Corpus linguistics and pragmatics

Abstract

Pragmatics and corpus linguistics were long considered mutually exclusive because of their stark methodological differences, with pragmatics relying on close horizontal reading and qualitative interpretation and corpus linguistics typically scanning texts vertically and processing data quantitatively. In recent years, however, corpus linguists and pragmaticists have discovered common ground thus paving the way for the advent of the new field of corpus pragmatics. This chapter takes a meta-methodological approach aiming to show that corpus pragmatics integrates the horizontal (qualitative) methodology typical of pragmatics with the vertical (quantitative) methodology predominant in corpus linguistics. To illustrate, we examine the choice between past indicative was and past subjunctive were in as-if clauses in the Corpus of Contemporary American English (COCA). The methodology integrates quantitative collocational analyses as well as qualitative examination of the two forms in concordance lines and larger contexts. The findings suggest that the choice is co-determined by two factors: the speaker's attitude to the factuality of the comparison expressed in the as-if clause and the speaker's choice of one of two major syntactic functions for the as-if clause to perform in the matrix clause. We finally discuss the possibility of viewing the was/were alternation as a form of negative empathetic deixis (Lyons 1977).

1. Introduction

Corpus linguistics, a relatively young linguistic discipline though its roots can be traced back as far as the thirteenth century (see McCarthy and O'Keeffe 2010), came to the fore in the latter half of the twentieth century. This time period saw the development of more powerful computers and linguists, fascinated by the potential for the analysis of language, began to amass large, principled collections of language in an electronic format, and these became known as *corpora*. A milestone in this process was the first of the modern, machine-readable corpora, the Brown Corpus, a corpus of American English written texts (see Francis and Kučera 1964). The next thirty years saw an explosion of interest in the use of computers to analyse language and a corresponding increase in the size of corpora. It was, however, the emergence of the more affordable, powerful and, in terms of space, manageable personal computers in the 1990s that revolutionised corpus linguistics and shaped it into the discipline we know today. This new-found ability to store large amounts of data that could be quickly analysed heightened the appeal of corpus linguistics and, driven on by the dictum there is 'no data like more data' (Sinclair 2001), the era of mega-corpora such as the Collins Corpus and Bank of English™ (approx. 2.5 billion words), the Oxford English Corpus and the Cambridge English Corpus (both approx. 2 billion words in size) arrived. Mega-corpora were initially compiled for lexicographical purposes; however, somewhat ironically, their development also coincided with the blending of a corpus linguistic methodology with other linguistic frameworks such as conversation analysis, spoken discourse analysis, sociolinguistics and, the focus of this chapter, pragmatics, which, in turn, has resulted in smaller, more domain-specific corpora. These smaller corpora facilitate a "constant interpretive dialectic between features of texts and the contexts in which they are produced" (Vaughan and Clancy 2013: 70) and are therefore ideally suited to the study of pragmatics.

Pragmatics, as we know, has its origins in the philosophy of language and this, coupled with a tendency to focus on individual texts, means that pragmatic research has been primarily qualitative (see Rühlemann and Aijmer 2015). The discipline has, however, always encouraged an on-going critique and redefinition of core concepts within the field (see Lindblom (2001) and Davies (2007) on Grice, for example) and this flexible approach could also be said to have been extended to methodological considerations. No one homogeneous methodology has emerged for the study of pragmatic phenomena. Instead, methodologies from areas such as discourse analysis or conversation analysis have been applied. Similarly, corpus pragmatics has emerged from the blending of pragmatics with a corpus linguistic methodology. Arguably, the most important methodological benefit of corpus linguistics for pragmatics is the empirical nature of many corpus

studies. Although it is acknowledged that corpus studies can also be qualitative, it is the quantitative element provided, for example, by concordancing software that arguably defines the discipline. Corpora exist as electronic text files and are therefore suitable for analysis via concordancing or other corpus software. The corpus entry point for many researchers is the word frequency list. In general, a word frequency list appears visually as a list of all the types in a corpus coupled with the number of occurrences of each type. These frequency counts are referred to as 'raw' and can, in turn, be normalised so that they might be compared to other frequency results from other bigger and/or smaller (in size) corpora. It is also possible to compare one corpus to another using keyword lists. Keyword lists, generated by corpus software, feature items that occur with unusual in/frequency when one corpus is compared to another. Keyword lists provide a measure of saliency rather than simple frequency due to the statistical nature of the process (see Baker 2006). The procedure involves generating a word frequency list for the target corpus and then a word frequency list for a larger reference corpus (for example, the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA)). The computer then processes the statistical significance of difference between the two corpora using chi-square or log-likelihood tests.

The most familiar representation of a corpus linguistic 'vertical reading methodology' (Rühlemann and Aijmer 2015: 7) is the KWIC (key word in context) or concordance line format. Characteristic of the KWIC format is that the search item, or *node*, is presented visually in the centre of the line(s) surrounded by a number of words on either side:

Table 1: Ten concordance lines for *as if he was* generated using the BNC

1	A0E 295	but he looks at him disrespectfully	**as if he was**	applying a Brechtian alienation-effec
2	A52 352	to finger, hand to hand, he looked	**as if he was**	auditioning for the lead role in The C
3	A54 26	expression entered Jackson's eyes	**as if he was**	wondering why nothing ever proved
4	A73 464	comer's face was green. He looked	**as if he was**	about to be sick. 'Will you go down
5	A7A 2635	ed at Marx who stood before them	**as if he was**	quite happy to stand there in silence
6	A7J 1476	st with a curious air of detachment	**as if he was**	an observer from another civilisatio
7	A7L 948	ures. The Americans must have felt	**as if he was**	taking revenge on them for what ha
8	AB5 1288	h of Ziggy Stardust and treated him	**as if he was**	Ziggy Stardust. We began to treat hi
9	ABX 980	Why were they all treating him	**as if he was**	five years old. He felt like screaming
10	ABX 2659	awkward. She was talking to him	**as if he was**	grown-up and it made him feel unco

This format allows for the formulation of initial hypotheses based on patterns that might be determined in relation to the node. For example, if we look at the co-text to the left of *as if he was*,

the verb *look* appears in lines 1, 2 and 4 and *treat* appears in lines 8 and 9. In terms of patterning to the right of the search item, the *V-ing* form occurs in lines 1, 2, 3 and 7. These patterns can be further explored using corpus software collocational tools. Table 2 provides support for the initial observations based on the concordance lines. In COCA, the most frequent collocate of *as if he was* is *looked*. The COCA results also contain an MI score. The MI (mutual information) score is a statistical measure of the strength of collocation. The higher the MI score generated, the stronger the collocation – an MI score of higher than 3 is usually indicative of a strong collocation (for more information on MI scores see Baker (2006: 100-104)). As we can see, *sounded*, though not the most frequent, is the strongest collocate of *as if he was*.

Table 2: Top five collocates for *as if he was* in the spoken component of COCA (L4-R4)

Number	Word	Frequency	MI
1	looked	70	5.09
2	felt	54	4.96
3	trying	38	4.22
4	looking	28	3.82
5	sounded	25	6.94

There are a number of other reasons why a corpus linguistic methodology is suitable for the study of pragmatics. Pragmatics, in addition to syntax and semantics, emerged in the fifties and sixties as an element in the tripartite approach to the study of language was initially preoccupied with invented or intuited language data. As pragmatics has evolved it has embraced other methods of data collection such as discourse completion tasks in order to elicit, for example, data on speech acts that are relatively rare in everyday conversation. Corpus linguistics offers another data source to researchers studying pragmatic phenomena – naturally-occurring, spontaneous, uncensored, real-life data increasingly freely available to researchers. In addition, context is of great importance to pragmatics in that elements such as the language user(s), social situation and activity type all play an important role in how communication unfolds. Some modern-day corpora provide extensive contextual information in the form of a range of sociolinguistic metadata. For example, written corpora frequently contain information about text type and date of publication which provides opportunities for the study of diachronic historical pragmatics. However, it is modern spoken corpora that provide the richest vein of social variables. These corpora are increasingly characterised by their attention to contextual metadata – many spoken corpora contain speaker information such as place of birth, age, gender, level of education, religion or social class.

