Beyond expectations: (Applied) corpus linguistics and a framework for the study of spoken professional talk

Eric Friginal
The Hong Kong Polytechnic University
eric.friginal@polyu.edu.hk

Abstract

The analysis of real-world, recorded, and transcribed texts (i.e., corpora) of professional, spoken communication in the workplace has been conducted quite successfully through corpus-based approaches (Friginal, 2024). Corpus linguistics is primarily a methodological research approach to the study of language, and specifically, discourse structure, patterns, and use (Biber et al., 2010; Thompson & Friginal, 2020), with corpora serving as datasets of systematically collected, naturally-occurring registers of texts utilized for a variety of purposes. The use of corpora has become a popular approach in the quantitative analysis of the linguistic characteristics of written and spoken language, in general, and sub-registers such as oral communication in the workplace, in particular (Egbert et al., 2022; Staples, 2015). Various findings have pedagogical and, more importantly, language policy applications. This paper focuses on the important contributions of an iterative corpus-based framework to examine linguistic patterning in telephone/telephony-mediated professional discourses so as to obtain novel understandings of how talk is used and construed in these domains. Current limitations, emerging contributions from generative AI applications, and a call to action proposing training and assessment models will be discussed.

Keywords: Professional workplace corpora, spoken communication, applied corpus linguistics, telephone-based interaction.

Resumen

Más allá de las expectativas: Lingüística de corpus (aplicada) y un marco para el estudio de la interacción oral en el ámbito profesional

https://doi.org/10.17398/2340-2784.47.43

El análisis de textos reales, grabados y transcritos, de comunicación oral profesional en el lugar de trabajo se ha llevado a cabo con cierto éxito desde un enfoque basado en corpus (Friginal, 2024). La lingüística de corpus es una orientación metodológica para la investigación lingüística, en especial en lo que respecta a la estructura discursiva, patrones y uso (Biber et al., 2010), que se basa en corpus que funcionan como conjuntos de textos compilados de forma sistemática pertenecientes a géneros reales que se emplean con diferentes objetivos comunicativos. El uso de corpus se ha popularizado para el análisis cuantitativo de las características lingüísticas de textos escritos y orales, en general, y, en particular, de subregistros como los que se producen en la interacción oral en el lugar de trabajo (Egbert et al., 2022; Staples, 2015). Diversos hallazgos obtenidos desde esta perspectiva presentan aplicaciones no solo pedagógicas, sino también para la política lingüística. El presente artículo se centra en las importantes contribuciones de un marco iterativo basado en corpus para examinar los patrones lingüísticos en el discurso profesional por teléfono para comprender de qué manera se usa y se construye la interacción oral en este espacio profesional. Se discuten, asimismo, las limitaciones actuales de este tipo de investigación, así como las contribuciones emergentes de las aplicaciones de inteligencia artificial generativa, y se hace una llamada a la acción proponiendo modelos de entrenamiento y de evaluación.

Palabras clave: Corpus de interacciones en el ámbito laboral, comunicación oral, lingüística de corpus aplicada, interacción telefónica.

1. Introduction: Corpora and spoken, professional interaction

Over the years, a methodical and systematic description of a range of broad-to-specific linguistic features of spoken discourse has been achieved through corpora and corpus analysis. Corpus-based comparisons from transcribed texts have shown variations in the use of lexical and syntactic choices of participants across registers, especially of professional workplace interactions, which I study. These domains have traditionally been explored by applied linguists by looking at, for example, the structure of talk through the analysis of socio-phonetic features of speech (e.g., Orr, 2003), transactional and interactional dialogues (e.g., Cheepen, 2000; Cheepen & Monaghan, 1990), and how interlocutors navigate and complete specific tasks through turn-taking and turn markers such as interruption, latching, and overlaps (e.g., Gardner & Wagner, 2004; Schegloff, 2001). In addition, many socio-pragmatic patterns of workplace interactions have been examined with a substantial degree of interest by discourse or conversation analysts. Seminal works by Cameron (2001, 2008) have considered top-down talk in service encounters and investigated language that is regulated and standardized.

Although there were initial challenges in collecting and quantifying spoken corpora, innovative approaches in the past 20 years have paved the way for what now have become relatively easier comparisons of quantitative information and linguistic distributions of transcribed speech. Quaglio (2009) and Al-Surmi (2012), for example, identified the linguistic characteristics of speech from a television sitcom and selected soap operas for comparison with face-to-face conversations. These studies revealed important functional differences between television dialogues and naturally-occurring "real-world" conversation. In the early 2000s, when corpus-based approaches were steadily progressing together with advancements in computing and affordances from the Internet, more studies started to focus on professional and workplace domains, matching theoretical developments in the area of English for Specific or Occupational Purposes (ESP/EOP). Adolphs et al. (2004) explored the application of corpus methodologies in healthcare encounters in order to describe the characteristics of communication events in clinical settings. Using a corpus of staged telephone conversations between patients and clinicians, they were able to show several linguistic characteristics of the strategies used by healthcare professionals in addressing caller and patient needs. More recently, Staples (2015, 2019) analyzed the spoken discourse characteristics of patient-provider interactions in healthcare, demonstrating how to apply corpus findings and resources to classroom settings, including pronunciation training for learners.

One very clear strength of corpus-based methods is that the largely quantitative analysis of language allows for linguistic features *in use* to be found and disclosed, that would otherwise remain hidden or undetected by speakers' perceptions. Macro analyses to address groups of people, various demographics, registers, or situational contexts could then be conducted to produce a range of numerical data for (qualitative) interpretation and potential application in practical contexts. Related studies in spoken academic settings have analyzed the distribution of linguistic features of speech, e.g., stance expressions in classroom management (Biber, 2006), features such as *so* and *oh* in social interactions (Bolden, 2006), or markers of accommodation and involvement in class lectures (Barbieri, 2006). Results from these analyses have shown unique distributional data of speech characteristics and linguistic strategies employed by speakers across roles and demographic characteristics—applicable in other settings and all very relevant for language learners.

