Interpreting Quantitative Data in Corpus Linguistics

Susan Hunston[1]

[1] University of Birmingham, UK

Abstract

Quantitative information has become increasingly important in Corpus Linguistics, and increasingly sophisticated as measures that are sensitive to how language works have become more readily available. Questions around the use of quantitative information are driven by the need in Corpus Linguistics to innovate methodologically and theoretically.

In 'phase 1' studies, corpora from different geographical areas, or chronological times, or registers, are compared by quantifying the relative frequency of given grammatical or semantic categories. Such methods have underpinned substantial advances, for example the Longman Grammar of Spoken and Written English, work on Systemic-Functional Linguistics, and work comparing learner varieties of English, among many others.

'Phase 2' studies prioritise lexis over grammar and individual wordforms over categories of form or meaning. In these studies, frequency is often reduced to a concept of 'typicality' or 'centrality'. Comparison between corpora is usually not the identifying feature of such work. Examples include Sinclair's work on Units of Meaning, or Frances and Hunston's work on grammar patterns. The key aspects of phase 2 studies are its exploratory, 'bottom-up' approach and the novelty of its insights.

A challenge for Corpus Linguistics is to marry the rigour of quantitative measures with innovation in insight. One way of doing this is to allow numbers to drive the way that information in the corpus is organised. This is what I term 'phase 3' studies. These are illustrated by a study of adjectives in a corpus of comments about university teaching staff (see Millar and Hunston 2015), and by a study of lexis in a corpus compiled from an interdisciplinary academic journal (see Murakami et al in press). In both cases the initial corpus work treats the corpus as a 'bag of words', allowing co-occurrence calculations to organise the data before linguistic considerations are brought to bear. Phase 3 studies remain true to a data-driven approach to corpora. They achieve a sketch rather than an analysis of a corpus.

Keywords

Lexis, grammar, patterns, bag of words, data driven corpus linguistics

Introduction

There are two distinguishing features of Corpus Linguistics as a field of research. Firstly, it involves naturally-occurring discourse, and in relatively large quantities. What counts as 'relatively large' depends on the individual study and can be anything from a few hundred thousand words to hundreds of millions of words. Among the assumptions that lie behind

Corpus Linguistics, though, is the view that there are aspects of language use that are important but that are invisible to the human reader of texts. In particular, the relative frequency with which words, phrases, and grammatical categories are used is of importance but can be established only with the help of search software. In very basic cases, the language of the corpus is rearranged so that a reader is presented with an altered and focused view. Taking this one or more steps further, quantitative information is given that replaces the human reading process.

Secondly, Corpus Linguistics attempts to make contributions to linguistic theory that are informed by quantitative information. As will be noted below, those contributions may relate to the mechanisms by which language changes over time, or to the nature of the difference between registers, or to the relationship between lexis and grammar. The question of 'what is language like' is one that Corpus Linguistics seeks to answer. In some cases, that answer is very much aligned with other approaches to language; in other cases less so.

Work in Corpus Linguistics has grown exponentially over the last three decades, and the quantitative tools it routinely uses have become more sophisticated. An area of constant exploration is the role that a technical expertise in language structure – lexis, grammar, discourse – plays in relation to expertise in quantification. As the discussion below will indicate there are various types of combination, from 'mainly language with some numbers thrown in' to 'mainly numbers with little regard for language'. This is not simply a matter of finding a balance, but constantly exploring new ways of approaching language so that possibilities of finding new knowledge constantly come into view. Whatever Corpus Linguistics is, it is not static.

In this paper I shall present a view of Corpus Linguistics that conceptualises it in three phases, distinguished by the use made in each phase of quantitative data. I am going to be talking about corpora of English, though much of what I say will be applicable to other languages.

1. Laying the foundation: quantifying language categories

One of the greatest contributions of corpus linguistics to theoretical linguistics is the opportunity it affords for quantifying the comparative frequency of various linguistic categories that are identified and tagged in corpora. This activity has not been without criticism. Chomsky, for example, has famously offered the view that simple quantity adds nothing to the explanatory function of theory. It is true that establishing comparative frequencies does not change our knowledge of what can be said, but it does alter our understanding of what is typically said, and under what circumstances. It also permits comparison between corpora and as a consequence geographical varieties can be compared, changes in language over time can be observed, and the stages of language development (in children or in learners) can be described.