Pragmatics also offers reciprocal benefits to corpus linguistics. One of the most valuable of these has emerged, again ironically, from a criticism of the use of a corpus linguistic methodology in the study of pragmatic phenomena – namely the relationship between linguistic form and function. When approaching a corpus, a researcher's first point of entry is often a lexical 'hook' (Rühlemann 2010) such as, for example, the pragmatic marker *you know* or *sorry* as a representation of the pragmatic act of apology. This enables the researcher to access their corpus via the tools outlined here. However, the relationship between linguistic form and function is characterised by ambiguity and unpredictability (Mey 2001) and it is the unfolding of dynamic contextual elements that facilitates the correct understanding of the function a particular form fulfils. This ambiguous, unpredictable relationship hampers the process of automatic retrievablity through corpus software. Corpus linguistics has responded to this challenge by developing methodologies and pragmatic annotation schemes that allow for a more sophisticated utterance-by-utterance analysis (see, for example, Archer and Culpeper 2003; Kohnen 2008; Rühlemann and O'Donnell 2012). An apposite example of this is the Irish component of the International

Corpus of English (ICE-Ireland; Kallen and Kirk 2008) which has been annotated to display aspects of discourse, prosody and pragmatics (SPICE-Ireland; Kallen and Kirk 2012). An example of this system is shown here:

<P1A-033$C> <#> <rep> <[> Exactly exactly </[> </{> </rep> <#> <rep> Oh I-know* I-mean* like* <,> the way I should say to them you-know* <,> at the end of the day you have to respect your privacy <{1> <[1> </rep>
(Kallen and Kirk 2012: 43 [original emphasis removed])

Although on the surface the tags give the extract quite a dense visual appearance, it demonstrates many of the features that makes the corpus searchable and, therefore, of import to the study of pragmatics. For example, <rep>…</rep> marks the beginning and end of a representative act (the corpus is also tagged for directives, commissives, expressives and declaratives) and * is used to tag discourse/pragmatic markers. Our attention now turns to a specific consideration of corpus pragmatics in light of this discussion of mutual benefits of the intersection of the two disciplines.

Corpus pragmatics is a relatively recent development at the intersection of the fields of corpus linguistics and pragmatics and a growing number of corpus pragmatic studies have recently emerged that highlight the fruitfulness of the synergy (Romero-Trillo 2008; Aijmer and Rühlemann 2015). Corpus pragmatics represents a highly iterative approach to the study of pragmatic concepts that integrates the more traditional qualitative or 'horizontal' approach in pragmatics with the more quantitative or 'vertical' nature of corpus linguistics. This combined methodological approach is represented diagrammatically in Figure 1:
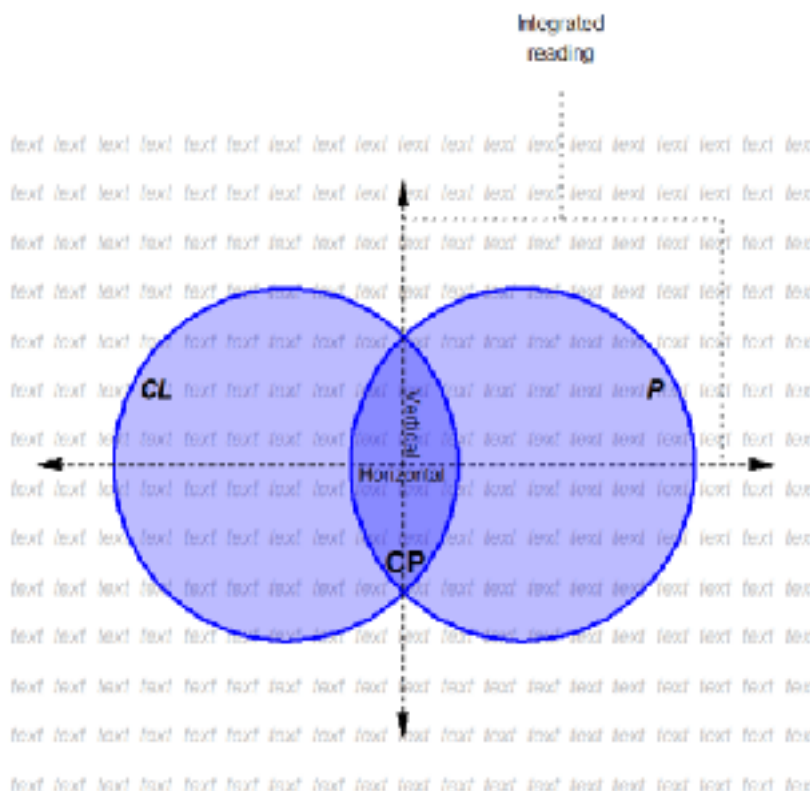

Figure 1: Integrated reading methodology in corpus pragmatics (adapted from Rühlemann and Aijmer 2015: 12)

In this way, corpus pragmatics takes a more nuanced approach to the consideration of lexico-grammatical patterns that characterize a text, something which has long been the focus of both pragmatics and corpus linguistics in general (Clancy and O'Keeffe 2015). Corpus pragmatic

studies are, then, characterized not only by a focus on form and function, but also on patterns of variation at social, cultural and regional levels. The studies also offer new insights into pragmatic principles such as speech acts and (im)politeness through re-evaluation and re-investigation. In order to illustrate corpus pragmatics, we have selected a brief number of studies from the canon to represent what is fast becoming a substantial body of research.

The focus on interpreting patterns of form and function according to their interactional and situational context of occurrence is most evident in one of the largest bodies of work in the corpus pragmatic realm – that of pragmatic markers (PMs). For example, Fung and Carter (2007) explore PM form and function in two pedagogical corpora – a learner corpus from Hong Kong and the pedagogical subcorpus from the Cambridge and Nottingham Corpus of Discourse in English (CANCODE). Aijmer (2013) uses the suite of ICE corpora in order to compare form, function and frequency across a range of World Englishes. Amador-Moreno and McCafferty (2015) examine pragmatic markers diachronically using a corpus of letters written 1750-1940. Millar (2015) focusses on those used on an Irish beauty website. Finally, Clancy (2016) compares pragmatic markers used in the discourse of family and friends. These studies highlight the array of factors that corpus pragmatics pays attention to in its consideration of pragmatic markers. A large variety of corpora have been used to examine pragmatic marker frequency, form and function across different language varieties, time periods, text types, activity types, speaker roles and relationships, etc. Corpus pragmatics has also challenged the membership of that canonical grouping of PMs (*well, you know, like, actually, just*, etc.) by arguing for the inclusion of features such as tag questions (Barron 2015), vocatives (Clancy 2015) or interjections (Norrick 2015). While the field of PMs is arguably the most extensive field of corpus pragmatic research, corpora have also been exploited in the quest for a better understanding of pragmatic phenomena far beyond pragmatic markers. In fact, a growing number of corpus-pragmatic studies are concerned with core-pragmatic concerns such as reference, deixis, speech acts, and turn taking, to name only a few. For illustration, a few such studies are cited in the following. Biber et al. (1999: 263) propose a variational-pragmatic case study on reference patterns with the definite article *the* across the four registers, conversation, fiction, news reportage and academic writing. Among the many discoveries is that "[a]lthough anaphoric reference may intuitively seem to be the most basic use of the definite article, other uses are in fact equally or more common" (Biber et al. 1999: 266); for example, while anaphoric reference accounts for less than 30% of all uses in all four registers, situational reference, which "ranges from reliance on the immediate speech situation to dependence upon the larger shared context" (Biber et al. 1999: 264), stands out in conversation where it accounts for more than half of all occurrences. Rühlemann and O'Donnell (2015) investigate 'introductory *this*' in storytelling concluding that the usage serves as a theme marker announcing the protagonist of an upcoming story early on and can hance be seen as a form of discourse deixis. Tao (2003) examines 'turn initiators', that is, the first verbal element occurring in a speaking turn, and finds that "the function of turn beginnings may be characterized as mainly to link back to prior turns" (Tao 2003: 203; cf. Heritage 2015 on *well*-prefaced turns). Rühlemann and Gries (2015) study turn order patterns in multi-party storytelling; their findings suggest that turn order in storytellings with more than two participants is essentially structured as if there were only two participants, adding support to recent proposals that see multi-party conversation as "built for two" (Stivers 2015). Speech acts have seen a large number of corpus-based examinations, both in a synchronic and a diachronic perspective. Synchronic corpus studies include, amongst others, Aijmer (1996) on thanking, apologies, requests, and offers; Adolphs (2008) on the distinction between questions (i.e. requests for information) and suggestions with 'why don't you'; Jucker et al. (2008) on compliments; Allister (2015) on directives in academic contexts. Diachronic speech act studies based on corpora include, inter alia, Jucker and Taavitsainen (2000) on insults; Kohnen (2008) on directives, Jucker et al. (2008) on compliments, Jucker and Taavitsainen (2008) on apologies (for a comprehensive overview of diachronic corpus research into speech acts, see Kohnen 2015).