1.1. Telephone-mediated professional talk

Utilizing corpora, I have been studying professional talk, in particular, telephone-mediated interactions, in several contexts, and especially in cross-cultural workplaces, for over 25 years now. Telecommunication in professional domains is unique, as speakers navigate interactions with interlocutors without immediate visual cues, not sharing the same space (or time) during the speech event. The telephone and radiotelephony have become essential parts of modern business operations (Friginal, 2024). They are designed to provide participants with a quick and easy way to address concerns, queries, or issues with a product or service remotely. This phenomenon has produced a growing interest in research on aviation, maritime, and customer service call center discourses from sociolinguistic and pragmatic perspectives, and, especially, in ESP/EOP. Large-scale, location-based comparisons and turn-by-turn micro-analyses of communication have looked at the complex power relations and face-considerations in dealing with interlocutors' language and behavior (Bieswanger, 2016; Estival et al., 2016; Friginal, 2022). Telecommunications have evolved significantly, incorporating new technologies such as artificial intelligence (AI), web chats or chatbots, and video-based platforms, among others, to provide opportunities to further enhance professional talk (Friginal & Friginal, 2023; Lockwood, 2022).

My studies of (outsourced) customer service interactions and pilot-controller on-the-job communication have been supported by partner multinational corporations operating in the Philippines, India, Costa Rica, and the United States (U.S.), as well as aviation universities, tech companies, and staff training centers. I have collected an outsourced call center corpus from recordings of interactions between international call-takers and callers (i.e., customers) from the U.S. These interactants engage in various types of communicative tasks, e.g., troubleshooting a technical problem or processing orders for a wide range of products, with defined speaker roles similar to a business service encounter (i.e., server vs. servee or agent vs. customer). My primary foci include the dynamics of cross-cultural communication between participants, gender of speakers, call-takers' experience in phone support and quality of service performance, and the linguistic structure of communicative tasks in customer service interactions. Many other studies of globalized call center interactions have been conducted in the past several years matching the growth of the outsourcing industry in the Philippines and India and other parts of the world (e.g., Cowie, 2007; Friginal & Friginal, 2023). Among these, Poster (2007) and Taylor and Bain (2005) look at labor practices in Indian call centers that require Indian call representatives to pose as Americans for American call centers, or British for those that serve companies located in the United Kingdom.

Related to telecommunications in call centers are the intricate turns and exchanges in pilot and controller communication in aviation. Aeronautical radiotelephony encompasses what is known as standardized phraseology and plain English (ICAO, 2010), prescribing norms and use of language, typically English for global flights. Routine aviation operations are covered by standardized phraseology, which is prescribed in the United Nation's International Civil Aviation Organization (ICAO) Document 9432: Manual of Radiotelephony (2007). Standardized phraseology does not adhere to the grammar rules of common English, omitting many extraneous function words and using only a set of about 400 lexical items (Philips, 1991). In addition to its limited lexicon and syntactic structures, standardized phraseology is unique semantically in its rejection of ambiguity, and phonetically in its standardization of pronunciation. Standardized phraseology is the preferred register of use, but as its components are limited, it cannot be used in all situations. Using corpora, Bieswanger (2016) found that the register of plain English also maintains structural conciseness and a restricted lexicon (in general, similar to standardized phraseology), but he argued that these two are distinct registers which both need to be explicitly taught in schools and training facilities.

My argument here is that the use of corpora and associated corpus-based approaches has successfully advanced the linguistic analysis of professional communication in face-to-face settings as well as those mediated by the telephone and radiotelephony. By harnessing the power of large-scale data sets, I have been able to delve deeper into the intricacies of language use, uncovering patterns, nuances, and pragmatic aspects that were previously elusive. The availability of comprehensive corpora has not only facilitated more accurate and reliable linguistic investigations but has also paved the way for the development of innovative computational tools and techniques that enhance our understanding of professional communication in diverse contexts. As we continue to explore and refine these approaches, it is clear that corpora and corpus-based methodologies will remain invaluable assets, shaping the future of linguistic analysis and contributing to our knowledge of professional communication dynamics.

2. Corpus linguistics *applied*

Corpus linguistics is a methodological research approach to the study of language, and specifically, discourse structures, patterns, and use (Biber et al., 2010). Corpora serve as datasets of "systematically collected, naturally-occurring registers of texts" (Friginal & Hardy, 2014, p. 20), which are electronically stored, analyzed, and utilized for a variety of purposes. Bowker and Pearson (2002) identify four primary characteristics of a corpus as: (1) authentic, (2) relatively large, (3) electronic, and (4) conforms to specific design criteria. There are corpora containing a variety of registers (also referred to as "text types") including academic and professional English, spoken English in job interviews, newspaper articles, learner language, or chatbot interactions with direct policy or teaching applications. There is no specific rule regarding the size of a corpus but it should be large enough to promote a systematic analysis of relevant, target linguistic patterns, especially when utilized for materials design in the classroom (Friginal & Hardy, 2014). With the advent of audio recording and transcription software and state-of-the art programming tools, more specialized spoken corpora have been compiled and also freely shared and explored for research and teaching purposes. One clear benefit of this is that corpora facilitate targeted observation and study of authentic language use and more opportunities for triangulation and continuing research.

2.1. Teaching and policy implications

Specifically, *applied* corpus linguistics, utilized in language and social research, has contributed important linguistics-based explications of discourse with critical language policy and pedagogical implications (Thompson & Friginal, 2020). Biber et al. (2010) noted that corpora have been held to be default resources in linguistic research, and various stakeholders of a particular domain, therefore, benefit from the practical and pragmatic applications of corpus data. For example, corpora have contributed immensely to studies of phraseological and collocational patterns of English, illustrating how such patterns can inform language training for specific purposes. In a field like aviation, phraseology is a very important area of study, and corpus approaches have enhanced the ability of pilots and their controllers to understand and utilize prescribed forms of utterances successfully. As Römer (2009) observes, "language is highly patterned" (p. 140), and often, these patterns are important to highlight and teach in the training classroom, ensuring safety aviation operations (Friginal et al., 2020).

The application of corpus linguistics to language policy and pedagogy has become an increasingly important area of study in recent years. One of the key ways in which corpus analysis can inform language policy is by providing empirical evidence to support or challenge existing assumptions about language. For example, corpus data can reveal patterns of language change over time, such as the emergence of new vocabulary, changes in grammatical structures, or shifts in the relative frequencies of different linguistic features (Grieve, 2016). This information can be crucial for policymakers and language planners as they seek to update and revise language policies to reflect the evolving nature of language. Similarly, corpus analysis can shed light on the linguistic diversity within a given speech community, illustrating differences in language use across various social, regional, or demographic groups. This information can be particularly important for language policies that aim to promote linguistic equality, protect minority languages, or address issues of language variation and standardization (Friginal & Hardy, 2014; Grieve, 2016).