A key example of this kind of work is the Longman Grammar of Spoken and Written English (Biber et al 1999), which is based on the grammatical description of English established by Quirk et al (1972). As well as quantifying a whole range of categories across the whole corpus, LGSWE gives separate figures for different registers: conversation, fiction, news and academic. The traditional categories applied to the verb phrase (tense, aspect, voice) are quantified across the registers. Biber et al show, for example, that verb phrases as a whole are most frequent in conversation and least frequent in academic prose (p456), that present tense verbs are proportionally more frequent in conversation and academic prose while past tense is most frequent in fiction (p456), that present progressive is particularly frequent in

conversation while past progressive is proportionally more frequent in fiction (p462), and that whereas active voice is massively more frequent than passive overall, passives are found proportionally most frequently (comprising about 25% of all verb phrases) in academic prose (p476).

An example of comparative work in the same tradition is the paper by Leech and Smith (2006), which is representative of a wider body of work by the same authors comparing two corpora from the early 1960s comprising written British and American English respectively (LOB and Brown) and their counterparts from the early 1990s (FLOB and Frown). Leech and Smith call attention to the reduction in frequency of many core modal verbs (e.g. *shall* has reduced in frequency by 44% in both US and British English; *would* has reduced by 6% in US English and by 12% in British English), a concomitant rise in semi-modals such as *be going to* (a rise of 54% in US English but no such rise in British English) and *be supposed to* (a rise of 15% in US English and 113% in British English). Other notable features include a rise in both varieties of the present progressive, including progressive uses of verbs often said to lack a progressive form, such as BE, and a fall in the overall use of the passive voice, even in academic prose (down 26% in US English and 17% in British English academic writing). Leech and Smith point to two apparent influencing factors: colloquialisation (written language becoming more like speech) and Americanisation. They point out, however, that these two trends do not always point in the same direction (the increased use of the subjunctive in British English, for example, going against the colloquialisation trend) and that British English does not always follow the American lead.

Quantitative corpus work of this kind serves to answer a number of questions, such as: To what extent do situational varieties of English differ from one another, and what are the main parameters of difference between them? To what extent do geographical varieties differ? In what respects has English changed over 30 thirty years, and what seem to be the main drivers of that change? It does so, quite simply, by counting the instances of a given category of use in two or more corpora and identifying amount of similarity or difference.

A similar methodology, though from a different theoretical perspective, is represented by, for example, Matthiessen (2006), who works within the Systemic-Functional Linguistics tradition. One reason for mentioning this work is that relative frequency is central to SFL in a way that it is not central to other models. Halliday conceptualises grammar as comprising a series of paradigmatic systems, the branches of which are weighted probabilistically. Meaning is created by the contrast between the selected item and its contrasting possibilities; their relative probabilities determine the degree of 'markedness' of the selection. Registers, for Halliday, are distinguished by the specific probability weightings that apply to them. It is clear, therefore, that the frequency of given categories in individual texts, in registers (that is, collections of texts) and in the language overall, will be crucial to SFL.

Annotating a corpus for the categories central to SFL (for example, distinguishing process types, or types of Theme) remains a largely manual enterprise, though assisted by mark-up and quantification software such as the UAM Corpus Tool (O'Donnell). Once annotation has been carried out, however, quantification can be carried out. Matthiessen demonstrates that registers are indeed distinguished by the relative frequency of grammatical categories in them, and he also shows the variation of individual texts within the register.

2 'Quantifying' phraseology: a lexical approach

The second 'phase' I draw attention to adopts a lexical view of language, and is often treated as qualitative rather than quantitative. This approach to Corpus Linguistics is based on the observation of pattern in concordance lines, where a word or short phrase is the node of the line and the few words occurring before and after the phrase are shown for each occurrence. The job of the researcher is to identify patterns of use. The attitude towards quantification is relatively casual and implicit, but is nonetheless of importance. Two examples will be given here, both using the Bank of English corpus (HarperCollins Publishers and the University of Birmingham).