Corpus pragmatic studies offer a new perspective on traditional or classical pragmatic (and indeed corpus linguistic) phenomena. Corpus pragmatics has contributed to a re-evaluation of the use of elicited data in pragmatics through a comparison with corpus data. Schauer and Adolphs (2006) compared data in relation to the speech act of gratitude generated from both DCTs and the CANCODE corpus. At the level of form, they found that in both the DCT and corpus data, *thank**

and *cheers* were the most frequent expressions of gratitude. However, at an interactional level, where 'speech acts combine into larger units of discourse' (Schneider 2012: 1027), the results converged. The corpus data showed that *cheers* primarily functioned not as an expression of gratitude, but as a *response* to such an expression. In addition, the corpus data highlighted that the speech act of gratitude takes place across an extended series of speaker turns as part of a process of collaborative negotiation. This attention to spoken language as an interactional, collaborative process is a feature of other corpus pragmatic research. For example, Clancy and McCarthy (2015) examine the phenomenon of co-construction across two corpora – CANCODE and the Limerick Corpus of Irish English (LCIE). Co-construction is an inherently pragmatic activity in that it involves the creation of meaning in and through context-specific interaction (see Kereckes, 2007; Rühlemann, 2007). A corpus pragmatic approach, in addition to confirming again that meaning is a dynamic and emergent phenomenon, allowed Clancy and McCarthy to quantitatively demonstrate the frequency of co-construction in relation to the items *if*, *when* and *which* and also exemplify extended patterns by which interactants sanction one another's participation in this phenomenon.

In the following we will present a case study on the choice between indicative *was* and subjunctive *were*. Our aims are two-fold: first we wish to illustrate the integrated corpus-pragmatic methodology; second we propose an interpretation of our findings in the light of Lyons' (1977) notion of 'empathetic deixis', thus contributing to the small, yet growing body of corpus reseach on deixis, undoubtedly one of the key areas of pragmatic interest.

2. Case study: indicative *was* and subjunctive *were* in *as-if* clauses — a case of empathetic deixis?

2.1 Introduction

The distinction between indicative *was* and subjunctive *were* expresses a contrast of mood. Mood "refers to the factual or non-factual status of events" (Carter and McCarthy 2006: 307). Mood, as well as the distinction between indicative *was* and subjunctive *were,* therefore needs to be appreciated within the broader perspective of modality. Modality "construes the region of uncertainty that lies between 'yes' and 'no'" (Halliday and Matthiessen 2004: 147). Its primary function is that of assessing "the intermediate degrees, various kinds of indeterminacy" (Halliday and Matthiessen 2004: 146) that fall in between polar choices such as 'yes' and 'no'. More specifically, 'uncertainty' is an epistemic attitude related to a range of intermediate values on the continuum spanning between extreme values of 'factual' and 'true' on the one hand and 'non-factual' and 'untrue' on the other. It is obviously not the case that states or events *are* factual, near-factual, probable, possible, hypothetical, unlikely, or non-factual in and by themselves. They are assigned values on the factual/non-factual cline by speakers. The expression of (un-)certainty is thus a deeply pragmatic category: the degree to which information is marked as certain and hence reliable or less certain and hence less reliable depends on the speaker and the degree of his/her knowledge or belief that what he/she says, reflects the true state of affairs.

Regarding the distinction between the indicative mood and the subjunctive mood, the former is often considered "a factual mood" (Carter and McCarthy 2006: 3007), whereas the latter is treated as the non-factual one. As Mindt (2000: 165) observes with regard to subjunctive *were*: "[n]on-real states are generally expressed by the subjunctive form *were*." These observations, however, are merely preliminary orientations. The assumption of a convenient division of labor between indicative *was* used for expressing factuality and subjunctive *were* used to express lack of factuality is simplistic and not borne out by authentic speech data. Even a quick search in corpus data instantly returns counter examples where *was* is used to express counterfactuality as well as examples where *were* is used in the context of relative certainty. In (1), an excerpt from an investigation into the murder of Angie Samota, Angie's friend Anita who spent a night out with Angie and Russell, is reported as *feel[ing] as if she were along for appearance's sake only*. Here, it appears that Anita was under the firm impression that the sole purpose of her being taken along by Angie and Russell was *for appearance's sake*. For Anita, then, the state of affairs described in the *as if* clause was not only possible but even likely. In (2), by contrast, the arrest of people planning

to blow up JFK airport is certainly not a trivial event; despite this certainty, the speaker uses indicative *was* in the *as if* clause.

(1)   Now, 28 years after Angie Samota was murdered, Russell is talking once again about what happened that night and about Angie. (Photo-of-Samota; MANKIEWICZ: Her friends describe her as the kind of girl that guys get crushes on. Mr-RUSSELL-BUCHANA: Maybe so. MANKIEWICZ: Possible that you had a crush on her? Mr-BUCHANAN: Oh, no, no. Not at all. I hardly knew her. MANKIEWICZ: (Voiceover) But after questioning Angie's friend Anita about that shared night out, investigators wrote that she told them that the evening centered around Russell and Angie and that Anita felt **as if she were along for appearance's sake only**. (NBC_Dateline, 2012)

(2)   When the FBI announced the arrest of people who wanted to blow up JFK Airport, New York Times buried that story on page A-37 **as if it was**, you know, trivial. (NPR_TalkNation, 2007)

Rather than assigning the two forms to the opposite ends of the factuality/non-factuality continuum, Quirk et al. (1985: 1110) observe that subjunctive *were* and indicative past may be used as 'alternatives' in hypothetical *as if* clauses. The examples given by Quirk et al. include:

(3)   She treats me as if I was a stranger.
(4)   She treats me as if I were a stranger.

The choice between indicative past and subjunctive *were* may hence not be a grammatical one; that is, it may not be one that is coerced on the speaker by the grammatical system that reserves one form for one type of modal meaning and another form for another modal meaning. Rather, the two forms seem to be competing with one another in the same meaning area of uncertainty referred to above.

   The question we are going to address in this case study is related to the factors co-determining the choice of the one form over the other. While the *was/were* alternation occurs in a variety of constructions, particularly in hypothetical *if*-clauses (cf. Quirk et al. 1985: 1091ff.), we will be concerned here only with *was/were* alternation in *as-if* clauses.

   We base our examination on the hypothesis that the choice of the subjunctive form is a case of negative empathetic deixis. While deixis is in general "one of the most empirically understudied core areas of pragmatics" (Levinson 2004: 97), empathetic deixis is even less well-studied; only very few relevant studies come to mind, e.g., Rühlemann's (2007) case study on 'introductory *this*' in storytelling. Empathetic deixis occurs "when the speaker is personally involved with the entity, situation or place to which he is referring or is identifying himself with the attitude or viewpoint of the addressee" (Lyons 1977: 677). Related terms include 'inner deixis' (Caffi & Janney 1994) and 'emotional deixis' (Lakoff 1974). Lakoff subsumes under emotional deixis uses of *this* and *that* that are "generally linked to the speaker's emotional involvement in the subject-matter of his utterance" (Lakoff 1974), noting that emotional deixis is used for 'vividness' and typically occurs in colloquial contexts.

   Lyons (1977) emphasizes that empathetic deixis typically concerns binary choices, as between *this* and *that, here* and *there, now* and *then,* where, depending on the speaker's involvement (or dis-involvement) with the entities referred to, or identification (or non-identification) with the addressee's attitude or viewpoint, a shift can be observed from origo-farther reference (e.g., *that*) to origo-nearer reference (e.g., *this*), or the other way round, from reference that is closer to the speaker's deictic center to reference that is further removed from it. In the context of the choice between *this* and *that*, it is instructive to note that the near/distant polarity commonly assumed for the pair (with *this* indexing nearness while *that* indexes distance) is best seen as "a matter of *psychological* rather than real distance" (Quirk et al. 1985: 374; our emphasis). It is this psychological potential inherent in the choice that enables speakers to alternate between the binary forms as required not (only) by physical or factual proximity but by their psychological

situation (attitude, epistemic status, level of (un-)certainty, etc.) thereby also manipulating origo-proximity (cf. Rühlemann 2007: 222).