In both aviation and outsourced call centers, corpus data shed light on the challenges that non-English first language (L1) interlocutors typically experience as they navigate discourses that have rules and expectations more aligned with English L1 speakers. By identifying the areas of language that are particularly difficult or problematic for second language (L2) speakers of English, resulting patterns can then be used to develop more specific and effective instructional strategies, such as the incorporation of explicit pattern-based instruction or the use of authentic language samples for discussion and practice in the training classroom. Moreover, corpus-based research can also inform the development of language assessment tools, by identifying the linguistic features and patterns that are most relevant and important for successful communication in a given context. This information can then be used to design assessment tasks and rubrics that more accurately reflect the language skills and abilities that are valued in real-world situations (Garcia & Fox, 2020). Pilot and controller language assessment has slowly incorporated corpus data, although the ICAO language requirements are still strictly modelled on L1 norms (Estival et al., 2016; Friginal et al., 2019).

It is important to note that the application of corpus-based data and findings to language policy and pedagogy is not without its challenges. For instance, the interpretation and application of corpus data can be complex and context-dependent, requiring a deep understanding of the linguistic and sociocultural factors that shape language use. Moreover, the availability and accessibility of high-quality corpus data can vary significantly across different languages and contexts, which can limit the generalizability of corpus-based findings (Egbert et al., 2022). Despite these challenges, the potential benefits of applying corpus-based research to language policy and pedagogy are significant and far-reaching. By providing empirical evidence about language use and language learning, corpus linguistics can help to enlighten the development of more informed, effective, and equitable language policies and teaching practices that better serve the needs of diverse language communities and learners. As the field of applied corpus linguistics continues to evolve and expand, it is likely that we will see an increasingly close and productive collaboration between corpus researchers, language policymakers, and language educators. This collaboration has the potential to drive significant improvements in the way we understand, plan, and teach (spoken) language in professional contexts.

2.2. Analyzing professional spoken corpora: A framework

I have developed an iterative, applied corpus linguistics framework which allows me to build upon my initial findings, ask new research questions, and employ a combination of quantitative and qualitative methods to gain a comprehensive understanding of the spoken discourse phenomenon I am investigating. By following this cycle, I believe that I can systematically explore the linguistic patterns, variations, and contextual factors that shape professional talk, ultimately contributing to the field of applied corpus linguistics and spoken discourse analysis. As shown in Figure 1, the cyclical process of hypothesis formation, quantitative exploration, contextual analysis, and refinement of research questions is well-suited to uncover nuanced patterns in spoken discourse. Quantitative analyses of corpus frequencies and distributions and resulting statistical tests can reveal large-scale trends, but qualitative techniques are definitely needed to understand pragmatic meanings and how linguistic features intersect with social context (i.e., to answer the question: *So what?*). This framework acknowledges that both angles are necessary.

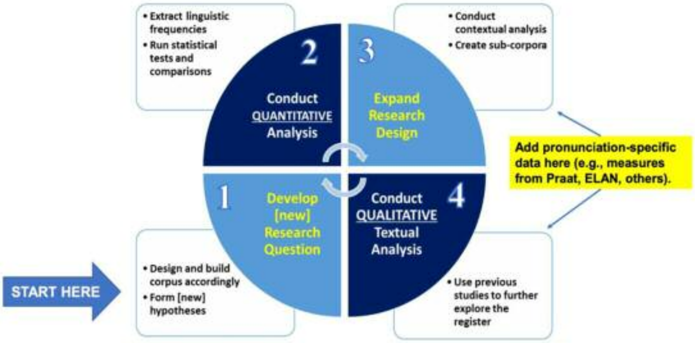
• Extract linguistic frequencies
• Run statistical tests and comparisons
2
Conduct QUANTITATIVE Analysis
3
Expand Research Design
• Conduct contextual analysis
• Create sub-corpora
Add pronunciation-specific data here (e.g., measures from Praat, ELAN, others).
1
Develop [new] Research Question
Conduct QUALITATIVE Textual Analysis
4
START HERE
• Design and build corpus accordingly
• Form [new] hypotheses
• Use previous studies to further explore the register

Figure 1. Stage cycle for corpus-assisted discourse analysis of spoken (professional) corpora.

Overall, starting with rigorously developed research questions, mixed (quantitative/qualitative) analyses, and examination of both linguistic distributions and contextualized language in an iterative process offers a model for how large corpus data can be synthesized with fine-grained linguistic and socio-pragmatic details. This approach reflects a more holistic understanding of obvious and underlying language patterns. Its systematic, evidence-based methods are well-suited for comparing studies across domains and populations. A summary of the steps and components recommended by the framework is provided in Table 1.

Steps		Components/Descriptions
1a	Developing Research Questions	• This is the starting point of the cycle where I identify the research gaps or areas of interest that I want to explore. • Formulating relevant and meaningful research questions is crucial as it will guide the subsequent steps of my analysis and especially the design and collection of my corpus (next step).
1b	Corpus Design and Collection	• Based on the research questions, I will then need to design and collect an appropriate corpus of spoken discourse data. • Careful consideration should be given to factors such as the target population, context, and data collection methods to ensure the **corpus is representative** and relevant to my research objectives.
1c	Hypothesis Formation	• After designing and collecting the corpus, I then can start forming hypotheses about the linguistic patterns, variations, or relationships I expect to observe in the data. • These hypotheses will serve as the foundation for my quantitative analysis (i.e., what statistical test/s will I run?).
2	Quantitative Analysis	• In this step, I will start by processing (e.g., tagging, annotating), cleaning, and analyzing my corpus, obtaining various frequency data and linguistic distributions. • I then conduct statistical tests and comparisons to explore the linguistic frequencies and patterns within the corpus. • Techniques such as frequency analysis, collocations, and statistical significance testing can be employed to identify any notable patterns or deviations from my hypotheses.
3	Expanding Research Design and Contextual Analysis	• Based on the findings from the quantitative analysis, I can then expand my research design by asking new research questions or refining my existing ones. • Incorporating contextual analysis, such as examining the social, cultural, or situational factors that may influence the linguistic patterns, can provide a more nuanced understanding of the spoken discourse being studied.
4a	Qualitative or Textual Analysis	• To complement the quantitative analysis, I can incorporate qualitative or textual analysis approaches, such as using pronunciation-specific data from tools like ELAN or Praat. • This step allows me to delve deeper into the linguistic features, patterns, and their potential underlying meanings or pragmatic functions, mixing transcribed texts with annotations specific to speech characteristics.
4b	Interpretation/Conclusion and Repeating the Cycle	• After the qualitative or textual analysis, I can synthesize my observations and develop concluding remarks. • I then could repeat the cycle by revisiting my research questions, refining my corpus design, and conducting further quantitative and contextual analyses. • This iterative approach enables me to continuously expand my understanding of my corpus, uncover new phenomena, and refine my research methodology.