The first example uses the quintessentially British phrase 'cup of tea' (a phrase used as a demonstration by Sinclair and by Danielsson). The Bank of English contains over 2,400 instances of this phrase, of which 100 random examples are selected for this illustration. Preceding 'cup of tea' are a number of different elements:

- Indefinite article: 'a cup of tea' (65 instances)
- Indefinite article + adjective: 'a nice/quick/calming cup of tea' (9 instances)
- Possessive: 'my/your/Linda's cup of tea' (7 instances)
- Other determiners: 'the cup of tea', 'another cup of tea', 'every cup of tea' (6 instances)

The obvious observation here is that 'cup of tea' is used with the indefinite article in nearly 75% of cases. But a further observation that can be made is that 'a cup of tea' and 'my cup of tea' are dissimilar in meaning ('They planned to celebrate with a quiet cup of tea' as opposed to 'She's not my cup of tea'). In other words, 'a cup of tea' is a container with brown liquid in it, whereas 'my cup of tea' refers to one's preference or otherwise for an individual.

To what extent is it true that 'possessive + cup of tea' has this metaphoric use exclusively? To test this, instances of 'my/his/her/their cup of tea' in the BoE are identified (160 in total), and 100 random lines selected. The ones that do not indicate preference are then identified; there are 27 of them. Thus, 73% of instances of 'possessive + cup of tea' have a non-literal interpretation. A further observation is that the non-literal instances tend to include either a negative or a comparator ('not my cup of tea' or 'more/just/entirely my cup of tea'). A further count is then carried out on the 100 lines, identifying those including a negative or comparator and those with literal or non-literal meaning.

	Literal	Non-literal	Total
Negative	1	58	59
Comparator	1	12	13
Neither	25	3	28
Total	27	73	100

The usual conclusion from this kind of study is that the non-literal meaning of 'cup of tea' is reliably associated with negative and comparative phraseology, while the literal meaning is

rarely used with these words. More significantly, information such as this forms the basis of Sinclair's concept of Idiom Principle (1991; 2004), Hoey's (2005) proposal of lexical priming, and Deignan's work on metaphor and phraseology. The essence of the theory is that meaning is expressed in phraseologies (as opposed to phrases) rather than in words, so that a particular phraseology such as 'x BE negative / comparator possessive cup of tea' has one meaning whereas 'determiner (adjective) cup of tea' has another. Importantly, though, the association is not one-to-one. It is entirely open to speakers of English to produce an utterance that has the formal features of the literal use but has the non-literal meaning, and vice versa. Sinclair argues that much of English operates within this somewhat grey area between the absolute fixed phrase and the entire open choice. It is often exploited in jokes, such 'Chocolate is not my cup of tea'. The second point to make is that the precise numbers involved here are not of particular interest. The difference between the relevant numbers have to be significant, not only in the statistical sense but sufficiently to give confidence that the generalisations drawn from them are accurate.

My second example is a study of verb complementation (Hunston 2003). The study was based on a preliminary observation that when the lemma DECIDE is followed by a that-clause, the verb wordform is likely to be *decided*, whereas when it is followed by a wh-clause the predominant wordform is *decide*. Taking 100 instances of 'decide / decided' and 'that / whether', there are 7 instances of 'decide that', 16 instances of 'decide whether', 4 instances of 'decided whether' and 73 instances of 'decided that'.

	that	*whether*
decide	7	16
decided	73	4

Furthermore, the wordform 'decide' when followed by 'whether' is frequently preceded by indications of obligation, necessity or volition, such as 'will decide', 'has yet to decide', 'was forced to decide' and so on. This was later formalised with a study that looked at 10 verbs, each occurring with both that-clauses and wh-clauses. The conclusion was that non-finite verb-forms co-occur with wh-clauses that construe hypothetical actions whereas finite verb-forms co-occur with that-clauses that construe actual situations. In addition, a concept known as 'semantic sequences' was developed – this suggested that concordance lines can be read to identify 'what is often said', thus moving from lexis to grammar to discourse. Further examples are given in Hunston (2011), where phraseology is used to identify how concepts such as 'ideas', 'claims' and 'discoveries' are evaluated in popular science discourse.

