# Theory and Methods in Pragmatics

Dr Vittorio Tantucci
*Senior Lecturer in Linguistics*
*C45 County South*
*Department of Linguistics and English Language*
*Lancaster University*
*LA1 4YL*
*http://www.lancaster.ac.uk/linguistics/about-us/people/vittorio-tantucci*

## Table of contents

# Chapter 6
# Corpus and computational pragmatics

## 6.1 Theory

In this chapter we will learn about **corpus pragmatics** (**CrP**) and **computational pragmatics** (**CmP**). For CrP, we will first need to have a grasp of the main characteristics of corpus linguistics, and then see how these can be applied to pragmatics' research. This will mainly involve the way linguistic data is retrieved and selected for a research project. The latter notion of CmP will then come in handy for the quantification and statistical analysis of the data that we collected and its relationship with the context in which it was produced.

### 6.1.1 Corpus Linguistics

**Corpus linguistics** (**CL**) is quite a young discipline. This is due to the increasing availability of computers starting from the second half of the twentieth century. CL is the study of written or spoken linguistic material that has been recorded in the form of one or a plurality of texts. The main tenet of CL – one that makes it different from any other approach to language use – is that it provides the analyst with linguistic material that is collected in contexts with minimal experimental interference. This means that the language contained in a corpus is not, or only marginally, elicited.

The first electronic corpus, the Brown Corpus, was compiled in the 1960s. It included one million words and was divided into a number of genres that were meant to represent the written American English language of the time (Francis & Kučera 1964). In recent years, also due to constant advances in computer technology, corpora started to become "freely available online to the casual browser, language learner and relatively novice student" (Anderson & Corbett 2009). Even Google published its own large-scale corpus, the Google Ngram Viewer, an online resource containing hundreds of millions of books in a number of languages (see Michel et al. 2011). A relatively accessible range of different kinds of English corpora is available on the platform english-corpora.org[1]. The architecture and web interface of these corpora include both diachronic and synchronic texts that comprise a diversity of genres and styles, such as the Early English Books Online, the Corpus of US Supreme Court Opinions, the Corpus of Historical American English, the Coronavirus Corpus and so on. Corpora can be composed of raw or annotated text. For instance, annotation entails adding linguistic and interpretive data to an electronic corpus of spoken and/or written language data. In most cases, corpora are annotated based on part-of-speech (POS) tagging, as each word is computationally assigned to its grammatical class.

#### 6.1.1.1 The importance of frequency

After the 1960s, corpora rapidly became an invaluable tool for linguists working in a variety of fields, including lexicographers, grammarians, discourse analysts, sociolinguists, but also scholars interested in language teaching, literary studies, translation studies, forensics and, most crucially, pragmatics (cf. O'Keeffe & McCarthy 2010). This "corpus revolution" (e.g. Crystal 2003: 448; Tognini Bonelli 2012: 17) led to a quantitative turn in linguistic analysis, as language could be now searched on a large scale so that patterns that could not be seen before could now be identified and studied (Tognini Bonelli 2012: 18).

Corpus data immediately revealed a fundamental aspect of language that was not at the core of linguistic enquiry prior to the 60s: **frequency**. In particular, high frequencies of patterns of use showed how language is mostly formulaic, that is made of highly conventionalised expressions. This was captured by Sinclair's **idiom principle** (1991) which emphasises that texts are largely composed of ready-made expressions that are chosen as single units, even though they might appear to be analysable into segments (Sinclair 1991: 110) or words. This means that languages are organised in a way that certain expressions favour certain collocates rather than others (a concept that also justifies the existence of linguistic constructions as conceptual entities, e.g. Goldberg 1996, 2018; Croft 2001). Think about the speech act of a greeting, for instance. English speakers favour the expression *good morning* as a greeting when people meet before noon. However, this is not as obvious as it may seem, as many other possibilities could be semantically as plausible, such as *a pleasant*, *fine*, or *enjoyable morning* (e.g. Chanturia & Martinez 2015). And yet, the following dialogic pair, despite being perfectly understandable, is completely unidiomatic, or in other words not a conventional way to perform a greeting among native speakers of English:

(1)
[*Two university colleagues bumping into each other on their way to their offices*]
? A: Pleasant morning!
? B: Pleasant morning!

[1] Last accessed: 10/03/2023.

Despite the countless combinations among words in a language, the expressions that speakers use are for the most part restricted to fixed or semi-fixed formulas (or constructions), which are, in most cases, idiomatic. These are effectively conventions of use, what people do. They 'sound right' because most people talk/write like that. Doing it differently sounds inaccurate or even incorrect, such as writing *Great wishes* instead of *Best wishes* as a salutation at the end of a formal email, or *responding the door* rather than *answering it* when receiving a parcel. One of the greatest contributions of corpus linguistics has been the one of showing empirically – and on a large scale – that linguistic conventions and formulaic expressions are not exceptional, but rather the pillars upon which texts and interactions are organised and performed. This is true for any register, both written and spoken.

In some way, corpora based on spoken data are perhaps the most valuable ones when it comes to the study of language use, as they include interaction that occurred during the here-and-now of a speech event. However, as one may imagine, their compilation is much more challenging: they are more time consuming (Rühlemann & Aijmer 2015), they involve a number of ethic considerations such as speakers' privacy and consent (e.g. Mcenery & Hardie 2012), just to name a few. For these and many other reasons, spoken corpora traditionally tended to be much smaller in size. One early example is the London–Lund Corpus of Spoken English (LLC) from the 1960s and 1970s, notoriously used to study discourse markers and conversational routines (Aijmer 2002), but which was limited to half a million words. We can now find very large collections of spoken data, such as the spoken components of the British National Corpora (BNC and the BNC14) or the Corpus of Contemporary American English (COCA). Research based on multimodal corpora is also growing exponentially, with a clear focus on the interplay between linguistic and non-linguistic semiotic systems (e.g. Carter & Adolphs 2008).

### 6.1.2 Corpus Pragmatics (CrP)

Corpus pragmatics (CrP) – as the name suggests – combines pragmatics' theory with a corpus linguistics' methodology. As for any other approach to pragmatics, CrP focuses on the relationship between language and context. However, it does so not simply with 'sound' examples that are invented by the researcher. It is rather based on naturalistic utterances, i.e. ones that have been realised spontaneously in a written or a spoken form and that have been subsequently recorded in a corpus. Many would argue that CrP is but one possible method to do pragmatics. However, we can also to see it as a unique theoretical perspective that focuses on 'what people actually do' with language in context, rather than 'what would we suppose them to do' – in the case of data generated by intuition – or 'ask/bring them to do' – in the case of experimental data.