Table 1. Steps and components of the stage cycle for corpus-assisted discourse analysis of spoken (professional) corpora.

3. Sample analysis and comparison: Call centers and aviation

To illustrate the application of my framework, I present the following sections exploring the combined linguistic characteristics of talk in customer service call centers and aviation, based on specialized corpora of interactions between customer service representatives (or "agents") and their callers, and groups of pilots communicating with their respective air traffic controllers. I have been collecting both specialized corpora with my students and collaborators – Cross-Cultural Aeronautical Communication Corpus (CCACC) and Corpus of Outsourced Customer Service Calls (Co-CSC) from an overarching research goal of describing a range of discourse characteristics of these two domains. CCACC and Co-SCS are exploratory corpora, annotated across socio-cultural structures and task dimensions of interaction in these two settings, focusing especially upon speakers' first language background (L1), role-relationships, discoursal goals and objectives, and cultural identities. For this paper, I decided to compare these two similar, but certainly distinct, telephone-mediated registers to show how my framework is able to produce teaching, assessment, and policy-based implications to these industries. A full-paper version of this comparison is found in Friginal (2024), but the results presented here included more recent corpus data and additional participants.

For Step 1a of the framework (Developing Research Questions), I pursued a multi-dimensional (MD) analytical approach developed by Biber (1988), together with pre-identified linguistic features from Friginal (2009, 2013) in order to analyze the characteristics of outsourced call center and aviation discourse in general, and in particular, potentially culture-specific differences between speakers' utterances. Implications related to macro language policy and culture-based training for agents and pilots, future correlational studies, and the sustainability of these two industries are briefly discussed below.

3.1. Linguistic co-occurrence and MD analysis

The concept of linguistic co-occurrence, which is the foundation of MD analysis (Step 2 of the framework), can be introduced by pointing out intuitively the common differences in the linguistic composition of various types of registers. For example, spoken registers are different from written registers because of factors such as dysfluencies and the presence of numerous linguistic features that show immediate interactivity (e.g., questions and responses, speech-act formulae, or inserts). With computational tools such as Biber's grammatical tagging program, it is then possible to statistically identify and establish these sets of co-occurring linguistic features and compare how they are used by different groups of speakers. In a call center corpus, for example, a comparison of how groups of U.S.-based, Indian, and Filipino agents make use of these statistically correlating features is possible, and then attempt to describe their unique functions derived from these agents' distinctive demographic characteristics. The same process applies to (non-English L1) international pilots navigating required radiotelephony in completing their tasks with U.S.-based controllers on the ground. The emerging sets of features tell something about the detailed intercultural, linguistic composition of the discourse which is not normally seen in qualitative observations. An extensive discussion of the statistical procedure and interpretation of corpus-based, MD analysis can be found in Biber (1988, 2006), Conrad and Biber (2001), Friginal (2009; 2013), and Berber-Sardinha and Veirano Pinto (2014, 2019). Table 1 shows the composition of specialized corpora used in this study.

Corpora		Number of texts (i.e., speakers)	Number of words
CCACC	International pilots (non-English L1 speakers) (INT-P)	220	42,000
	U.S. pilots (U.S.-P)	100	18,500
	U.S. pilot trainees (U.S.-PT	80	12,000
	Total	**400**	**72,500**
Co-CSC	Philippine agents (PHIL)	400	120,000
	Indian agents (IND)	300	86,000
	U.S.-based agents (U.S.)	300	82,000
	Total	**1,000**	**288,000**

Table 2. Composition of the Cross-Cultural Aeronautical Communication Corpus (CCACC) and Corpus of Outsourced Customer Service Calls (Co-CSC) used for this study (adapted from Friginal, 2024).

The parallel corpora of U.S.-based, Indian, and Filipino call center agents (N=1,000 total texts or total individuals) from two main types of tasks (troubleshooting and product inquiry/order) from Co-CSC was provided by four U.S.-owned call center companies primarily for research and training purposes. Texts from the CCACC were extracted from several sources including those provided by airlines operating in Asian and South American countries with service to U.S. locations. Training and simulation texts from an aeronautical training company were also included in the exploratory corpus, together with texts from VASAviation's YouTube channel (search for "vasaviation" from https://www.youtube.com), with publicly-available audio files (most with accompanying transcripts) of authentic materials that feature a sampling of actual language used by pilots and controllers in in emergency situations. All recordings were transcribed into machine readable text files by trained transcriptionists following conventions used in the collection of the service encounter corpus of T2K-SWAL (TOEFL 2000 Spoken and Written Academic Language, see Biber, 2006 for a description of this corpus). Personal information about the interlocutors, if any (e.g., names, addresses, phone numbers, credit card or social security numbers, etc.) was consistently replaced by different proper nouns or a series of numbers in the transcripts. No attempt was made to annotate phonetically and the transcribed texts were manually checked for format and accuracy (Friginal, 2024).

3.2. Featured result: Task-oriented, polite utterance vs. Involved talk

For analysis and comparison across Steps 3 and 4a of the framework, the linguistic features of the primary dimension (DIM 1: Task-oriented, polite utterance vs. Involved talk) statistically represented 46% of variance in agent and pilot utterances interpreted in this sample comparison. The linguistic composition of DIM 1 is shown in Table 3.

Dimension	Features
Dim 1: Task-oriented, polite utterance vs. Involved talk	Positive: *OK*, second person pronouns (*you*, *your*), average world length, *please*, nouns, possibility modals, *do/don't*, *could* (*could you*), nominalizations, *let's* (*let us*), questions, average length of turns, *thanks*, *ma'am/sir*, time adverbs, *now*, pronoun *they* ⇳ Negative: Pronoun *it*, first person pronouns (*I*, *my*), past tense verbs, *that* deletion, private verbs, WH clauses, perfect aspect verbs, *I mean/you know*, verb *do*

Table 3. Linguistic dimensions of outsourced call center and aviation English interactions.