We finally note that empathetic deixis, although hardly studied, seems to be observable in a great many fields. These would include not only the above-mentioned use of 'introductory *this*' but also 'attitudinal *that*', as in *Janet is coming. I hope she doesn't bring that husband of hers* (Quirk et al. 1985: 374), where *that* is used to "imply dislike or disapproval" (Quirk et al. 1985: 374). Also, empathetic deixis may be manifest in the 'subjective progressive', exemplified in *The silly cow. She's always trying to tell me things* (BNC: HGL 3271), a usage which often "suggests a hyperbolic tone of disapproval" (Leech et al. 2009) and where the choice of the progressive aspect over the simple aspect seems to suggest a move from origo-neutral to origo-near territory. Empathetic deixis may also be at play in Historic Present, a switch from past tense to present tense in storytelling, as in *we're driving down the west coast and there was like some rocks* (Narrative Corpus) where the switch into the origo-near present tense in *we're driving* could be seen as indexing the narrator's heightened involvement with the events he/she is relating (cf. Rühlemann 2007: 192); its effect is to "produce a more vivid description" (Biber et al. 1999: 454). Empathetic deixis also may account for occurrences of indirect speech where the decision to backshift or retain the tense of the original utterance is contingent on the speaker's "doubt as to [the reported utterance's] present validity" (Quirk et al. 1985: 1028). Choosing to backshift, where non-backshifting is perfectly legal, then, would indicate the speaker's disinvolvement with the reported proposition via use of origo-far past tense. Finally, we wish to list 'alternative recognitionals' among the possible candidates for empathetic-deictic choices. Alternative recognitionals are person references that occur in categorical shifts, i.e., in "environments where the unmarked category of reference [a simple name] was entirely possible" (Stivers 2007: 77). They most commonly take the form of "a descriptive recognition instead of a name" (Stivers 2007: 77), for example, 'yer sister' instead of 'Alene' (to use the choice described in Stivers 2007). Alternative recognitionals too are psychology-driven as they are "commonly used in complaints" (Stivers 2007: 82) where they serve to place "the referent in the domain of the responsibility of the addressee" (Stivers 2007: 81) rather than the speaker, thus putting the referent at a distance from the latter.

To return to the *was/were* alternation, it should be obvious that the choice between indicative *was* and subjunctive *were* too represents a binary choice. The question arising then is whether it can also be used for empathetic deixis. This seems at least possible if we consider the semantics of *as-if* clauses and how this underlying semantics is exploited pragmatically in context.

There is broad agreement among grammarians that *as-if* clauses typically contain a strong element of similarity and comparison. Quirk et al. (1985: 1110) list them as instances of 'clauses of similarity and comparison'. Biber et al. (1999: 840) note with regard to the subordinators *as if* and *as though* that they "indicate that the adverbial clause is showing similarity but is not to be taken factually." Carter & McCarthy (2006: 774) observe that *as-if* clauses can operate "as the second element of comparisons of similarity". The first element of comparison resides in the matrix clause usually preceding the *as-if* clause. So, to return to Quirk et al.'s (1985) example

(5)   [C1 She treats me] as if [C2 I was a stranger]

the way *she treats me* (as comparative element C1) is compared to the treatment of *a stranger* (as comparative element C2), with the conjunction *as if* conjoining the two elements. Or, to use examples from COCA data:

(6)   [C1 When Grant was a young child, he looked] as if [C2 he was older than he was].
         (CNN_SunMorn, 2002)

(7)   MARTIN-LUTHER-KING# Like anybody, I would like to live a long life. CHRISTIANE
         AMANPOUR (ABC NEWS) (Off-camera) [C1 Martin Luther King, saying that just before he
         died], as if [C2 it was a premonition]. (ABC, 2012)

(8)   [C1 Mitt Romney spent much of the day campaigning] as if [C2 he were the nominee]
         (CBS_NewsMorn, 2011)

(9)   [C1 Clinton was serving up the partisan red meat], almost as if [C2 she were a butcher].
(ABC_Nightline, 2007)

In (6), the way the basketball player Grant Hill looked as a young child is compared to the way older children look; in (7), King's assertion on the day before his assassination that he would like to live a long life like anybody is compared to a premonition; in (8), Mitt Romney's way of campaigning is likened to the campaign style of a party nominee; and, in (9), Hillary Clinton made a series of partisan ('red meat') remarks in a way similar to the way a butcher serves up red meat. While, then, *as-if* clauses invariably serve as the second element of comparisons of similarity, it will be obvious that the actual 'comparability' of the two comparative elements varies a great deal. The variation is in terms of degrees of factuality. In (6), Grant Hill's height as a young child was in actual fact the height of an older child; so the comparison encoded in the *as-if* clause builds on a factual base and may hence be labeled 'factual'. In (7), King's assertion he would like to live a long life just hours before he was shot has indeed the 'aura' of a premonition; however, it is unclear whether there are in fact premonitions and it is unclear whether King's assertion that he would like to live a long life was in fact (intended as) one of them. The comparison thus seems based on a premise that is at best 'near-factual'. A more obvious case of lacking factuality is (8). Here, Mitt Romney's campaigning looked like the party nominee's at a time in 2011 when he had not yet become the offical GOP nominee for president; the comparison hence builds on a premise that is simply incongruent with the facts; it is 'counterfactual'. In (9), finally, the comparison of Hillary Clinton with a butcher is very obviously counterfactual. Indeed, considering the extreme disparity between the concepts of 'politician' and 'butcher' (in a literal sense) the comparison is more than just contrary to facts and best understood as 'absurdly counterfactual'. In other words, semantically *as-if* clauses build on comparison and similarity; pragmatically, speakers have room to exploit the comparability of the compared elements to stake out their stance on the factuality cline ranging from factual to absurdly counterfactual.

Having drawn the outlines of empathetic deixis, the perspective in which we are going to evaluate our results, and having sketched out how semantic and pragmatic aspects interact in *as-if* clauses, in the following sections we aim to report on a case study on factors co-determining the choice between indicative *was* and subjunctive *were* in *as-if* clauses. As will be shown the study is limited in scope and intended strictly as a pilot study. We will describe the data and methods used in the next section (Section 2.2) before we present the results (Section 2.3), discuss the findings (Section 2.4) and, finally, draw some tentative conclusions (Section 2.5).

2.2 Data

This case study is based on data from the COCA corpus (cf. Davies 2008), a very large corpus of contemporary American English. Being a monitor corpus, it is updated regularly; at the time we used COCA for this study, the word total was 520m words. While the corpus hosts a number of written registers including Fiction, Magazine, News, and Academic, its spoken data are limited to transcribed speech from a broad range of unscripted TV programs. Even in that spoken sub-corpus, *as if* clauses seem to be very frequent. A search for *as if * was/were*, where the asterisk * is used as a wild card for any potentially intervening item, e.g. *that, there, he, she, it, they*, etc. returns 1198 hits. This is a large number — too large to allow any in-depth (corpus-)pragmatic analysis. Consequently, analyzing all available occurrences of *as if* was far beyond our aims. To keep the analysis manageable, we decided to exclude *as-if* clauses with nominal subjects and plural pronominal subjects and instead to focus on *as if* followed by singular personal pronouns only; thus, the data set subjected to further analysis included clauses with *as if I was/were*, *as if he was/were*, and *as if she was/were*. This raw data set included 322 hits. Five of them turned out to be duplicates, which were discarded leaving us with 317 hits for further processing.

2.3 Methods

The 317 hits selected for close examination were coded for two factors, one pragmatic, one syntagmatic.