In this section I have drawn on a tradition which, as noted above, is often described as qualitative rather than quantitative. What I have tried to point out, however, is that this qualitative work does rely on measures, sometimes intuitively rather than formally established, of relative and comparative frequency. What is valued in this kind of work is observation or noticing, that is, the identification of classes of object that may not exist as a class outside that context and which have a meaning- or function-related definition rather than a formal one.

3. Enhancing innovation

All the kinds of quantitative approaches outlined above continue to be used in Corpus Linguistics. If we try to combine the benefits of phases 1 and 2 we might arrive at the following desiderata:

- The statistics should be robust and should be applicable to language data;
- The method of working should rest on as few preconceptions about language as possible, and be as exploratory as possible;
- The outcome should offer genuine insight into language and discourse.

I am now going to give three examples of what might be considered to be phase 3 research. I would describe this work as quantity-led. Like the phase 2 work outlined above, there is an attempt to rely on information that emerges from the text ('trust the text', as Sinclair says), rather than on information that is presupposed. Phase 2 work is based on the human observation of the behaviour of large numbers of a single word or phrase ('decide' or 'cup of tea') and builds theory bottom-up from such observations. Concordancing software rearranges the data – the texts – to permit that observation. In phase 3 work, statistical packages take over the role of rearranging the data. So although the approach is quantitative rather than qualitative, and relies on numbers rather than observation, to my mind it has the same bottom-up approach that moves from evidence to theory.

One example of this phase comes from many decades ago; the others are more recent.

Example 1: Multi-Dimensional Analysis (Biber, 1988 and subsequent publications)

The essence of Biber's MDA is the observation that language is different in different contexts. Academic prose is different from newspaper prose; political speeches are different from casual conversation, and so on. The differences are easily recognisable, but rest on a multiplicity of frequency variations. Biber is not the only linguist to observe this (Halliday does this in a more theoretically-informed way, and Matthiessen, as noted, has added statistical rigour to that model), but Biber has carried out more extensive investigations of this type than anyone else. Unlike Halliday, he begins with a 'common sense' notion of register and with an eclectic mix of language features. The mix is intentionally eclectic because unlike Halliday he does not make presuppositions about which features will be significant in distinguishing between registers. The various sub-corpora are then tagged with the language features, and the strength of co-occurrence of those features is calculated. The result is a number of factors, each consisting of a set of features that either attract or repel each other. The factors are then interpreted in terms of what they mean in terms of discourse. 'Informational' versus 'involved' is one factor; 'narrative' versus 'non-narrative' is another. At this point the factors are renamed 'dimensions'. Corpora of texts belonging to two different registers may be alike on one dimension and different on another. Plotting registers as being more or less like each other therefore requires multiple dimensions.

The original five dimensions proposed by Biber have been widely used in subsequent work by him and others, but in other work the dimensions have been worked out anew, permitting the approach to be applied to texts that may not be categorised in traditional register categories. Further refinements to the set of language features have also been made.

A project carried out 2013-15, led by Thompson and advised by Biber ('Interdisciplinary Research Discourse' or IDRD), used the current set of language features from Biber's studies,

but attempted two innovations. The first was to generate new dimensions from a corpus of articles from a single academic journal (Global Environmental Change). We identified, for example, 'system-oriented' versus 'action-oriented'; 'explicit argumentation' versus 'implicit argumentation'; greater or lesser degrees of 'spoken-ness'. The second was to avoid a prior division of our corpus into registers. We had deliberately selected an interdisciplinary journal for our project, with the aim of investigating ID discourse. Although we hypothesised in advance that different disciplinary discourse styles would emerge from the journal, we were not able to, and indeed did not want to, divide the articles in the journal between those disciplines in order to arrive at sub-corpora that could then be compared. What we did instead was to assign each article – each text – a value on each of the identified dimensions. We then derived clusters of texts (or 'constellations') that shared values on those dimensions. Depending on how the figures are interpreted – with greater or lesser granularity – we identify 3 or 6 constellations. We are able then to identify the type of research being reported in each, basically on stylistic grounds. Figure 1 shows the output of this project: each constellation mapped against the six dimensions.