The combination of positive and negative features in DIM 1 illustrates a linguistic dimension that differentiates between transactional (e.g., *OK*), addressee-focused (e.g., use of second person pronouns *you/your*), polite (*thanks, sir/ma'am*), and elaborated discourse (e.g., longer average length of turns, nouns, and nominalizations) and involved narrative (first person pronouns, past tense verbs) portraying how informational content is produced by agents and pilots in customer service and aviation communication. Figure 2 shows the average dimension scores of Philippine (PHIL), Indian (IND), and U.S.-based (U.S.) agents as well as International (INT-P) and U.S. pilots (U.S.-P) and U.S. pilot trainees (U.S.-PT) along a positive and negative scale.

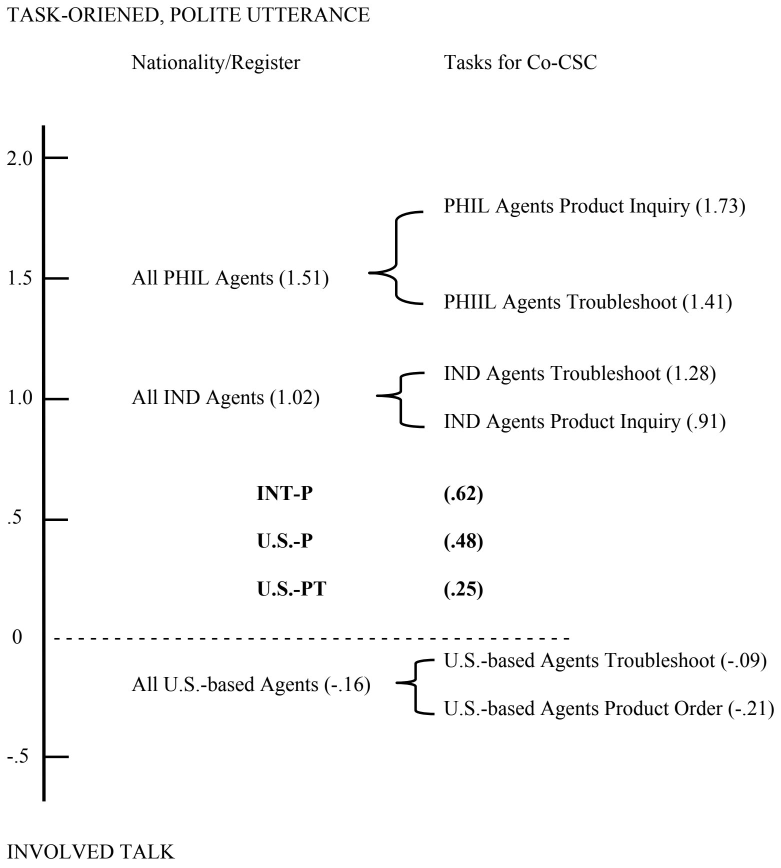

Figure 2. Comparison of average dimension scores for DIM 1: Task-oriented, polite utterance vs. Involved talk (Friginal, 2024).

Comparative group dimension scores revealed differences in the way agents (from two task groups: troubleshooting and product inquiry) and English L1 and L2 pilots and student pilots make use of these co-occurring linguistic features. For Co-CSC, PHIL and IND agents plot on the positive side of the scale, while U.S.-based agents are on the opposite. All aviation speaker groups are on the positive side of DIM 1, with International Pilots with slightly higher average dimension scores than the two U.S.-based groups. The consistent use of addressee-focused, involved production features, and politeness and respect markers establishes the linguistic preferences of PHIL and IND agents compared to U.S.-based agents. In general, service encounters commonly allocate for courteous language and the recognition of roles (e.g., server vs.

servee), and call center agents are expected to show respect and courtesy when assisting customers (D'Ausilio 1998; Friginal 2009). However, of interest here is the variation between these groups of agents who are, in fact, dealing with similar contexts or tasks. PHIL agents overwhelmingly prefer a collective use of these features in responding to their callers. PHIL agents also differ dramatically from their American (presumably English L1 speaker) counterparts, which poses a series of relevant research questions for future detailed investigation (e.g., Do these differences correlate with quality of service? Should PHIL and IND agents be strictly trained to mirror distributions and features of U.S.-based agents in this dimension?).

3.3. Comparison of politeness markers across groups of pilots

Figure 2 shows that INT-P slightly differed from their U.S.-P counterparts and an in-depth analysis of data showed that the major difference is accounted for by distributions of markers identified as lexical politeness features (e.g., *thanks/thank you, appreciate, sorry, apologize, please, could you** [*please*]). The use of politeness features in aviation is a more complicated issue than in call centers. As prescribed by various ICAO language-related policies, training procedures, and manuals used during student pilot exercises require that routine communication between pilots and controllers be conducted solely and strictly in prescribed standard phraseology. As politeness is not part of standard phraseology, aviation interlocutors have been trained to understand and consider that these features may be identified as superfluous and unnecessary in the interaction (Ishihara & Prado, 2021). Politeness markers are also deemed as impeding communication efficiency and detracting from conciseness in communication. In addition, training manuals have also explicitly highlighted that pilot and controller utterances introducing a new topic are more likely to fail if they are mitigated (i.e., they are indirect) than if they are direct (Linde, 1988). The use of these polite markers is identified, therefore, as also representations of mitigation in discourse, and that these turns are more likely to fail if they are mitigated than if they are direct (Linde, 1988). Figure 3 shows that International Pilots have close to four (3.89) of these polite markers normalized per 1,000 words, compared to .5 and .88 for U.S. pilot trainees and U.S. pilots respectively.

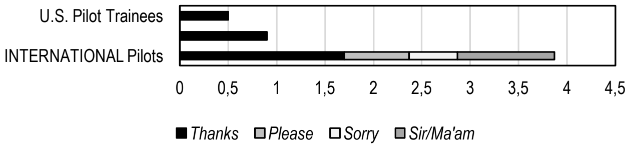

Figure 3. Comparison of politeness markers across groups of pilots (Friginal, 2024).