It is only recently that approaches from corpus linguistics have been used in research on pragmatics, with influential edited collections (e.g., Felder et al. 2011; Taavitsainen et al. 2014; Aijmer & Rühlemann 2015) and the creation of a new dedicated journal: *Corpus Pragmatics*. Despite this recent transition, corpus data have already become one of the preferred sources for publications in linguistic venues that involve pragmatic analysis, e.g. *Journal of Pragmatics, Pragmatics, Intercultural Pragmatics, Pragmatics, Journal of Historical Pragmatics* and so on. This is partly due to the development of new corpora which are tagged in a way that pragmatic phenomena can be studied in different text types and situations, such as the British Component of the ICE family (ICE-GB, as in International Corpus of English), that is 'parsed' so that all texts are automatically segmented into constituents, such as clauses and phrases (Hunston 2002: 19).

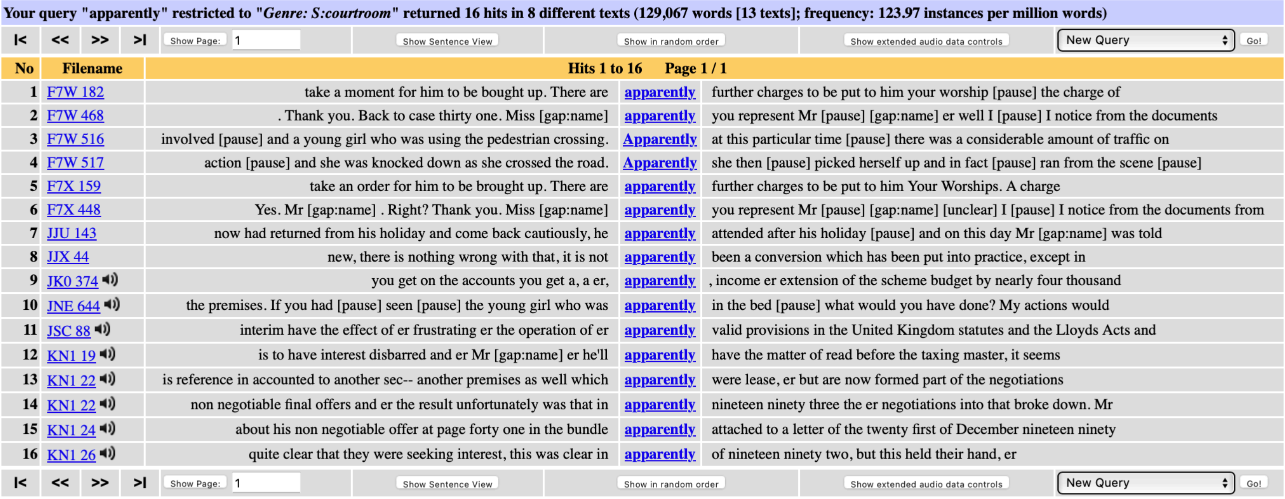
Another factor for why CrP took some time to gain popularity is that corpus linguists often approach language data in a bottom-up manner: they start with forms rather than functions

(Rühlemann & Aijmer, 2015; Aijmer 2018; O'Keffe 2018; Clancy & O'Keffe 2019). Imagine that your aim is to find out how Members of Parliament express uncertainty in a parliamentary corpus, e.g. the Hansard corpus of British Parliament speeches[2]. One thing that you cannot do is searching for *expressions of uncertainty*, as if you were looking for the noun *dog* or the adjective *interesting*. A traditional CL approach would then be one of using a corpus software to automatically generate a list of adjectives and adverbs that are semantically related to uncertainty (e.g. *unlikely, possible, difficult, uncertain, feasible, possibly, maybe* and many others), ranked by frequency. A second step would then analysing the relationship between those forms and their contexts of use. However, some pragmaticians might not find this strategy to be completely satisfactory. In fact, one drawback is that we would only be able to capture parts of speech – in this case adjectives and adverbs – rather than any complex expression or specific strategy that may convey the same function, such as *it is quite a possibility that* or *I suspect that*, but also expressions of certainty such as *surely* which, in turn, may be used in contexts of sarcasm or mockery. Similarly, certainty may also arise out of conversational implicatures that could be inferred only due to ad hoc contextual cues, e.g. after the speaker has provided a number of factual examples that may strongly suggest that s/he is certain of what s/he is claiming. Finally, even more challenging would be searching for other pragmatic phenomena that are often prone to conversational implicatures, such as indirect speech acts, perceptions of (im)politeness or engagement, to name a few. There is thus the necessity to exploit corpus linguistic techniques to study pragmatic phenomena that are formally not restricted to formulaic or fixed expressions. In other words, as language use and 'implied meaning' are at the core of empirical work in pragmatics, it is key to think of effective methods to accomodate both form to function and function to form approaches to language corpora (cf. sections 6.3.1, 6.3.2).

### 6.1.3 Vertical and horizontal reading in CrP

CL is traditionally centred on a 'vertical reading' (Tognini Bonelli 2010) of the hits that are generated from a specific query. Vertical-reading approaches are often performed to spot **key words in context** (**KWIC**). This is also a method that is referred to as concordance line display. Corpus software is thus used to scan through the texts in search for a word (or a larger expression) and displays it in the center of the concordance line along with limited amounts of **co-text** (a surrounding portion of text) to either side. This method is particularly useful to study the immediate co-text of formulaic expressions or single words, as well as their context-driven polysemies. Each line that is generated by a corpus search, is called a **concordance**. Consider the snapshot about the concordances of *apparently* from the courtroom speech section of the BNC in Figure 1:

[2] https://www.english-corpora.org/hansard/. Last accessed: 02/02/2023.