The pragmatic factor is the degree of factuality of the comparison expressed in the *as-if* clause. Four degrees were distinguished on the factuality continuum: 'factual', for clauses expressing circumstances that are taken to be true; 'near-factual', for clauses describing circumstances that are presented as 'close' to being true; 'counterfactual', for circumstances that are clearly at odds with the facts; and, finally, 'absurdly counterfactual', for circumstances extremely (sometimes hilariously) out of touch with the facts. In nearly all cases, it was necessary to consult larger contexts than the concordance line. In many cases, the degree of factuality only became clear from very large extended contexts and required substantial inference. A case in point is example (10):

(10)   Even if you know Mr. Govan's condition in advance, you can be startled to enter a room of
        Chicago's Schwab Rehabilitation Center and see the upper body of a strong young man
        poking out of the center of a padded platform, **as if he were** simply standing up out of the
        sunroof of a sports car. Then you realize, Mr. Govan is standing up on his hips, a stuffed
        pair of blue pants legs from a warm-up suit, socks and unsmudged running shoes are
        propped up on the bottom half of a nearby wheelchair while the technicians twist, test and
        tinker with Howard Lee Govan's new legs (NPR_Weekend, 1995)

Here, the speaker initially describes the impression of 'seeing the upper body of a strong young man poking out of the center of a padded platform, as if he were simply standing up out of the sunroof of a sports car'. The picture of a strong young man standing up out of the sunroof of a sports car is nothing unusual, so the comparison expressed in the *as-if* clause seems merely counterfactual (since the 'padded platform' *is* not a car's 'sunroof') but it does not seem to violate factuality in any outrageous way. However, Mr. Govan's 'condition', it transpires from the end of the extract, is that he apparently has artificial legs that can be removed from the body. So, the comparison of Mr. Govan who is in actual fact standing up on his hips while technician are tinkering with his new legs with a strong young man standing up out from a car's sunroof is wildly disparate. The hit was hence coded 'absurdly counterfactual'.

Given examples such as these, it will not be surprising that in a number of cases the degree of factuality could not be determined with sufficient confidence. These cases were excluded from further examination.

In pragmatic studies, the focus of attention is often on processes in the speaker's mind, such as inference, intention, evaluation, (un-)certainty, and, in our case, factuality — processes which are often not encoded in the language and thus remain in the unsaid or, at best, leave behind traces, or indices, in the said. In corpus linguistic studies, by contrast, the focus shifts to processes directly observable in the speaker's language, that is, in the text consisting of the string of words actually produced — processes which are then manifested in the said. These textual processes can be described as syntagmatic processes governed largely by what Sinclair (1991) termed the 'idiom principle'. This principle holds that the choice of any one word in co-text raises the odds that certain other words will be co-selected in that co-text. Corpus linguistics has been prolific in discovering fundamental types of syntagmatic patterning, such as collocation, colligation, semantic prosody, and so forth. In this spirit, corpus linguists have targeted binary choices such as the dative alternation (e.g., Bernaisch et al. 2014) or the genitive alternation (e.g., Gries & Wulff 2013). The analyses of these choices have invariably suggested that none of them can be explained by a single factor alone but that the choices owe to multiple factors, many of them being syntagmatic factors. We take up this multi-factorial perspective in our analysis of the *was/were* alternation in *as-if* clauses. This uptake is motivated by collocational analyses. These brought to light that copular verbs were by far the most frequent group of collocates of *as-if* clauses. Since the copular verbs invariably demand subject complements for their complementation we hypothesized that the syntactic function the *as-if* clause enters into is another co-determining factor in the *was/ were* alternation. We already note at this stage that if we discover a significant attraction of either

or both the alternative forms to one or more syntactic functions it will be interesting to see whether such an attraction can be understood as (a sub-type of) colligation, an association pattern of a lexical item in/with a grammatical category (cf. Hoey 2005).[1]

Based on the data set of 317 hits and with all codes for Factuality and Syntactic Function in place, we carry out a linear probability model, a type of logistic regression model, to test the following hypotheses:

H_0: The choice of indicative 'was' and, respectively, subjunctive 'were' in *as-if* clauses does not depend on any of the independent variables Factuality and Syntactic Function and their pairwise interactions; adjusted $R^2 = 0$.

H_1: The choice of indicative 'was' and, respectively, subjunctive 'were' in *as-if* clauses does depend on at least one of the independent variables Factuality and Syntactic Function and their pairwise interactions; adjusted $R^2 > 0$.

(R^2 is a measure which "quantifies the proportion of the variance in the data that is captured and explained by the regression model" (Baayen 2008: 88) and which can thus serve as a diagnostic for the goodness-of-fit of the model (cf. also Gries 2009: 260).)

2.4 Results

As noted, following standard corpus linguistic procedure, we started out by analyzing collocational profiles of *as-if* clauses. Table 3 lists the top ten collocates in L3-R3 (i.e., three items to the left of the node, three items to the right):

Table 3: Collocates of as if i|he|she was|were in the spoken subcorpus of COCA (L3-R3)

Rank	Word	Freq	MI score
1	almost	24	5.40
2	felt	16	5.48
3	looked	11	5.10
4	trying	9	3.16
5	sounded	8	7.90
6	seem	6	4.38
7	treat	5	5.98
8	treated	5	5.76
9	seems	5	4.62
10	reading	4	4.72

The first observation to be made is that nine out of the top ten most frequent collocates are verbs. The second observation is that *felt, looked, sounded, seem*, and *seems* are among typical

---

[1] Colligation is an interesting 'idiomatic' patterning in that it involves a lexical item (or several such items) in frequent co-occurrence with an abstract, non-lexical paradigmatic 'slot' (e.g., the subject, the theme, etc.). In other words, colligation is not a purely syntagmatic phenomenon but rather an intertwining of the syntagmatic and the paradigmatic axes.

copular verb forms, an observation suggesting the possibility that the *as-if* clause acts as a subject complement constituent. To examine this hypothesis, we did another collocational search, this time for verbs only and for the left-hand range L1-L3 only. This latter restriction is justifiable considering that if the *as-if* clause acts as a subject complement then it will have to appear after ('on the right' of) the copular verb. The results of this search (not shown here) fully confirmed the hypothesis: the number of typical copular verbs increased. Finally, to ascertain that the verbs were indeed used as copular verbs and not otherwise (e.g., as mono-transitive verbs, which is possible in the case of *felt, looked, sounded*) we inspected concordance lines. Again, the hypothesis was confirmed: in the large majoritiy of cases the verbs in question did act as copular verbs and, hence, the *as-if* clause did act as subject constituent. Given this apparent co-occurrence of copular verbs and *as-if* clauses as subject complement constituent, we decided to code all hits in the sample for the syntactic function the *as-if* clause fulfils in the matrix sentence.

It was found that *as-if* clauses essentially perform two distinct syntactic functions, as a subject complement and as a manner adverbial. The two functions are illustrated in (11)-(16):

Subject complement:
(11)   … They looked **as if it was** falling on deaf ears, Dr. Drew. (CNN, 2013)

(12)   … It was **as if I were** having a conversation with a person that I both trusted and loved (CNN_Crossfire, 1990)

(13)   … I felt **as if he was** saying goodbye to me … (CBS_Rather, 2004)

Manner adverbial:
(14)   … some of the workers touched the steel beam **as if it was** a coffin. (NBC_Today, 2002)

(15)   refused to bring them back, lied about it and then began posing **as if she were** the mother. (CNN_Grace, 2011)

(16)   … So the President was behaving **as if he was** the underdog. (ABC, 2012)

In (14) through (16), the *as-if* clause does qualify as a manner adverbial in that, as is characteristic of this type of adverbial, it "can be paraphrased by *in a … manner* or *in a … way*" (Quirk et al. 1985: 557). So, for example, (16) can be paraphrased as *So the President was behaving in a manner/way similar to an underdog*.

Beside these two major syntactic categories we also detected a few minor ones. These include uses of the *as-if* clause as an object complement, a metapragmatic comment, a kind of inversed concessive adverbial,[2] as a metalinguistic comment, and as part of a discontinuous modification of an adjective phrase:

Object complement:
(17)   he wasn't treating him as if he was gay and insulting him **as if he was** gay (MSNBC_Carlson, 2006)

Concessive adverbial:
(18)   Gene Randall, CNN, Chicago. CATHERINE CRIER, Anchor: Well, **as if he wasn't** getting enough flak from the Bush camp, Clinton got some lip (CNN_Politics, 1992)

Meta-pragmatic comment:
(19)   … the statement where he said he chose to appear - appear, **as if it was** an appearance. (CNN_Showbiz, 2009)

---

[2] We refer to this concessive adverbial as 'inversed' because the negative polarity expressed in the *as-if* clause is inversed, via inference, into a positive polarity. So, for example, *as if he wasn't getting enough flak from the Bush camp* translates into *although he was getting enough flak from the Bush camp*.