There were no occurrences of *please* (including *could you*), *sorry* (*apologize*), and especially *sir*/*ma'am* in the turns by U.S. pilots and trainees. Interestingly, International Pilots used *sir*/*ma'am* once per 1,000 words in their collective turns. At times, *sir* co-occurs with *Roger*, which indicates that a message had been heard and understood (*Roger, sir*). Several occurrences of *thank you* and *appreciate* are found in the CCACC. Text samples of occurrences of polite markers in international pilots' turns are shown below:

(1) [airline] OK, hold short of Mike Alpha, roger **sir** (source: CADS-1244)

(2) Roger to the gate, **thank you**. (source: CADS-1146)

(3) Oh, negative **sir** we're on two two right holding short of foxtrot. (source: CADS-3556)

(4) Roger, **sir**, we just exit the runway and we're holding short of […] (source: CADS-3556)

(5) I'm not on the ramp yet, **sir**. (source: CADS-3556)

(6) Yes, **sir**, we'll follow the Asiana, and next time **I would like you** to be polite with me. **Thank you**. (source: CADS-3556)

(7) Holding short of Hotel, **sir. Appreciate** it. (source: CADS-6781)

3.4. Policy and training implications

For Step 4b of the framework, my exploratory, cross-register analysis of intercultural interaction in outsourced call centers and international aviation using MD analysis revealed several interesting characteristics of the discourses in general, and in particular, potentially culture- and task-specific differences between English L1 and non-English L1 interlocutors (agents and pilots). In customer service, agents make use of politeness markers frequently as they engage the callers and monitor the flow of conversation, but there are clear differences potentially contributed by speakers' L1 and cultural background, and DIM 1 differences between PHIL, IND, and U.S.

agents are certainly important topics to further examine. In aviation INT-P's use of DIM 1 features is generally similar to U.S.-P and U.S.-PT, with interlocutors plotting around a comparable range, with only minor dimension score differences. Pilots' turns and questions are within the expected turn-taking sequences, with comparable distributions for nouns, nominalizations, *OK*, and questions (average frequency). Major differences are observed in INT-P's use of lexical politeness markers, especially *sir*, *thank/s/appreciate*, and *please*.

3.4.1. Implications for agent training

It is crucial to establish the nature of intercultural linguistic variation in outsourced call center transactions managed by offshore and inshore/U.S. agents. It is evident that there are systematic patterns and discourse features favored by offshore agents, influenced by factors like their L1 background, customer service norms in their respective countries, as well as training practices in their local call centers. These patterns from corpora can be used for correlational studies, examining variables such as service assessment scores and customer satisfaction survey results. For instance, do higher scores in DIM 1 by PHIL agents correlate positively or negatively with customer satisfaction scores?

Considering the language proficiency of PHIL and IND agents in (American) English, service accuracy, rapport with U.S. callers, and workflow compliance is also important in determining the characteristics of successful or unsuccessful transactions handled by offshore agents compared to their U.S.-based counterparts. The study raises other questions based on its results, including:

- Do Filipinos have an advantage over Indian agents in effectively relating to American callers due to the historical and cultural affinity Filipinos have with Americans and American English?
- Indian agents, based on their average DIM 1 scores and use of polite features, align more closely with U.S.-based callers than Filipino agents. What does this outcome signify for business-specific plans and decisions?

To gain a better understanding of intercultural communication in outsourced call centers, providing linguistic information to these questions and correlating service assessment scores with agents' characteristic discourse patterns is highly relevant and useful. Additionally, qualitative survey results on callers' awareness of accents and how they impact their customer service experience would offer insights into the role of cultural factors and linguistic perceptions in determining the success or failure of outsourced call center communications.

3.4.2. Implications for pilot training

As noted in Friginal (2024), in the domain of call center communication, business talk frequently employs politeness markers, engages callers by providing sufficient or detailed information and explanation, and utilizes discourse markers to monitor the flow of conversation. These patterns, however, are not necessarily encouraged or even necessary in aviation phraseology, as most U.S. pilots, trainees, and controllers do not typically employ such techniques. Nonetheless, given the intercultural and global nature of aviation discourse, alternative approaches to delivering instructional and task-focused language may warrant closer examination and discussion in pilot/controller training and materials design initiatives.

The ICAO's Language Proficiency Requirements (LPRs) have faced notable criticisms (Douglas, 2004) due to issues such as the broad definition of aviation English, which contrasts with the intended scope of the LPRs: radio communications in situations where standard phraseology is unavailable (ICAO 2010). The choice of the term "plain English" to identify the scope of language beyond standard phraseology, as well as the practice of assessing pilots' English skills, have led to varying conceptions of the utterances to be taught and evaluated. The emphasis on grammar and pronunciation has marginalized not only real-world communications, but also other linguistic areas or topics covered by the LPRs (Alderson, 2011; Kim 2018; Kim & Elder, 2009). To support this observation, text samples from the CCACC, particularly those intended for non-English L1 pilots, have shown a preference for personalized and polite support, considering the characteristics of these pilots to ensure accurate and collegial communication with their controllers. Prado and Tosqui-Lucks' (2019) study, also using corpora, reinforces the call to strictly distinguish standard phraseology from plain English in how they are taught and assessed (Bieswanger, 2016). Furthermore, a conversation analysis (CA) of scripts from the Hudson River accident (Garcia & Fox, 2020) suggests that the transition between phraseology and plain language is manifested through the use of "indexical references" (e.g., *you, we, this, here*) and words like *sir* and *okay* (as captured in DIM 1 of this study), which may help signal to all radio frequency users that they should listen attentively and build a collaborative relationship. These DIM 1 features can potentially aid pilots and controllers in fostering a positive relationship and achieving mutual understanding, which is crucial for the management of successful flights and emergencies.

4. Current limitations and a way forward

Corpus-based analysis of spoken professional discourse is a valuable approach for gaining insights into the language use and communication patterns within various professional contexts. However, there are still several significant limitations that researchers like myself often face when undertaking such studies. I summarize my observations below:

- *Difficulty in data collection and corpora building*: One of the primary challenges in this field is the difficulty in collecting suitable data or building comprehensive corpora. Spoken professional discourse often occurs in private, sensitive, or confidential settings, making it challenging to gain access and obtain consent from participants. Professionals, especially in high-stakes or regulated industries, may be reluctant to allow recordings of their conversations, fearing potential repercussions or a breach of client confidentiality.
- *Strict approvals and permissions*: Collecting data for corpus-assisted analysis of spoken professional discourse often requires rigorous approval and permission processes. Researchers must navigate complex institutional review boards, organizational policies, and legal frameworks to ensure compliance with **ethical standards** and privacy regulations. The time-consuming nature of these approval processes can significantly delay or hinder research projects.
- *Unclear ethical and privacy guidelines*: The ethical and privacy considerations surrounding the collection and analysis of spoken professional discourse are not always well-articulated or standardized. Researchers must carefully navigate issues such as participant consent, data anonymization, and the protection of sensitive information. The lack of clear and widely accepted

> guidelines can create uncertainty and hesitation among researchers, potentially limiting the scope and depth of their investigations.