Your query "apparently" restricted to "*Genre: S:courtroom*" returned 16 hits in 8 different texts (129,067 words [13 texts]; frequency: 123.97 instances per million words)

| No | Filename | Hits 1 to 16 Page 1 / 1 | | |
|---|---|---|---|---|
| 1 | F7W 182 | take a moment for him to be bought up. There are | apparently | further charges to be put to him your worship [pause] the charge of |
| 2 | F7W 468 | . Thank you. Back to case thirty one. Miss [gap:name] | apparently | you represent Mr [pause] [gap:name] er well I [pause] I notice from the documents |
| 3 | F7W 516 | involved [pause] and a young girl who was using the pedestrian crossing. | Apparently | at this particular time [pause] there was a considerable amount of traffic on |
| 4 | F7W 517 | action [pause] and she was knocked down as she crossed the road. | Apparently | she then [pause] picked herself up and in fact [pause] ran from the scene [pause] |
| 5 | F7X 159 | take an order for him to be brought up. There are | apparently | further charges to be put to him Your Worships. A charge |
| 6 | F7X 448 | Yes. Mr [gap:name] . Right? Thank you. Miss [gap:name] | apparently | you represent Mr [pause] [gap:name] [unclear] I [pause] I notice from the documents from |
| 7 | JJU 143 | now had returned from his holiday and come back cautiously, he | apparently | attended after his holiday [pause] and on this day Mr [gap:name] was told |
| 8 | JJX 44 | new, there is nothing wrong with that, it is not | apparently | been a conversion which has been put into practice, except in |
| 9 | JK0 374 🔊 | you get on the accounts you get a, a er, | apparently | , income er extension of the scheme budget by nearly four thousand |
| 10 | JNE 644 🔊 | the premises. If you had [pause] seen [pause] the young girl who was | apparently | in the bed [pause] what would you have done? My actions would |
| 11 | JSC 88 🔊 | interim have the effect of er frustrating er the operation of er | apparently | valid provisions in the United Kingdom statutes and the Lloyds Acts and |
| 12 | KN1 19 🔊 | is to have interest disbarred and er Mr [gap:name] er he'll | apparently | have the matter of read before the taxing master, it seems |
| 13 | KN1 22 🔊 | is reference in accounted to another sec-- another premises as well which | apparently | were lease, er but are now formed part of the negotiations |
| 14 | KN1 22 🔊 | non negotiable final offers and er the result unfortunately was that in | apparently | nineteen ninety three the er negotiations into that broke down. Mr |
| 15 | KN1 24 🔊 | about his non negotiable offer at page forty one in the bundle | apparently | attached to a letter of the twenty first of December nineteen ninety |
| 16 | KN1 26 🔊 | quite clear that they were seeking interest, this was clear in | apparently | of nineteen ninety two, but this held their hand, er |

Figure 1.
Concordances of *apparently* from the Courtroom section of the BNC

The adverbial *apparently* is polysemous as it can either refer to what is available ‘at sight’ or to what can be inferred or reported based on some external evidence, thus expressing indirect or interpersonal evidentiality (Aikhenvald 2004; Tantucci 2013, 2016). Now, corpus evidence in Figure 1 clearly shows that in a context that is highly dependent on probatory evidence, such as Courtroom speech, *apparently* shows a clear tendency to express an evidential meaning rather than a sensorial one, as the co-text of all the 16 concordances demonstrates. Importantly, the concordance results also include statistical information such as **raw frequencies** (how many times that form appears in that corpus-section) and **normalised frequencies**, how many times one would ‘virtually’ expect to find the same form in a corpus with same characteristics but with a size that is kept constant, e.g. one made of 10,000 or 1,000,000 words (cf. section 6.3.1.1). In the case of the CQPweb version of the BNC, this is reported at the top of the screen, as shown in Figure 1.

The KWIC method of concordance visualisation is also useful to identify the type of collocates that are favoured by some words or formulaic patterns. Consider the case of the expression _Shut the **_[3]_!_, which can be used as a directive face threat when collocating with *fuck up*, but can also be used as an expressive speech act to positively convey surprise when collocating with *front door*. The use of each collocate may be highly dependent on both co-text and context and therefore being remarkably informative once these are taken into account as part of the corpus search.

You may now be ready for Exercise 1 at the end of this chapter.

In corpus pragmatics the traditional vertical reading needs to balanced with further horizontal reading of contextual and co-textual aspects of speech events (Rühlemann & Aijmer 2015). This involves taking into account larger portions of text in which an utterance is located, the genre and situatedness of speech event, but also – whenever possible – the demographics of the interlocutors, their social and epistemic status (e.g. Heritage 2012; Tantucci et al. 2022) and multimodal aspects of the interaction such as gestures, gaze, prosody and so on.

[3] Each asterisk * expresses an empty collocate in a corpus query search, as for any word that could appear in that specific position.

### 6.1.4 Computational Pragmatics (CoP)

A simple yet sound definition of **computational pragmatics** (**CmP**) is pragmatics with computational means (Bunt 2017: 326). Intuitively, CmP has much in common with corpus pragmatics, as in most cases it is adopted for inferential analysis and making predictions out of data that have been gathered via CrP means. It typically includes corpus data and computational modeling of context-dependent utterance generation and interpretation. Depending on the variety of information contained in a corpus, CmP allows the researcher to account for variables that contribute to the emergence of meaning and behaviour that are inherently dependent on context. Now, the notion of context is never an easy one, as it involves cultural, institutional, interpersonal, demographic and co-textual variables of any sort. Capturing context in its entirety is therefore perhaps an impossible task. However, computational methods of analysis and data manipulation can be key for modeling as many aspects of context as corpus data can provide. The relationship between CrP and CmP could then be typically be seen as a sequential one. A researcher gathers some data that include pragmatic information from a corpus first (CrP), and then may want to rely on computational analysis to draw inferences or generate predictions about certain characteristics of such context-dependent data (CoP).

Since the late 80s, there has been a vast body of work on the design of predictive models of logic simulations that generate inferences based on certain propositional premises. For instance, Stickel (1988) implemented a model for the weighted generation of propositional inferences. Similarly Hobbs et al. (Hobbs 1990; Hobbs et al. 1993) applied a similar methodology to a variety of context-dependent semantic phenomena. Research on natural language processing and engineering has been geared towards the development of systems that could reason “in a way that allows machines to interpret utterances in context” (McEnery 1995: 12). Machine-learning approaches have been adopted to produce rules to be applied to large bodies of data (cf. Jurafsky et al. 1997). However, one limitation of this is the reduction of computational processing to some kind of black box, with the risk of overgeneralising (Weisser 2016) ad hoc implicatures. Bayesian models have recently been adopted to model uncertainty and the updating of beliefs in light of new information based on the listener’s knowledge about the context of the speech (e.g. Franke & Jäger 2016; Goodman & Frank 2016).

