is a limited set of language units used for realizing turn construction (McCarthy & Carter, 2002).

When taken together, the findings from the studies of spoken corpora are indicative of peculiar nature of spoken language, which implies the need for developing an innovative approach to teaching spoken English through integrating these findings into the teaching practices. Following section presents the pedagogical implications arising from the spoken corpus research.

4 Major Implications Arising From Research

Spoken corpus based research especially those comparing spoken English with the written language clearly shows that spoken language is quite different from written one; therefore, teaching speaking requires a different approach that could be reinforced through corpus-informed pedagogy. A corpus-informed approach to language teaching can be defined as translating the findings obtained through corpus analysis into language teaching practices both in the form of developing materials and designing instructional activities.

Holistic evaluation of the literature on spoken corpus analysis as sketched out above implies that corpus-informed pedagogy should be introduced to the field of language teaching, which has the potential to bring out numerous benefits for the language learners. Although there is a renewal of interest in this direction (e.g. Touchstone series (McCarthy et al., 2005), current practices point out a lack of application especially in terms of corpus-informed language materials (Caines et al., 2016). Therefore, the first implication arising from the research is that spoken corpus findings should be represented sufficiently in the materials targeting to teach oral skills to language learners. In the current situation, most of the spoken language materials are designed in a mono-directional format; however, spoken corpus analysis clearly shows that conversation is of bi-directional nature requiring language uses accordingly. This is also the case for the materials aiming to teach listening skills as well. Listening instruction, in many cases, focuses on comprehension and require learners to complete the tasks after listening to an audio. However, corpus based analysis demonstrates that listening in real life settings involves responding and constructing turns (Caines et al., 2016). Therefore, the findings from corpus research suggest that instructional materials on listening and speaking skills should be adjusted in such a way to include real-life practices and teach learners actual language use.

Another important implication drawn from the spoken corpus research, especially from those comparing learner speech with native speaker speech is that language learners should be made aware of the idiosyncratic properties of spoken English. As shown by most of the research, learners are generally biased towards written English and registerinterference which refers to learners' use of patterns from written language in their speech or vice versa is at work most of the time (Gilquin & Paquot, 2008) (Aijmer, 2002). To offer a solution for this, learners could be provided with consciousness-raising activities in speaking classes. As an activity of this type, Huang (2013) suggests comparing an academic word list and frequent words in spoken English. Instructing students on how to use online corpora for language learning purposes could also produce fruitful results in this direction. When the students are equipped with the necessary skills to search a spoken corpus, they could discover the properties of spoken English without being dependent on the teacher or classroom material. The literature shows that there is an increasing tendency among the language practitioners in consulting online corpora for language learning; however, it has not gained the popularity it deserves.

Final implication suggested by the relevant literature is that a special attention should be paid to development of pragmatic competence in EFL learners. An important part of learning a second language is learning the pragmatic properties of the target language not just to maintain successful communication but to develop related vocabulary. Wei (2009) observes that lack of pragmatic knowledge of the target language potentially results in disfluency in speech and sounding non-native like. To overcome this problem of L2 learners, pedagogical actions could be taken in instructional settings. Utilization of chunks arising from the phraseology research could make a starting point in this direction. Wei (2009) suggests analyzing "appropriate use of functional chunks in teaching English conversation" (p. 292), and performing such an analysis with the students would provide them with insights on pragmatics of English conversation.

5 Concluding Remarks

In conclusion, corpus based analyses of spoken English, though developing relatively slowly, have provided increasingly adequate description of spoken language and have brought about new perspectives to linguistics and language teaching. As noted by Gabrielatos (2005), "corpora have made it possible to compare native intuitions with actual use, and move from prescription to description (p.22)" which has the potential to inform the practices in language teaching.

Despite the increasing interest in spoken English through corpus methods, there is still much room for improvement. First, the fact that existent spoken corpora make up a relatively smaller portion of all corpora implies the need for developing spoken corpora with greater amount of data. Equally important is the inclusion of audio and video sources in accompany with the scripted speech. As emphasized by Caines et al., (2016) availability of multimodal corpora reinforced through video and audio materials is likely to enable the EFL learners to discover the properties of spoken language (e.g. gesture, body language etc.) which are otherwise difficult to access. Additionally, given the limited amount of research on spoken corpus, it is hardly surprising that the findings from the extant studies have had the noteworthy impact on pedagogic materials. It could therefore be concluded that there is a need for more corpus-based teaching materials. Finally, further research is necessary to discover the uncharted features of spoken language.

References

Aas, H. L. (2011). Recurrent word-combinations in spoken learner English : A study of corpus data from Swedish and Norwegian advanced learners. Ph.D. Thesis, University of Oslo.

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In: S. Granger, J. Hung, and S. Petch-Tyson, eds. *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins Publishing Company, pp. 55–76.