To address these limitations, I believe that researchers and policymakers could collaborate to develop more streamlined and standardized data collection protocols, establish clear ethical and privacy guidelines, and explore innovative methods for capturing authentic spoken professional discourse while respecting the concerns and needs of participants (especially customers in service encounters). Increased funding, interdisciplinary collaborations, and the development of user-friendly data collection tools could also help to overcome these challenges and enable more comprehensive and reliable corpus-assisted analyses of spoken professional discourse. It is also clear that corpus-based methods are still limited when it comes to studying the socio-phonetic features of speech. Pronunciation, including such features as intonation, rhythm, pitch, volume, and stress of words and discourse is complex and difficult to easily program or capture through algorithms. However, there are advancements in the use of computational tools, dictation and transcription software, qualitative coding programs, and automated sentiment analyzers (e.g., those utilized in customer service and social media platforms) that may serve as models for a robust collection of a new generation of specialized spoken corpora especially developed for pronunciation teaching and learning. The annotation of spoken corpora for prosody, for example, the Hong Kong Corpus of Spoken English (HKCSE) (Cheng, et al., 2008) and more detailed contextual transcriptions and annotations of spoken texts suggest prospects for capturing some phonetic features of speech in orthographic transcripts.

4.1. Generative AI and analyzing speech patterns

A way forward is to leverage the current set of tools that have been developed with Generative AI Large Language Models (LLMs). This rather overwhelming flow of computing applications and accessible programming platforms may eventually lead us to multiple options in processing recorded speech. The key potential benefit is automation of routine analysis tasks, allowing linguists to focus on higher-level interpretation of generative models' outputs and discovery of deeper linguistic patterns in speech. Of course, oversight would still be required to ensure LLMs contribute insights appropriately. Automatically transcribing and annotating large volumes of audio/video data would accelerate the analysis by handling time-consuming transcription work. Eventually, new tools may be able to easily generate synthetic speech data to augment existing corpora of various specialized registers. This could help address (pronunciation) gaps in the data, as well as other important demographic characteristics of spoken discourse. More metadata information could also be obtained further by inferring new attributes not present in original recordings, like inferred age/gender of speakers, emotion/sentiment, topics of discussion etc. In general, I am cautiously optimistic about the integration of Generative AI tools in corpus-based research although there are several unknowns as to accuracy and ethical concerns, but I see how this new innovation will allow us to successfully facilitate multi-modal analysis by generating text transcripts to accompany audio/visual data, or synthesizing speech to accompany text corpora.

5. Concluding remarks

The widespread application of corpora and corpus-based methodologies has significantly advanced our understanding of spoken professional interaction. As I have shown in this paper, the use of large-scale transcribed datasets has enabled me to systematically examine the linguistic features, patterns, and pragmatic functions that characterize a diverse range of professional communication contexts, including those from telephone-mediated talk. From the fine-grained analysis of aviation radiotelephony and call center interactions to the broader exploration of business and academic spoken registers, corpus linguistics has provided invaluable insights. Here, I emphasized that one of the key strengths of the corpus-based approach is its ability to uncover linguistic phenomena that may be overlooked or underappreciated by individual speakers' perceptions. By harnessing the power of quantitative data, I have been able to identify subtle variations in lexical choices, syntactic structures, and pragmatic strategies employed by professionals across different settings. This has led me to a more nuanced understanding of how language is used to navigate the complexities of workplace communication, from the highly standardized phraseology of aviation to the more fluid, context-dependent interactions of call center agents and customers.

Furthermore, the corpus-based framework outlined in this paper demonstrates the potential for a systematic, iterative approach to studying spoken discourse. By combining quantitative and qualitative methods, I have been able to move beyond mere descriptions of linguistic features to explore the deeper sociocultural and pragmatic underpinnings of professional communication. The cyclical process of hypothesis formation, data analysis, and contextual examination allows for the progressive refinement of research questions and the uncovering of new, unexpected patterns. This flexibility is crucial, as the dynamics of spoken interaction are inherently complex, requiring a multifaceted approach to fully capture their essence. Finally, the implications of corpus-based research on professional interaction extend well beyond the academic realm. By providing empirical evidence of language use in these domains, corpus linguistics can significantly inform language policy and pedagogy. As we move forward, integrating AI and new models, the continued development and refinement of corpus-based approaches hold the promise of even greater advancements in our understanding of the nature of professional communication. The future of this field is indeed bright, and the potential for further discoveries and innovations is truly exciting.

Article history:
Received 13 May 2024
Received in revised form 17 May 2024
Accepted 20 May 2024

References

Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, *1*, 9-28. https://doi.org/10.1558/japl.v1.i1.9

Alderson, J. (2011). The politics of aviation English testing. *Language Assessment Quarterly*, *8*(4), 386-403. https://doi.org/10.1080/15434303.2011.622017

Al-Surmi, M. (2012). Authenticity and TV shows: A multidimensional analysis perspective. *TESOL Quarterly*, *46*(3), 472-495. https://doi.org/10.1002/tesq.33

Barbieri, F. (2006). *Patterns of age-based linguistic variation in American English conversation*. Unpublished Doctoral Dissertation, Northern Arizona University.

Berber-Sardinha, T., & Veirano Pinto, M. (Eds.) (2014). *Multidimensional analysis 25 years on: A tribute to Douglas Biber*. John Benjamins.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.) (2019). *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.

Biber, D., Reppen, R., & Friginal, E. (2010). Research in corpus linguistics. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd Ed.) (pp. 548-570). Oxford University Press.

Bieswanger, M. (2016). Aviation English: Two distinct specialised registers? In C. Schubert & C. Sanchez-Stockhammer (Eds.), *Variational text linguistics: Revisiting register in English* (pp. 67-86). De Gruyter. https://doi.org/10.1515/9783110443554-005

Bolden, G. (2006). Little words that matter: Discourse markers “So” and “Oh” and the doing of other-attentiveness in social interaction. *Journal of Communication 56,* 661-688. https://doi.org/

10.1111/j.1460-2466.2006.00314.x

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora.* Routledge.

Cameron, D. (2001). *Working with spoken discourse*. Sage.

Cameron, D. (2008). Talk from the top down. *Language and Communication*, *28,* 143-155. https://doi.org/10.1016/j.langcom.2007.09.001

Cheepen, C. (2000). Small talk in service dialogues: The conversational aspects of transactional telephone talk. In J. Coupland (Ed.), *Small talk: Professional and commercial applications* (pp. 288-311). Pearson.

Cheepen, C., & Monaghan, J. (1990). *Spoken English: A practical guide*. Pinter Publishers.