While the logic modeling of implicatures and inferences is a compelling area of CoP, statistical methods have also been adopted for the analysis of other pragmatic phenomena, such as conventionalised speech act types, (im)politeness reciprocity, engagement and so on. Tantucci & Wang (2018, 2020a, 2020b) developed a computational framework for the study of illocutional concurences, which relies on machine learning and classificatory models, such as conditional inference trees or hierarchical clustering (Levshina 2015, 2021) to identify behavioural and linguistic associations that lead to context-situated speech acts (see also Van Olmen & Tantucci 2022; Põldvere et al. 2023). A computational approach to engagement in interaction has been developed by Tantucci & Wang (2021, 2022a, 2022b, 2022c) so that utterances that repeat and re-adapt components from an interlocutor’s turn at talk (e.g. words or structures) can be quantified on a large scale (cf. section 6.4.1.1). Large-scale annotation of resonance can be quantified numerically and therefore fitted as the outcome variable of multiple linear regression modeling (Tantucci & Wang 2022a).

### 6.2 Methods

In this section, I will illustrate the two main methodologies that characterise corpus pragmatics, namely, the form to function approach (section 6.3.1), and the function to form approach (section

6.3.2). Each method will be first introduced theoretically, and then put into play with an illustrative case study from the literature (respectively in sections 6.3.1.1 and 6.3.2.1).

### 6.2.1 Form to function approach

Corpus methods that search for specific formal expressions and subsequently analyse their functional characteristics are called **form to function** approaches. These have been – and still are – quite central in most corpus linguistic enquiries (O'Keffe 2018). In pragmatics' research, these involve the analysis of expressions that can be easily queried in a corpus search browser, such as response tokens (e.g. *really, yes, fine*), pragmatic markers (*actually, believe it or not, as it seems*), vocatives (*Mary, professor*), and so on. The focus of form to function approaches is thus on specific lexemes (or more abstract constructions that can be retrieved via corpus-queries), rather than the whole range of linguistic behaviour in context. The form to function framework allows to make generalisations about specific words and/or formulaic expressions, but does not allow the researcher to make broader claims about what people say or do in certain situations.

In defence of this methodology it can be said that while implied meaning often extends beyond what is said, nonetheless specific expressions often are associated with highly conventionalised implicatures or may also unambiguously express the intentions of the interlocutor (Rühlemann 2018). For instance, while *shut up* may be used as a polite response to tone down a compliment, it is most frequently uttered as a directive speech act of impoliteness, a usage where there is hardly any alternative meaning to be derived. In this sense, we can rely on specific queries to find instances where a particular expression may or may not generate context-based implicatures. The same can be said about pragmatic markers such as *well* (Levinson 2013: 108) or *look* (Tantucci 2021; Van Olmen & Tantucci 2022). For example, when *look* occurs in turn-initial position, it often functions as an 'attention getter' and projects a potential negative reaction to what is about to be said, as in *Look Dani, you don't know what you're speaking about* (Tantucci 2021: 26). If our aim was then to analyse the attention getting function of *look*, we could then first identify specific context types in a corpus (e.g. GP consultations, teacher-pupils exchanges in the classroom, and so on), search for usages of turn-initial *look* occurring in each context of interest, manually discard unwanted hits (most of which will be imperative usages of *look*), and, finally, investigate the kinds of speech acts that *look* contributes to express as a turn-initial pragmatic marker. We could similarly be interested in whether *look* tends to be used by speakers who are in a position of relative power, or whether it is bound to specific social distance conditions and so on.

All of these questions would be involved in a form to function approach to corpus pragmatics. This kind of analysis has been a stand stone in historical pragmatic research, as it allows to study the formation of procedural meanings of pragmatic markers out of words or larger constructions at different points in time (cf. Traugott & Dasher 2002; Traugott 2016, 2019; Tantucci 2017a, 2017b). It has also been used to study cross-cultural and cross-linguistic mismatches of forms originating from similar semantic sources in different languages (e.g. Van Olmen & Tantucci 2021). It has also been key for studying children's ontogenetic ability to master increasingly intersubjectified functions of specific forms, (e.g. from *look* used to direct visual attention to *look* used as a turn-initial pragmatic marker, or the shift from aspectual to evidential usages of the post-verbal marker 过 *guo* in Mandarin, Tantucci 2021).

#### 6.2.1.1 Form to function: Case study

Now it is time to put some the stuff we discussed into play. As a case study of a form to function approach, we can look at Torgersen et al. (2017) who examine the pragmatic marker *you get me* among young speakers in inner London. They used a 2.8 milion words corpus including two datasets: i. The English of adolescents in London (2004-2007, LIC), ii. The Multicultural London English (2007-2010, MLEC). Among other things, they were interested in the parts of the city where new usages of *you get me* take place and who tended to be the linguistic innovators. Extracts (2) to (5) display typical uses of *you get me* in the data:

(2) Dave: yeah and that they call me a mummy's boy. I don't care. it's my mum **you get me**.
Sue: [mm call me what you want...I'm the one that's still at home. all the luxuries and they're out there. no money yeah each week. scraping through.

(3) Ferda: <tsk> but she looked like twenty. cos if she was even though she was still thirteen even though I knew she was I would still go for her...yeah she might be thirteen but she's got the mouth. **you get me**?
Chelsea: true.
Lucinda: true.

(4) David: I don't care bruv...**you get me**? that's how cowardly you are you gonna stab me over a phone.

(5) Omar: I see where you're coming from and I see
David: [**You get me,** I see where you're coming from but.

(Torgersen et al. 2017:181)

In examples (2) and (3), *you get me* occurs as a marker of expected agreement (cf. Tantucci 2017a). Example (4) is somewhat different in that the hearer does not respond to the utterance. This is a clear indicator that *you get me* in this case has fully developed into a pragmatic marker, as it is no longer used as a separate proposition, but as a metalinguistic device, preemptively assuming that the hearer will be in agreement with what is said (Traugott, 2010; Fitzmaurice, 2004; Tantucci 2017b; Tantucci & Di Cristofaro 2022). Finally, in (5) the PM is used to comment on something another speaker has said. This is a case where *you get me* has lost its compositional meaning (Traugott & Trousdale 2013) as it is now used as a response token.