Discontinuous modification of adjective phrase (as-as construction):
(20)   It still is as real **as if it was** yesterday, and... (CBS_48Hours, 2008)[3]

As can be seen from Table 4, *as-if* clauses performing these minor syntactic functions were far less frequent than the two major types manner adverbial and subject complement. The hits coded for minor syntactic functions were therefore eliminated from further analysis.

Table 4:   Break-up of occurrences of syntactic functions of *as-if* clauses with 'was' and, respectively, 'were'

Syntactic function	*was*	%	*were*	%
Subject Complement	84	51.22	42	27.45
Manner Adverbial	54	32.93	93	60.78
Object Complement	8	4.88	8	5.23
Concessive Adverbial	2	1.22	0	0.00
Metalinguistic Comment	1	0.61	1	0.65
Discontinuous Modification	0	0.00	1	0.65
Unclear	15	9.15	8	5.23
Total	164		153	

The second variable for which we coded the data was factuality: whether the comparison expressed in the *as-if* clause is factual, near-factual, counterfactual, or absurdly counterfactual. A number of cases could not be determined with sufficient confidence. The hits were excluded from further analysis. Table 5 shows how the hits are distributed across the factuality continuum:

Table 5:   Break-up of degrees of factuality of *as-if* clauses with 'was' and, respectively, 'were'

Factuality	*was*	%	*were*	%
Factual	28	17.07	3	1.96
near-factual	40	24.39	21	13.73
counterfactual	59	35.98	78	50.98
absurdly counterfactual	15	9.15	43	28.10
unclear	22	13.41	8	5.23
Total	164		153	

---

[3] In data not included in the analysis (because the subject of the *as-if* clause is 'it') we found a case where the *as-if* clause seems to function as a reason adverbial:

But I felt that as if it was part of heritage and part of tradition, we should keep it. (CNN-News, 1996)

Here, *as if* could be replaced by the causal conjunction 'because' without forcing any change in meaning.

We see that the bulk of the hits for 'was' and 'were' alike are classed as 'counterfactual', but we also discover that the alternative forms are quite differently weighted across the factuality cline. This becomes obvious when the scores for the extreme values 'factual' and 'absurdly counterfactual' are considered: while the precentage of *as-if* clauses with 'was' coded as 'factual' is c. 17%, only 1.96% of *as-if* clauses with 'were' are coded 'factual'. The inverse relation holds for 'absurdly counterfactual': only 9.15% for 'was' but 28.10% for 'were'.

Using this data we defined a linear probability model to determine the extent to which the two factors Factuality and Syntactic Function and their potential interaction bear on the choice between indicative 'was' and subjunctive 'were'. As is standard practice in regression modeling, we began with the maximal model including all factors and their pairwise interaction, then eliminated insignificant predictors step by step until we arrived at the minimal adequate model which contains only significant predictors.

It turned out that no interaction between Factuality and Syntactic Function survived the elimination process. The minimal adequate model resulted in a very highly significant overall correlation (adjusted R-squared 0.1887, F-statistic: 15.77 on 4 and 250 DF, p-value: 1.602e-11). Its coeffcients are given in Table 6:

Table 6:   Coefficients of the linear probability model

	Estimate$_{was}$	Estimate$_{were}$	Std. Error	t value	Pr(>ItI)
(Intercept)	0.93493	0.06507	0.08159	11.459	< 2e-16 ***
SyntacticFunction_mannerA	-0.19659	0.26166	0.06094	-3.226	0.00142 **
Factuality_near-factual	-0.20921	0.27428	0.10285	-2.034	0.04299 *
Factuality_counterfactual	-0.40209	0.46716	0.09530	-4.219	3.43e-05 ***
Factuality_absurdly counterfactual	-0.56200	0.62707	0.10964	-5.126	5.93e-07 ***

Three aspects are key in interpreting the model's coefficients. For each predictor the coefficients are computed for a reference category; in the case of Syntactic Function, the reference category is 'subject complement'; for Factuality, the reference category is 'factual'. Second, the coefficients indicating the probability levels are the values listed under 'Estimate'. Third, positive 'Estimate' values indicate increased probability, negative values indicate decreased probability.

Thus, for the predictor Syntactic Function, where the 'Estimate' coefficient for manner adverbial is -0.19659, we see that the form 'was' is roughly 20% less probable in *as-if* clauses functioning as manner adverbials than in *as-if* clauses functioning as subject complements. By contrast, 'were' is roughly 26% more probable in manner-adverbial *as-if* clauses than in subject-complement *as-if* clauses.

Further, regarding the predictor Factuality we find perfectly complementary probabilities for 'was' and 'were'. The probabilities for 'was' to occur across the factuality continuum continuously *decrease*: compared to factual *as-if* clauses, the indicative form is c. 20% less likely to occur in near-factual ones, c. 40% less likely in counterfactual ones, and finally c. 56% less likely in absurdly counterfactual ones. In other words: the less factual the *as-if* clause, the less probable it is that a speaker will choose indicative 'was'. For 'were', conversely, the probabilities continuously *increase* across factuality: again taking the factual level as the baseline, the subjunctive form is c. 27% more likely to occur in near-factual clauses than in factual ones, 47% more likely in counterfactual ones, and 63% more probable in absurdly counterfactual ones. In other words: the less factual the *as-if* clause, the more likely speakers will choose 'were'.

2.5 Discussion

We investigated the *was/were* alternation in *as-if* clauses in COCA using a linear probability model. The model was built on the hypothesis that the alternation is influenced by two factors: Factuality and Syntactic Function. The model confirmed the hypothesis: both factors significantly predict the choices. The adjusted R-squared value of 0.1887 though was low, suggesting that the model is far from perfect: too much variance is still left unexplained. This underscores the pilot character of the present study. As noted, we did not include in the model hits for which the codes could not be assigned confidently. We also excluded from the model some minor syntactic functions. The model could possibly be improved if these variables are taken into account. However, this will only make sense if a larger sample is available where even very small categories are represented in sufficient numbers. A larger sample could also potentially highlight other factors driving the choice that have so far remained hidden. For all its limitations though, the present model is a reasonable starting point. Its main strengths are the following.

The results indicate that the choice of indicative 'was' and subjunctive 'were' is a multifactorial choice. We have identified two contributing factors. The first factor is the syntagmatic association between 'was' and the subject complement function of the hosting *as-if* clause on the one hand and the association between 'were' and the manner adverbial function on the other hand. This is no doubt a complex association. If it qualifies as colligation, it will certainly not represent a typical one. Typical colligations are frequent co-occurrences of a word/phrase and/in a grammatical function. In the case of the *was/were* alternation, the verbs 'was' and 'were' are embedded in the *as-if* clause, and consequently, the association does not directly hold between the verb and the syntactic function as either subject complement or manner adverbial but holds only indirectly via intermediation by the clause as the host structure. So rather than two elements in frequent co-occurrence we have three elements (the verb form, the clause, and the syntactic function) and, to compound matters, the three elements are not all discrete elements but the first two, verb and clause, are stacked into one another, with the verb being an element within the clause. Despite these intricacies, it seems clear that the association is a syntagmatic one ultimately obeying the idiom principle: if the *as-if* clause is selected in subject complement function, it is likely that the indicative form is co-selected within the clause; if the *as-if* clause is selected in manner adverbial function, it is likely that the subjunctive form is co-selected within it.

Second, the results suggest that the *was/were* choice is co-determined by the factuality of the comparison expressed in the *as-if* clause. Factuality, we noted, is a graded phenomenon, with some comparisons compatible with fact, some just close to but distinct from fact, others clearly distinct from fact, and again others utterly at odds with fact. The data showed that both verb forms are, in principle, possible with all four degrees of factuality. However, the model very strongly suggested that indicative 'was' and subjunctive 'were' differ with regard to the probability of co-occurence with each of the four levels of factuality: 'was' was predicted to co-occur with much greater probability with the levels closer to the factual end whereas 'were' was predicted to co-occur with greater probability with levels closer to the 'absurdly counterfactual' end. The different 'weights' of the two verbs on the factuality continuum are shown in Figure 2.