Cheng, W., Greaves, C., & Warren, M. (2008). *A corpus-driven study of discourse intonation*. John Benjamins.

Conrad, S., & Biber, D. (Eds.) (2001). *Variation in English: Multi-dimensional studies*. Longman.

Cowie, C. (2007). The accents of outsourcing: The meanings of "neutral" in the Indian call center industry. *World Englishes*, *26*(3), 316-330. https://doi.org/10.1111/j.1467-971X.2007.00511.x

D'Ausilio, R. (1998). *Wake up your call center: How to be a better call center agent.* Purdue University Press.

Douglas, D. (2004). Assessing the language of international civil aviation: Issues of validity and impact. In *IPCC 2004: Communication frontiers: Proceedings* (pp. 248-252). IEEE Professional Communication Society.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Routledge.

Estival, D., Farris, C., & Molesworth, B. (2016). *Aviation English, a lingua franca for pilots and air traffic controllers.* Routledge.

Friginal, E. (2009). *The language of outsourced call centers: A corpus-based study of cross-cultural interaction.* John Benjamins.

Friginal, E. (2013). Linguistic characteristics of intercultural call center interactions. In G. Nelson & D. Belcher (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 127-153). University of Michigan Press.

Friginal, E. (2022). *I'm sorry my what?* Understanding caller clarification sequences in outsourced call center interactions. *Sociolinguistic Studies*, *16*(1), 65-85. https://doi.org/10.1558/sols.42324

Friginal, E. (2024). The case of task-oriented, polite discourse in intercultural aviation and customer service interactions. *Journal of Corpora and Discourse Studies*, 7(1), 258-281.DOI: 10.18573/jcads.119

Friginal, E., & Friginal, R. (2023). Outsourced call centers. In A. Borlongan (Ed.), *Philippine English: Development, structure, and sociology of English in the Philippines* (pp. 340-352). Routledge.

Friginal, E., & Hardy, J.A. (2014). *Corpus-based sociolinguistics: A guide for students*. Routledge.

Friginal, E., Mathews, E., & Roberts, J. (2019). *English in global aviation: Context, research, and pedagogy*. Bloomsbury.

Friginal, E., Roberts, J., Udell, R., & Schneider, A. (2020). Pilot-ATC aviation discourse. In E. Friginal & J. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 70-85). Routledge.

Garcia, A., & Fox, J. (2020). Contexts and constructs: Implications for the testing of listening in pilots' communication with air traffic controllers. *The ESPecialist*, *41*(4). https://doi.org/10.23925/2318-7115.2020v41i4a4

Gardner, R., & Wagner, J. (Eds.) (2004). *Second language conversations*. Continuum.

Grieve, J. (2016). *Regional variation in written American English*. Cambridge University Press.

ICAO Procedures for Air Navigation Services—Air Traffic Management (Doc 4444) (2007). Montreal, Canada: ICAO.

ICAO Manual of Implementation of the Language Proficiency Requirements (Document9835-AN/453) (2nd Ed.) (2010). Montreal, Canada: ICAO.

Ishihara, N., & Prado, M. (2021). The negotiation of meaning in aviation English as a lingua franca: A corpus-informed discursive approach. *Modern Language Journal*, *105*(3), 639-654. https://doi.org/10.1111/modl.12718

Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? *Language Testing*, *35*(3), 403-426. https://doi.org/10.1177/0265532218758127

Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics*, *32*(3), 23.1-23.17. https://doi.org/10.2104/aral0923

Linde, C. (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. *Language in Society*, *17*(3), 375-399.

Lockwood, J. (2022). The design of a webchat assessment framework for contact centres in Asia. *Sociolinguistic Studies*, *16*(1), 39-63.

Orr, S. (2003). *Hanging on the telephone: A sociophonetic study of speech in a Glaswegian call center*. Unpublished MA Dissertation, University of Glasgow.

Philips, D. (1991). Linguistic security in the syntactic structures of air traffic control English. *English World-Wide*, *12*(1), 103-124. https://doi.org/10.1075/eww.12.1.07phi

Poster, W. (2007). Who's on the line? Indian call center agents pose as Americans for U.S.-outsourced firms. *Industrial Relations*, *46*(2), 271-304. https://doi.org/10.1111/j.1468-232X.2007.00468.x

Prado, M., & Tosqui-Lucks, P. (2019). Designing the radiotelephony plain English corpus (RTPEC): A specialized spoken English language corpus: Towards a description of aeronautical communications in non-routine situations. *Research in Corpus Linguistics*, *7*, 113-128. https://doi.org/10.32714/ricl.07.06

Quaglio, P. (2009). *Television dialogue: The sitcom Friends vs. natural conversation*. John Benjamins.

Römer, U. (2009). English in academic: Does nativeness matter? *Anglistik: International Journal of English Studies*, *20*(2), 89-100.

Schegloff, E. A. (2001). Accounts of conduct in interaction. Interruption, overlap, and turn-taking. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 287-321). Kluwer/Plenum. https://doi.org/10.1007/0-387-36274-6_15

Staples, S. (2015). Examining the linguistic needs of internationally educated nurses: A corpus-based study of lexico-grammatical features in nurse-patient interactions. *English for Specific Purposes*, *37*, 122-136. https://doi.org/10.1016/j.esp.2014.09.002

Staples, S. (2019). Using corpus-based discourse analysis for curriculum development: Creating and evaluating a pronunciation course for internationally educated nurses. *English for Specific Purposes*, *53*, 13-29. https://doi.org/10.1016/j.esp.2018.08.005

Taylor, P., & Bain, P. (2005). 'India calling to the far away towns': The call center labour process and globalization. *Work, Employment, and Society*, *19*(2), 261-282. https://doi.org/10.1177/0950017005505317

Thompson, P., & Friginal, E. (2020). Introduction to Applied Corpus Linguistics [online]. https://www.sciencedirect.com/journal/applied-corpus-linguistics

Eric Friginal is Professor and Head of the Department of English and Communication at The Hong Kong Polytechnic University. Before moving to Hong Kong, he was Professor of Applied Linguistics at Georgia State University. He specializes in applied corpus linguistics, language policy and planning, technology and language teaching, sociolinguistics, discipline-specific writing, and the analysis of spoken professional discourse in the workplace. He has two forthcoming edited volumes on aviation communication: *Global Aviation English Research and Teaching and Assessment in Global Aviation English* (Bloomsbury) co-edited with Malila Prado and Jennifer Roberts. He is founding co-editor-in-chief of *Applied Corpus Linguistics (ACORP) Journal.*