An important variable in the study by Torgersen et al. (2017) was the cultural background. Half of the speakers had a 'white London' background, as their families had lived in the area for at least three generations. This group of speakers was labeled as 'Anglo', whereas the second half were the children or grandchildren of immigrants and were termed 'non-Anglo'. The sociolinguistic distribution of speakers in MLEC2 and LIC3 is given in Table 1:

| | MLEC2 | LIC3 |
|---|---|---|
| No. of words | 194,236 | 457,812 |
| No. of speakers | 25 | 51 |
| Data collection period | 2008 | 2005 |
| Data collection method | Sociolinguistic interviews | Sociolinguistic interviews |
| Age | 16–19 (average 17) | 16–18 (average 17) |

| | | |
|---|---|---|
| Sex | female; male | female; male |
| Ethnicity | Anglo; non-Anglo (but different ethnicities from LIC) | Anglo; non-Anglo |
| Residence | North and West of Inner London (Hackney, Haringey, Islington) | Inner London (Hackney) |
| Social class | Working class | Working class |

Table 1.
Sociolinguistic distribution of speakers in the MLEC2 and LIC3 subcorpora

Adapted from (Torgersen et al. 2017: 182)

Qualitative analysis of the co-text is crucial for the identification of PM usages of different variants of *you get me* (e.g. *do you getting me, if you get me*, *get me* and so on). Based on the distribution of different variants of *you get me* as a PM in the two datasets, Torgersen et al. found that non-Anglo speakers showed a remarkable preference for the PM in contrast with Anglo speakers. In MLEC2 (North and West of Inner London) it was 93.8% and in LIC3 (Inner London) it was also very high: 82.3%. In Table 2 below are reported the raw frequencies of the variants of *you get me* both by Anglo and non-Anglo Speakers, together with the total number of words in each sub-corpus. In the last column are reported the per-milion-word (pmw) frequencies of *you get me*. As discussed in section 6.1.3, pmw frequencies are calculated by dividing raw frequencies by the total number of words of the (sub-)corpus, and then by multiplying the result per 1 million. For instance, the pmw frequency of Anglo speakers in the MLEC2 corpus is obtained by dividing the raw frequency (4) by the total number of words in that section of the corpus (56,010), and then by multiplying the result per 1 million, so as to obtain the pmw frequency of 71.4 (4/56,010*1000,000,000 = 71.4). Pmw frequencies are called normalised, in the sense that they provide a value that is constant independently from the size of the (sub)corpus of interest.

| | Population | Raw Frequency | (Sub)corpus size | Pmw frequency |
|---|---|---|---|---|
| MLEC2 | Anglo | 4 | 56,010 | *71.4* |
| | Non-Anglo | 120 | 138,226 | *868.1* |
| LIC3 | Anglo | 16 | 154,019 | *103.9* |
| | Non-Anglo | 125 | 303,793 | *411.5* |

Table 2.
Raw and pmw frequencies of Anglo vs Non-Anglo speakers in MLEC2 and LIC3

We are now at a stage where Torgersen et al. had some of their data classified and categorised. At this point, what they needed to do was to prove whether those differences in frequency are statistically significant, and, therefore, worthy of being reported in an academic venue.

A most common way to calculate statistical significance in form to function approaches is by inserting the raw frequencies of a form in a contingency table[4], together with total word counts of the (sub-)corpora where the construction is used. For instance, based on the data in Table 2, it is

[4] 2x2 contingency tables are used the most, but larger ones can also be used, e.g. when more than 2 (sub-)corpora are compared.

possible to compare all the usages of *you get me* by speakers from the MLEC2 corpus (124), with the ones by the LIC3 speakers (141) by accounting for each corpus size, respectively 194,236 and 457,812, as shown in Table 3 below:

| | |
|---|---|
| *You get me* in MLEC2 (124) | *You get me* in LIC3 (141) |
| Total words in MLEC2 (194,236) | Total words in LIC3 (457,812) |

Table 3.
Contingency table of the usages of *you get me* in the MLEC2 and LIC3

Once we have these 4 values, we can use various measures of association, e.g. the $X^2$ test, the Fisher exact test, the $G^2$ test and so on (cf. Levshina 2015). These are used to assess whether there is a significant association among the different frequencies that are reported. While differences exist among these various statistical algorithms (e.g. they can be based on real vs inferential computations, some perform worse with lower frequencies than others, and so on), the nature of this data is suitable for most of them. In the case of this study, Torgersen et al. found that MLEC2 speakers use much more *you get me* as a PM compared to LIC3 ($G^2 = 33.79$, $p < 0.01$), twice as much in fact. To explain, the $G^2$ value (33.79) is high enough to obtain a statistically significant result, which is conventionally achieved when the probability value of the computation is lower than 0.05 ($p < 0.01$ in this case). What this means is that there is (at least) 95% probability that a difference in frequency is not due to chance. In this case, it could be concluded that there is a significantly more widespread usage of the construction in the North and West of Inner London than the Hackney area. The Rstudio code to perform a $G^2$ test based on these frequencies is the following (all you need to do is to replace the frequencies in the third line of the code with the ones that you obtained in your own case-study):

```
> install.packages(“RVAideMemoire”)
> library(RVAideMemoire)
> mytable = matrix(c(124,141,141,457812),nrow = 2, ncol = 2, byrow = T)
> G.test(mytable)
```

You may now be ready for Exercises 2 and 3 at the end of this chapter.

### 6.2.2 Function to form approach

As noted by Landert et al. (2023), most of the tools that are used to search corpus data have been developed for the study of grammar and lexicon: e.g. concordances, collocations, n-grams and keywords. While these options inform traditional form to function approaches, the options for function to form approaches are currently much scarcer.

The **function to form** approach starts from a function and investigates the forms that are used to perform it. An example of relatively common form-to-function methods is in Garcia McAllister (2015), who performed “a line-by-line reading of the corpus conversations to identify speech acts within Searle’s speech act categories (i.e., directives, commissives, expressive, etc.) as they occurred in context” (Garcia McAllister 2015: 34). In this way, different speech acts can be found to be used in distinct subcorpora. Most importantly, unexpected ways to express those speech acts in context can be discovered.

Function to form approaches are adopted with corpora that have been previously annotated based on a number of pragmatic features, as in the co-called Narrative Corpus (Rühlemann & O'Donnell 2012). This, in addition to speakers' socio-demographic information and classic part-of-speech annotation, also includes tagging for quotatives, constructed dialog, and participant roles. Similarly, the SPICE Ireland corpus is famously tagged for prosody, speech acts, and pragmatic markers and includes meta-information related to speakers' socio-linguistic variables, among other things. This kind of corpus design has also facilitated a new socio-pragmatic turn in CrP, as more and more corpora started to provide information about the speakers' age, gender and class (Macaulay 2005). This strand of function to form approaches is a fascinating one, however it also has some disadvantages, such as the fact that the analyst must abide to categories that have been decided a priori by the corpus developers. It is also quite resource-costly and thus mostly amenable to relatively small corpora (Rühlemann 2018).