The histogram in the left panel shows the density curve for 'was' across the four degrees of factuality; the curve peaks right between the near-factual and the counterfactual levels. The middle panel represents the density curve for 'were' across the four levels; here, the peak is farther toward the opposite end of the continuum, at the counterfactual level. In the right panel the two density curves are overplotted to make them more easily comparable. The panel also shows the means (depicted in the dotted lines), which represent the central tendencies in the distributions: clearly, the two curves do not overlap and the central tendencies are differently located vis-à-vis the continuum, with the curve and the mean for 'was' leaning more toward the factual end of the continuum and the curve and the mean for 'were' moving farther away from the factual end.

So, the indicative form has a tendency to occur in *as-if* clauses closer to fact, whereas the subjunctive form tends to occur in *as-if* clauses farther removed from fact. Is this evidence to argue that the *was/were* choice is a case of empathetic deixis? It appears that the case can be tentatively
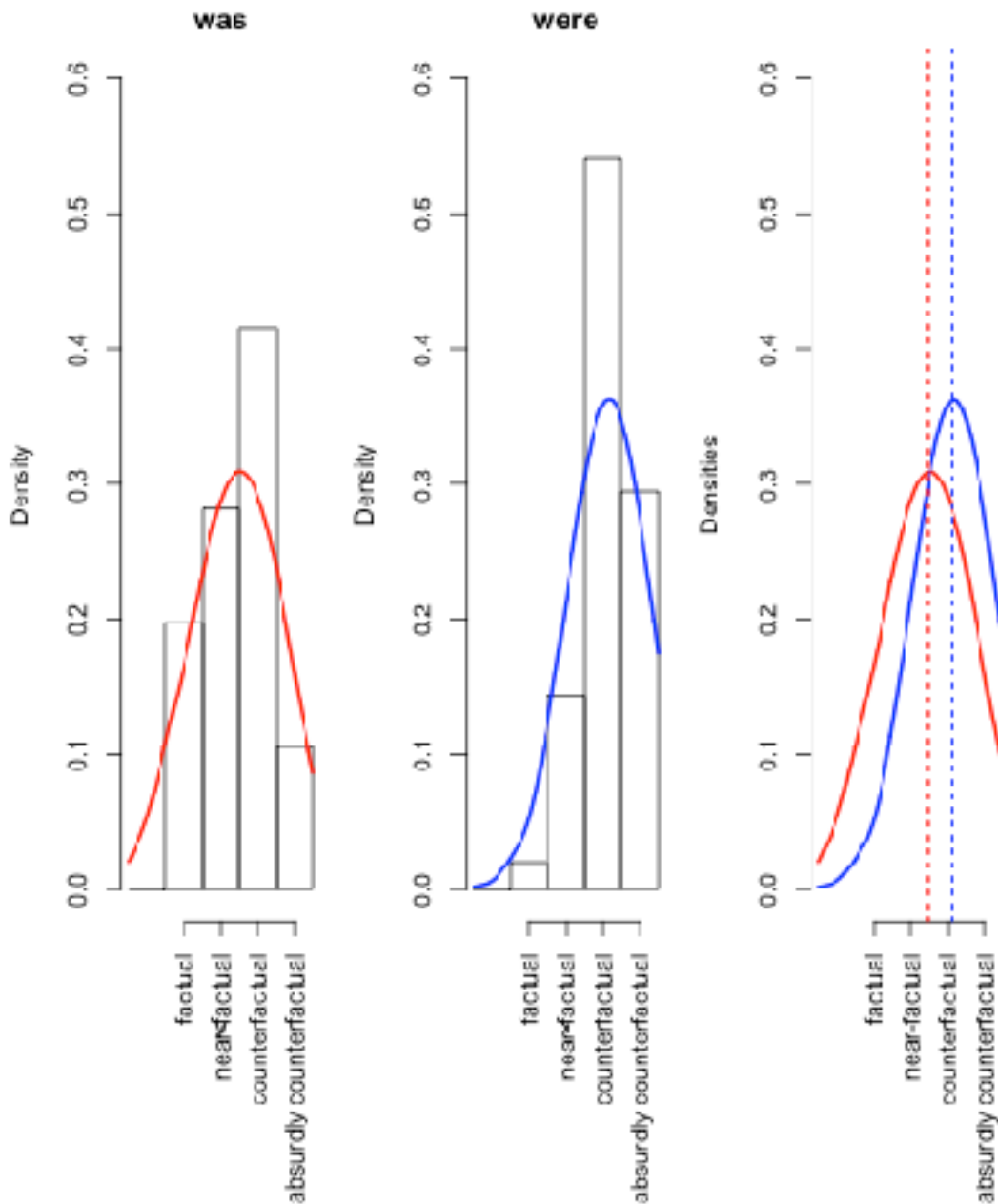
Figure 2:   Factuality clines for *as-if* clauses with 'was' and 'were' compared; left panel: histogram with density curve for 'was'; middle panel: histogram with density curve fore 'were'; right panel: density curves are overplotted, also shown are the means

made. Key to empathetic deixis is the notion of the deictic center. A typical empathetic choice is one where the speaker manipulates his/her proximity to that center, by indexing a reference as located farther away from, or closer toward, the center. In the case of the *was/were* alternation in *as-if* clauses it seems that this proximity variation is at work too: the use of 'was' indexes closer proximity to the factualness of the comparison in the *as-if* clause, whereas selecting 'were' indexes reduced proximity. In other words, while 'was' is an indexical 'pointing to' the speaker's relative lack of commitment, 'were' points out the speaker's utter non-commitment to the factualness of the comparison. It may thus represent a case of negative empathetic deixis.

Conclusions

This case study demonstrates the benefits of employing to the study of pragmatic phenomena a corpus pragmatic approach that integrates core methodologies of either discipline: the horizontal reading methodology typical of pragmatic analysis and the vertical reading methodology typical of corpus-linguistic analysis. Motivated by the observation that speakers routinely utilize both indicative 'was' and subjunctive 'were' in *as-if* clauses, applying this integrated methodology to corpus data yielded interesting insights. In particular, the findings presented seem to suggest that contrary to mainstream thinking in corpus linguistics, which prioritizes text-internal processes in the said over speaker-internal processes in the unsaid, and contrary to mainstream pragmatics doctrine, which prioritizes speaker-internal processes over text-based processes, the two worlds actually converge. In other words: linguistic choices may be due, not *either* to idiomatic patterning *or* speaker meaning, but to both at the same time. This convergence of influences is what the *was/ were* alternation in *as-if* clauses suggests.

Obviously, our study represents a tentative first step and will certainly require further research in order to interrogate the results presented in more detail. In prioritizing depth over breadth, as the nature of the research question required, we have analyzed a limited dataset, limited in terms of the size of sample collected (small), the type of speech examined (unscripted TV shows) and the variety targeted (AmE). Further research might compare larger samples and focus on private, spontaneous spoken data in other corpora of English such as the *International Corpus of English* suite or the *British National Corpus*. However, while we recognize and acknowledge these caveats, we argue for the corpus pragmatic approach. Corpus pragmatics, when harnessed with rigour, has the potential to offer new insights into both pragmatic and syntagmatic processes and the way the two come together in actual linguistic choice.

References

Adolphs, S., 2008. *Corpus and Context*. Amsterdam: John Benjamins.

Aijmer, K., 1996. *Conversational Routines in English: Convention and creativity*. London: Longman.

Aijmer, K., 2013. *Understanding Pragmatic Markers: A variational-pragmatic approach*. Edinburgh: Edinburgh University Press.

Aijmer, K. and C. Rühlemann. 2015. *Corpus pragmatics. A handbook.* Cambridge: Cambridge University Press.

Allister, P. G. 2015. Speech acts: a synchronic perspective. In: K. Aijmer and C. Rühlemann (eds.), *Corpus Pragmatics: A handbook*. Cambridge: Cambridge University Press, 29-51.

Amador-Moreno, C. and K. McCafferty, 2015. 'Sure this is a great country for drink and rowing at elections': Discourse markers in the *Corpus of Irish English Correspondence*, 1750-1940. In: C. Amador-Moreno, K. McCafferty and E. Vaughan (eds.), *Pragmatic Markers in Irish English*. Amsterdam: John Benjamins, 270-291.