Baker, P., Hardie, A. and McEnery, T. (2006.) *A glossary of corpus linguistics.* Edinburgh: Edinburgh University Press Ltd.

Bernardini, S. (2004). Corpora in the classroom: An over-view and some reflections on future developments. In: J. Sinclair, ed. *How to use corpora in language teaching.* Amsterdam: John Benjamins Publishing Company, pp. 15–38.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (2007). *Longman grammar of spoken and written English.* England: Pearson Education Limited.

Boxer, L. and Pickering, J. (1995). Problems in the presentation of speech acts in ELT materials: The case of complaints. *ELT Journal*. Vol. 49, pp. 99–158.

Buttery, P. and McCarthy, M. (2012). Lexis in spoken discourse. In: J. Gee, and M. Handford, eds. *The routledge handbook of discourse analysis.* New York, NY: Routledge, pp. 285–300.

Caines, A., McCarthy, M. and O'Keeffe, A. (2016). Spoken language corpora and pedagogical applications. In: F. Farr, and L. Murray, eds. *The Routledge handbook of language learning and technology.* London: Routledge, pp. 348–362.

Carter, R. (1998). Orders of reality: CANCODE, communication and culture. *ELT Journal*. Vol. 52, pp. 43–56.

Carter, R. and McCarthy, M. (2006). *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Chambers, A. (2005). Integrating corpus consultation in language studies, *Language Learning & Technology,* Vol. 9 No.2, pp. 111–125.

Chambers, A., Farr, F. and O'Riordan, S. S. (2011). Language teachers with corpora in mind: From starting steps to walking tall. *Language Learning* Vol. 39 No. 1, pp. 85–104.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*. Vol. New Series. No. 2, pp. 225–246.

Dutra, D. P. (2004). Bundles in learner corpora: What a type and token analysis can reveal [Online].Available at: http://www.nilc.icmc.usp.br/elc-ebralc2012/anais/completos/104066.pdf [Accessed 30 Augst 2016].

Fitzpatrick, E. (2007). *Corpus linguistics beyond the word corpus research from phrase to discourse.* The Netherlands: Rodopi.

Gabrielatos, C. (2005). *Corpora and language teaching: Just a fling, or wedding bells?*, *Tesl-Ej*, Vol. 8. No. 4, pp. 1–39.

Gilquin, G. (2012). LINDSEI [Online]. Available at: https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html [Accessed 25 May 2016].

Gilquin, G. and Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction.* Vol. 1. No. 1, pp. 41–61.

Huang, L. (2013). Pedagogical implications of the corpus-based investigation of discourse markers. *Learner Corpus Studies in Asia and the World. Vol.* 1, pp. 227–254.

Hughes, R. and McCarthy, M. (1998) 'From sentence to discourse: Discourse grammar and English language teaching', TESOL Quarterly, 32, pp. 263–287.

Kilimci, A. (2014). LINDSEI-TR: A new spoken corpus of advanced learners of English. *International Journal of Social Sciences and Education*. Vol. 4.No.2, pp. 401–410.

Leech, G., Rayson, P. and Wilson, A. (2001). *Word frequencies in written and spoken English based on British National Corpus.* London & New York: Routledge.

McCarthy, M. and Carter, R. (2002). This, that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *The Iris Association for Applied Linguistics. Vol.* 21, pp. 30–52.

McCarthy, M., McCarten, J. and Standtford, H. (2005). *Touchstone Student's Book 1.* Cambridge: Cambridge University Press.

O'Keeffe, A. and Farr, F. (2003). Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly*. Vol. 37. No. 3, pp. 389–418.

O'Keeffe, A., Mccarthy, M. and Keeffe, A. O. (2010). *The Routledge handbook of corpus linguistics.* Abingdon, UK: Routledge.

Ruhi, Ş. (2011). Creating a sustainable large corpus of spoken Turkish for multiple research purposes [Online]. Available at: https://std.metu.edu.tr/wp/wp-content/uploads/2009/05/ruhi_position_paper_gebze_110918.pdf [Accessed 18 June 2016]

Shirato, J. and Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*. Vol. 11. No. 4, pp. 393–412.

Sinclair, J. (1987). *Collins COBUILD English dictionary for advanced learners.* London: Harper Collins.

Şahin Kızıl, A. and Kilimci, A. (2014). Recurrent phrases in Turkish EFL learners' spoken interlanguage: A corpus-driven structural and functional analysis. *Journal of Language and Linguistic Studies*. Vol.10. No. 1, pp. 195–210.

Wei, N. (2009). On the phraseology of Chinese learners spoken English: Evidence of lexical chunks from COLSEC. In: A. Jucker, D. Schreier and M. Hundt eds. *Corpora: pragmatics and discourse. Papers from the 29th international conference on English language research on computerized corpora (ICAME 29).* Ascona, Switzerland: Rodopi, pp. 271–296.

Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (1997). *Teaching and language corpora.* London: Longman.