### 6.2.2.1 Form to function: Case study

We can now look at a recent method that is illustrative of the form to function framework, which is one accounting for engagement in dialogue by means of dynamic resonance (Tantucci & Wang 2022a; Tantucci et al. 2018). Dynamic resonance (a notion originally formalised in Du Bois 2014) occurs when speakers re-use words and expressions uttered by an interlocutor in order to express something new. This – on a large scale – is an important indicator of proactive engagement with other peoples' talk. Persistently low levels of dynamic resonance, on the other hand, underpin interactional detachment and are more distinctive of atypical speech (Hobson et al. 2012; Tantucci & Wang 2022b).

As a case study, we will focus on how resonance varies cross-culturally and cross-linguistically. In Tantucci & Wang (2022a) two sets of 1,000 utterances were compared, involving either agreement or disagreement from two comparable corpora of naturalistic interaction among family members in Mandarin Chinese and American English. These data were retrieved from the Callhome corpora of spontaneous interaction of Mandarin Chinese and American English, consisting, respectively, of 120 unscripted telephone conversations between native speakers, comprising 250,000 words each. A case of resonance involving disagreement is the one in (6) below:

(6) A: Because I don't have anyone to talk to.
B: Oh, come on, **you're kidding**, right?
A: No, **I'm not kidding**.

Callhome / Eng / 4485

In example (6), speaker A resonates with B in a way that formal similarity across turns is paired with specific rhetoric strategies. On the one hand, B's utterance is in disagreement with A, on the other, it overtly engages with A's speech, as it proactively resonates with what was said, i.e. it re-uses some of the words and/or constructions from the preceding turn. A way to quantify dynamic resonance in cases such as in (6) is to identify utterances where at at least one word is repeated after the preceding turn. In this case that would be *kidding*. Syntactic resonance can thus be assessed by counting the number of constituents of the construction that allows formal variation from the expression of speaker A to the one of speaker B, which in this case involves the [SUBJ BE *kidding*] structure, made of 3 constituents.

|  | SUBJ | BE | *kidding* |
|---|---|---|---|
| A: | *You* | *'re* | *kidding* |
| B: | *I* | *am (not)* | *kidding* |

Table 4.
Dynamic resonance of [SUBJ BE *kidding*]

Form to function approaches work best when they are based on multifactorial annotation, which means that some outcome variable (dynamic resonance in this case) can be predicted by two or more variables. In the case of Tantucci & Wang's study, the annotation included:

i. whether the utterance was one of agreement vs disagreement;
ii. the language (Chinese or English);
iii. whether or not pragmatic markers (PM) were present either at the beginning or at the end of the sentence;
iv. which PMs they were;
v. the source of resonance (i.e., whether speaker B would resonate with speaker A, with himself or herself, or with both);
vi. the degree of resonance occurring lexically (how many words were repeated), the one of resonance occurring syntactically;
vii. the distance from the original form and the resonating one (measured in intonation units, cf. Chafe 1994).;

Table 5 below provides a sample row (out of 2,000 for the two corpora) of the input of these dimensions with example (6) as a reference:

| (Dis)agreement | Language | PM | PMs | Source | Lexical Res | Syntactic Res | Distance |
|---|---|---|---|---|---|---|---|
| *Disagreement* | *Eng* | *No* | */* | *Other* | *1* | *3* | *2* |

Table 5.
Input sample for the annotation

Something similar to example (6) can be seen in the case of disagreement in Mandarin in (7) below:

(7) A: 给寄了, 他没收到, 不知道怎么回事儿。
gěijì le, tā méi shōudào, bù zhīdào zénme huí shìer
give send PF, he not receive, not know how CLAS[5] thing
'I sent it, he didn't receive it, I don't know why.'
B: 哦, 也不一定没收到呢 。
o, yěbù yīdìng méi shōudào ne
O, also not certain receive NE
'Oh, it might not be the case that he didn't receive it actually.'

Callhome / Chin / 0774

[5] Classifier.

| | SUBJ | PastNeg | 收到 |
|---|---|---|---|
| A: | 他 | 没 | 收到 |
| B: | / | (也不一定)没 | 收到 |

Table 6.
Dynamic resonance of [SUBJ PastNeg 收到]

In this case, B's turn here also involves disagreement and is marked with PMs occurring both at the left (哦 *o*) and right (呢 *ne*) sentence periphery of the utterance. The source of B's resonance is exclusively A's turn, which is then tagged as 'other'. The lexical resonance value of the construct is 2, with the bare repetition of the words 没 *méi* 'not' and 收到 *shōudào* 'receive' from turn A to turn B. The construction [Subj + PastNeg + 收到] includes three components, therefore the value for syntactic resonance is 3. Lastly, distance in this case is equal to 2, since it involves the occurrence of the first intonational unit (IU), 哦 *o*, and the second IU where B resonates with A. The violin plots in Figure 2 show the comparison between English and Chinese speakers with respect to lexical and syntactic resonance, as well as to the linguistic distance between the original construction and the resonating one.

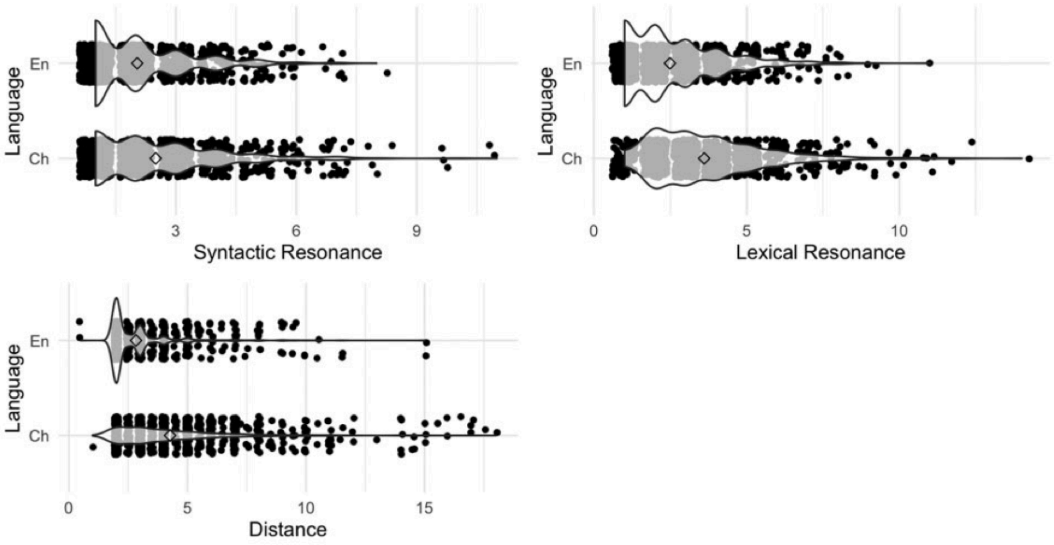


Figure 2.
Violin plots of the difference in the way American English and Chinese speakers resonate through interaction