Archer, D. and J. Culpeper, 2003. Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics. In: A. Wilson, P. Rayson and A.M. McEnery (eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt/Main: Peter Lang, 37-58.

Baayen, H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press

Baker, P., 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Barron, A., 2015. 'And your wedding is the twenty-second <.> of June is it?': Tag questions in Irish English. In: C. Amador-Moreno, K. McCafferty and E. Vaughan (eds.), *Pragmatic Markers in Irish English*. Amsterdam: John Benjamins, 203-228.

Bernaisch, T., S. Gries, and J. Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide*, 35(1), 7-31.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan, 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Caffi, C. and R. W. Janney. 1994. 'Toward a pragmatics of emotive communication.' *Journal of Pragmatics* 22, 325-373.

Carter, R. and M. McCarthy, 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Clancy, B., 2015. *Hurry up baby son all the boys is finished their breakfast*: Examining the use of vocatives as pragmatic markers in Irish Traveller and settled family discourse. In: C. Amador Moreno, K. McCafferty and E. Vaughan (eds.), *Pragmatic Markers in Irish English*. Amsterdam: John Benjamins, 229-247.

Clancy, B., 2016. *Investigating Intimate Discourse: Exploring the spoken interaction of families, couples and friends*. London: Routledge.

Clancy, B. and M. McCarthy, 2015. Co-constructed turn-taking. In: K. Aijmer and C. Rühlemann (eds.), 2015. *Corpus Pragmatics: A handbook*. Cambridge: Cambridge University Press, 430-453.

Clancy, B. and A. O'Keeffe, 2015. Pragmatics. In: D. Biber and R. Reppen (eds.), *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press, 235-251.

Davies, B., 2007. 'Grice's cooperative principle: Meaning and rationality.' *Journal of Pragmatics*, 39(12), 2308-2331.

Davies, M., 2008. The Corpus of Contemporary American English: 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

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

Francis, N.W. and H. Kučera, 1964. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University, Providence, RI.

Fung, L. and R. Carter, 2007. 'Discourse markers and spoken English: Native and learner use in pedagogic settings.' *Applied Linguistics*, 28(3), 410-439.

Gries, S. Th. 2009. *Statistics for linguistics with R. A practical introduction*. Berlin: Mouton de Gruyter

Gries, S. Th. and S. Wulff, 2013. 'The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research.' *International Journal of Corpus Linguistics,* 18(3), 327-356.

Hanks, W.F., 1992. The indexical ground of deictic reference. In: A. Duranti and C. Goodwin (eds.), *Rethinking context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press, 43-76.

Hanks, W.F., 2009. 'Fieldwork on deixis.' *Journal of Pragmatics,* 41,10-24.

Hanks, W.F., 2011. Deixis and indexicality. In: W. Bublitz and N. R. Norrick (eds.), *Foundations of Pragmatics* (vol. 1 of Handbooks of Pragmatics). Berlin: Mouton de Gruyter, 315-346.

Heritage, J., 2015. '*Well*-prefaced turns in English conversation: A conversation analytic perspective.' *Journal of Pragmatics*, 88, 88-104.

Jucker, A. and I. Taavitsainen, 2000. 'Diachronic speech act analysis: Insults from flyting to flaming.' *Journal of Historical Pragmatics*, 1(1), 67-95.

Jucker, A. and I. Taavitsainen, 2008. Apologies in the history of English. Routinised and lexicalised expressions of responsibility and regret. In: A. Jucker and I. Taavitsainen (eds.), *Speech Acts in the History of English*. Amsterdam: John Benjamins, 229-244.

Jucker, A., G. Schneider, I. Taavitsainen and B. Breustedt, 2008. Fishing for compliments: Precision and recall in corpus-linguistic compliment research. In: A. Jucker and I. Taavitsainen (eds.), *Speech Acts in the History of English*. Amsterdam: John Benjamins, 273-294.

Kallen J. and J. Kirk, 2008. *ICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banríona.

Kereckes, J., 2007. 'The co-construction of a gatekeepig encounter: An inventory of verbal actions.' *Journal of Pragmatics*, 39(11), 1942-1973.

Kohnen, T., 2008. Tracing directives through text and time.Towards a methodology of a corpus-based diachronic speech act analysis. In: A. Jucker and I. Taavitsainen (eds.), *Speech Acts in the History of English*. Amsterdam: John Benjamins, 295-310.

Kohnen, T., 2015. Speech acts: A diachronic perspective. In: K. Aijmer and C. Rühlemann (eds.), 2015. *Corpus Pragmatics: A handbook*. Cambridge: Cambridge University Press, 52-83.

Lakoff, R., 1974. Remarks on 'this' and 'that'. In *Proceedings of the Chicago Linguistics Society,* 10, 345-356

Leech, G., M. Hundt, C. Mair and N. Smith, 2009. *Change in contemporary English. A grammatical study*. Cambridge: Cambridge University Press

Levinson, S.C., 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S.C., 2004. Deixis. In: L. R. Horn and G. Ward (eds.), *The Handbook of Pragmatics*. Malden, MA: Blackwell, 97-121.

Lindblom, K., 2001. 'Co-operating with Grice: A cross-disciplinary metaperspective on uses of Grice's cooperative principle.' *Journal of Pragmatics*, 33(10), 1601-1623.

Lyons, J., 1977. *Semantics. Volumes I and II.* Cambridge: Cambridge University Press

McCarthy, M. and A. O'Keeffe, 2010. Historical perspective: What are corpora and how have they evolved? In: A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3-13.

Mey, J., 2001. *Pragmatics: An introduction*. Oxford: Blackwell.

Millar, S., 2015. Blathering beauties: The use of pragmatic markers in an Irish beauty blog. In: C. Amador-Moreno, K. McCafferty and E. Vaughan (eds.), *Pragmatic Markers in Irish English*. Amsterdam: John Benjamins, 292-317.

Norrick, N., 2015. Interjections. In: K. Aijmer and C. Rühlemann (eds.), 2015. *Corpus Pragmatics: A handbook*. Cambridge: Cambridge University Press, 249-273.

Romero-Trillo, J. (ed.), 2008. *Corpus Linguistics and Pragmatics: A mutualistic entente*. Berlin: Walter de Gruyter.

Rühlemann, C., 2007. *Conversation in Context: A corpus-driven approach*. London: Continuum.

Rühlemann, C., 2010. What can a corpus tell us about pragmatics? In: A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 288-301.

Rühlemann, C. and K. Aijmer, 2015. Corpus pragmatics: Laying the foundations. In: K. Aijmer and C. Rühlemann (eds.), *Corpus Pragmatics: A handbook*. Cambridge: Cambridge University Press, 1-26.

Rühlemann, C. and S. Gries, 2015. 'Turn order and turn distribution in multi-party storytelling.' *Journal of Pragmatics*, 87, 171-191.

Rühlemann, C. and 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.

Rühlemann, C. and O'Donnell, M.B., 2015. Deixis. In: K. Aijmer and C. Rühlemann (eds.), *Corpus Pragmatics: A handbook*. Cambridge: Cambridge University Press, 331-359.

Schauer, G. and S. Adolphs, 2006. 'Expressions of gratitude in corpus and DCT data: Vocabulary, formulaic sequences and pedagogy.' *System*, 34, 119-134.

Schneider, K., 2012. 'Appropriate behavior across varieties of English.' *Journal of Pragmatics*, 44(9), 1022-1037.

Sinclair, J.M., 2001. Preface. In: M. Ghadessy, A. Henry and R.L. Roseberry (eds.), *Small Corpus Studies and ELT: Theory and practice*. Amsterdam: John Benjamins, vii-xv.

Stivers, T., 2015. 'Coding social interaction: A heretical approach in Conversation Analysis.' *Research on Language and Social Interaction*, 48(1), 1-19.

Taavitsainen, I., A. H. Jucker and J. Tuominen (eds), 2014. *Diachronic Corpus Pragmatics* (vol. 243 of Pragmatics & Beyond New Series). Amsterdam: John Benjamins.

Tao, H. (2003). Turn initiators in spoken English: A corpus-based approach to interaction and grammar. In: P. Leistyna and C. Meyer (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 187-208.

Vaughan, E. and B. Clancy, 2013. 'Small corpora and pragmatics.' *The Yearbook of Corpus Linguistics and Pragmatics*, 1, 53-73.