Each instance of resonance in the two corpora corresponds to a dot in Figure 2. Areas with higher density are represented by wider portions of violin-shaped forms. A diamond-shaped point (◇) represents the mean distribution of the observations and is located in the middle of each plot. What the figure shows is that Chinese interaction in all three cases have a substantially longer distribution and higher means, indicating that resonance among Chinese family members tends occur to a larger degree syntactically, lexically and across longer stretches of interaction. In the study, Tantucci & Wang similarly show that engagement – as a byproduct of resonance – among

Chinese family members significantly correlates with presence of pragmatic markers, while this does not happen for American English speakers.

From this form to function study, we learned that Chinese interaction among family members involves speakers' higher degrees of overt engagement with the speech of the interlocutors in comparison with American English families.

You may now be ready for Exercise 4 at the end of this chapter.

## 6.3 Exercises

1. Go to the registration page of BNCweb (CQPversion) at http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php and register freely for your account. You are now ready to go to the spoken version of the BNC, search for the expression *good evening* and compare it in two different varieties: general conversation vs broadcast news (they appears as 'S:conv' and 'S:brdcast:news' under the genre section). Can you not answer the following two questions?
   - Is the expression always used as a greeting in the two varieties?
   - What formal differences can you identify concerning the way *good evening* is used in each context type?
2. Go to the spoken version of the BNC and select the dialogic context of general conversation (tick on 'S:conv' under the genre section) among male speakers (tick on 'Male' under the 'Sex' subsection). Now search for exclamation marks \! (yes, in the search box any punctuation mark, such as *, . ; :* needs to be preceded by a backward slash \). Now check among the collocates of your query all the adjectives that are used in contexts of appraisals (you can limit your sample to the first 20). Do the same for female speakers and compare the results (a $G^2$ test code for Rstudio, as given at the end of section 6.3.1.1, would be ideal).
3. Inspect Table 1 in section 6.3.1.1, which provides the data about the usage of the pragmatic marker *you get me* among teenage speakers in different parts of London. The main research hypothesis involves demographical differences in the use of *you get me*. By looking at the data appearing in the table, can you identify a potential problem for the following analysis?
4. Go to the spoken version of the BNC and select the dialogic context of interviews (tick on 'S:interview' under the genre section). Then search for any punctuation mark, which in this CQP version of the BNC, is obtained with the string _PUN. This will allow you to visualise all the dialogues that are included in the interviews section of the corpus. Based on section 6.3.2.1, annotate the first 10 complete interactions by quantifying the degree of lexical resonance that is expressed by those speakers, that is how many words are repeated from one interlocutor to another. Now try and do the same for the dialogic context of general conversation (tick on 'S:conv' under the genre section). You can now compare different degrees of lexical resonance in these two different contexts of use. A simple way to do so is to count the number of words of the first 10 exchanges in the interviews' section. Then, you can do the same for the exchanges of spontaneous conversation. You can finally create a 2x2 contingency table just like the one in section 6.3.1.1. In this case, you will insert resonance values at the top two quadrants, and total word counts at the quadrants at the bottom. Finally, you can compute a $G^2$ test based on the code at the end of 6.3.1.1.

## References

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford: OUP.

Anderson, W., & Corbett, J. (2010). Teaching English as a friendly language: lessons from the SCOTS corpus. ELT journal, 64(4), 414-423.

Aijmer, K. 2002. *English Discourse Particles: Evidence From a Corpus* [*Studies in Corpus Linguistics 10*]. Amsterdam: John Benjamins.

Aijmer, K. (2018). Corpus pragmatics: From form to function. In A. H. Jucker, K. P. Schneider & W. Bublitz (Eds.), *Methods in Pragmatics* (pp. 555–586). Berlin: De Gruyter Mouton.

Aijmer, K. & Rühlemann, C. (eds.) 2015. *Corpus pragmatics: A handbook.* Cambridge: Cambridge University Press.

Bunt, H. (2017). Computational Pragmatics. In Y. Huang (Ed.), *The Oxford Handbook of Pragmatics* (pp. 326–345). Oxford University Press.

Carter, R. & Adolphs, S. (2008) 'Linking the Verbal and Visual: New Directions for Corpus Linguistics', 'Language, People, Numbers', special issue of *Language and Computers* 64: 75–291.

Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive linguistics*, 25(3), 359-410.

Felder, E., Müller, M. & Vogel, F. (eds.) 2011. *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin and Boston: Walter de Gruyter.

Fitzmaurice,S. (2004). Subjectivity,intersubjectivity and the historical construction of inter- locutor stance: From stance markers to discourse markers. *Discourse Studies*, 6, 427–448.

Francis, W. N. & Kučera, H. (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Providence, Rhode Island: Brown University.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft,* 35(1), 3-44.

Garcia McAllister, Paula. 2015. Speech acts: A synchronic perspective. In Karin Aijmer & Christoph Rühlemann (eds.), *Corpus pragmatics: A handbook*, 29–51. Cambridge: CUP.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences,* 20(11), 818-829.

Heritage, J. (2012). Epistemics in action: Action formation and territories of knowledge. *Research on language & social interaction*, 45(1), 1-29.

Hobbs, Jerry (1990) An Integrated Abductive Framework for Discourse Interpretation. *Proceedings AAAI Spring Symposium on Abduction*, Stanford, pp. 10-12.

Hobbs, Jerry, Stickel, M. & Martin, P. (1993) Interpretation as abduction. *Artificial Intelligence* 63, 69-142.

Hobson, R. P., Hobson, J. A., García-Pérez, R., & Du Bois, J. (2012). Dialogic linkage and resonance in autism. *Journal of Autism and Developmental Disorders*, 42, 2718-2728.

Hunston, S. 2002. *Corpora in applied linguistics*. Oxford: Oxford University Press.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., ... & Van Ess-Dykema, C. (1997, December). Automatic detection of discourse structure for speech recognition and understanding. In 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (pp. 88-95). IEEE.

Landert, D., Dayter, D., Messerli, T., & Locher, M. (2023). *Corpus Pragmatics (Elements in Pragmatics)*. Cambridge: CUP. doi:10.1017/9781009091107

Levinson, S. C. (2013). Recursion in pragmatics. *Language*, 149-162.

Levshina, N. (2015). How to do linguistics with R: Data exploration and statistical analysis. Amsterdam: John Benjamins Publishing Company.

Levshina, Natalia. (2021). Conditional inference trees and random forests. In Magali Paquot &

Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 611–643. New York: Springer.

Macaulay, R. K. (2005). *Talk that counts: Age, gender, and social class differences in discourse*. Oxford: OUP.

McEnery, T. (1995) *Computational Pragmatics: Probability, Deeming and Uncertain References*. Unpublished PhD thesis. Lancaster University.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176-182.

O'Keefe, A., & McCarthy, M. (eds.). (2010). *The Routledge handbook of corpus linguistics*. Oxford, UK: Routledge.

O'Keeffe, A. (2018). Corpus-based function-to-form approaches. In A. H. Jucker, K. P. Schneider and W. Bublitz (eds) *Methods in Pragmatics.* Berlin: Mouton de Gruyter, 587 – 618.

O'Keeffe, A., Clancy, B., & Adolphs, S. (2019) [2nd ed.]. *Introducing pragmatics in use*. London: Routledge.

Rühlemann, C. (2018). *Corpus linguistics for pragmatics: A guide for research*. London: Routledge.

Rühlemann, C., & O'Donnell, M. B. (2012). Introducing a corpus of conversational stories. Construction and annotation of the Narrative Corpus. *Corpus Linguistics and Linguistic Theory*, 8(2), 313-350.

Stickel, M. E. (1991). A Prolog-like inference system for computing minimum-cost abductive explanations in natural-language interpretation. *Annals of Mathematics and Artificial Intelligence*, 4(1-2), 89-105.

Taavitsainen, I. and A. H. Jucker (eds.) 2014. *Diachronic corpus pragmatics (Pragmatics & beyond new series)*. Amsterdam: John Benjamins.

Tantucci, V. (2013). Interpersonal evidentiality: The Mandarin V-过 guo construction and other evidential systems beyond the 'source of information'. *Journal of pragmatics*, 57, 210-230.

Tantucci, V. (2016). Textual factualization: The phenomenology of assertive reformulation and presupposition during a speech event. *Journal of Pragmatics*, 101, 155-171.

Tantucci, V. (2017a). From immediate to extended intersubjectification: A gradient approach to intersubjective awareness and semasiological change. *Language and Cognition*, 9(1), 88-120.

Tantucci, V. (2017b). An evolutionary approach to semasiological change: Overt influence attempts through the development of the Mandarin 吧-ba particle. *Journal of Pragmatics*, 120, 35-53.

Tantucci, V. (2021). *Language and social minds: The semantics and pragmatics of intersubjectivity.* Cambridge: CUP.

Tantucci, V., & Di Cristofaro, M. (2020). Pre-emptive interaction in language change and ontogeny: The case of [there is no NP]. *Corpus Linguistics and Linguistic Theory*, 17(3), 715-742.

Tantucci, V., & Wang, A. (2018). Illocutional concurrences: The case of evaluative speech acts and face-work in spoken Mandarin and American English. *Journal of Pragmatics*, 138, 60-76.

Tantucci, V., & Wang, A. (2020a). Diachronic change of rapport orientation and sentence-periphery in Mandarin. *Discourse Studies*, 22(2), 146-173.

Tantucci, V., & Wang, A. (2020b). From co-actions to intersubjectivity throughout Chinese ontogeny: A usage-based analysis of knowledge ascription and expected agreement. *Journal of Pragmatics*, 167, 98-115.

Tantucci, V., & Wang, A. (2021). Resonance and engagement through (dis-) agreement: Evidence of persistent constructional priming from Mandarin naturalistic interaction. *Journal of Pragmatics*, 175, 94-111.

Tantucci, V., & Wang, A. (2022a). Resonance as an applied predictor of cross-cultural interaction:

Constructional priming in Mandarin and American English interaction. *Applied Linguistics*, 43(1), 115-146.

Tantucci, V., & Wang, A. (2022b). Dialogic priming and dynamic resonance in Autism: Creativity competing with engagement in Chinese children with ASD. *Journal of autism and developmental disorders,* 1-17.

Tantucci, V., & Wang, A. (2022c). Dialogic priming and dynamic resonance in Autism: Creativity competing with engagement in chinese children with ASD. Journal of autism and developmental disorders, 1-17.

Tantucci, V., Culpeper, J., & Di Cristofaro, M. (2018). Dynamic resonance and social reciprocity in language change: The case of Good morrow. *Language Sciences*, 68, 6-21.

Tantucci, V., Wang, A., & Culpeper, J. (2022). Reciprocity and epistemicity: On the (proto) social and cross-cultural 'value'of information transmission. *Journal of Pragmatics*, 194, 54-70.

Torgersen, Eivind, Costas Gabrielatos & Hoffmann, Sebastian. (2018). A corpus-based analysis of the pragmatic marker you get me. In Friginal Eric (ed.), *Studies in Corpus-Based Sociolinguistics*. Abingdon: Routledge, pp. 176–196.

Traugott, E. C. (2010). (Inter)subjectivity and (inter)subjectification: A reassessment. In Davidse, K., Vandelanotte, L., & Cuyckens, H. (Eds.), *Subjectification, intersubjectification and grammaticalization* (pp. 29–74). Berlin: De Gruyter.

Traugott, E. C. (2012). Intersubjectification and clause periphery. *English Text Construction*, 5(1), 7-28.

Traugott, E. C. (2016). On the rise of types of clause-final pragmatic markers in English. *Journal of Historical Pragmatics*, 17(1), 26-54.

Traugott, E. C., & Dasher, R. B. (2002). *Regularity in semantic change*. Cambridge: CUP.

Traugott, E. C., & Trousdale, G. (2013). *Constructionalization and constructional changes* (Vol. 6). Oxford: OUP.

Van Olmen, D., & Tantucci, V. (2022). Getting attention in different languages: A usage-based approach to parenthetical look in Chinese, Dutch, English, and Italian. *Intercultural Pragmatics*, 19(2), 141-181.

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis.* Chichester, UK. Wiley.