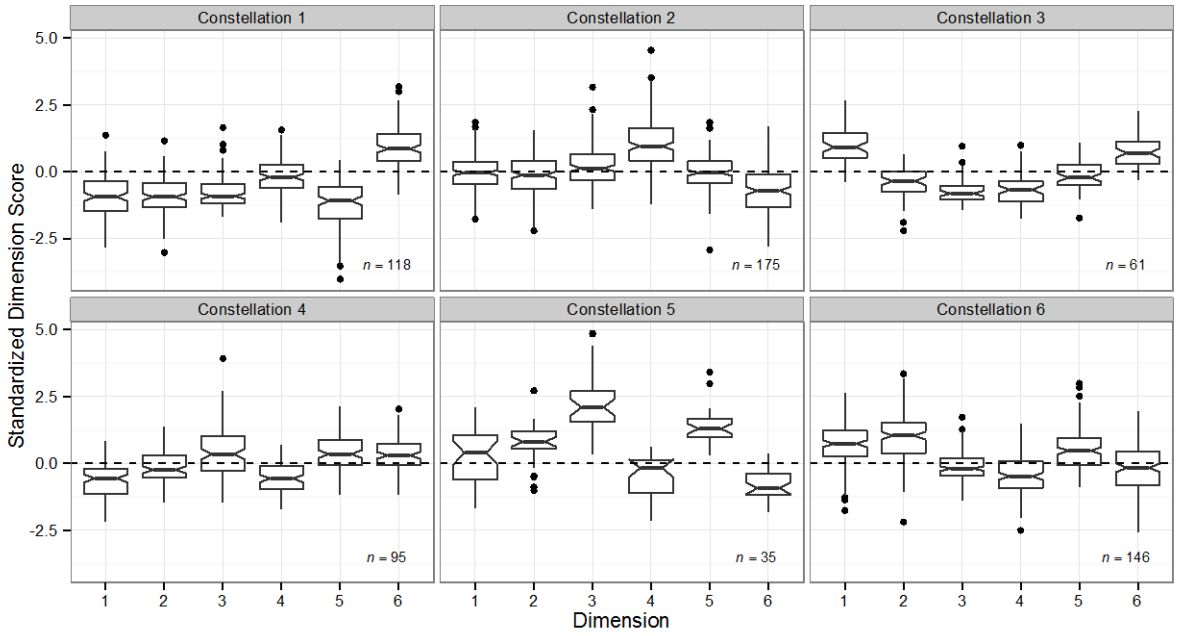

Figure 1: Constellations in the journal *Global Environmental Change*

In some senses, MDA is not 'bottom-up', because the language features are pre-determined and in most versions of MDA the registers are pre-determined too, on external criteria (for example, a newspaper text is found in a newspaper, whilst academic prose is found in academic journals). It is, however, exploratory in that which language features will be significant is not pre-determined by a theory of register. As the IDRD project shows, it is also capable of adaptation to a method that does not pre-determine register divisions. This has revived the exploratory nature of the approach.

Example 2: Collostructions (Stefanowitsch and Gries, 2003)

My second example uses quantitative information to enhance a linguistic theory. That theory is the theory of Constructions (e.g. Goldberg, 2006), which offers an alternative (to Chomsky's) view of how language is stored in the brain. What is stored, it is suggested, is not an underpinning 'deep grammar', but a very large number of constructions. Constructions might be divided into three types: very general constructions whose occurrence is not

restricted by lexis, such as the interrogative construction; constructions that typically co-occur with a restricted set of lexis, such as the ditransitive construction; and constructions that comprise phrases with a limited amount of variation, such as the 'accident waiting to happen' construction. Constructions have meaning, indeed they are defined as a recognisable connection between form and meaning, and Goldberg makes the point that constructions can over-ride the canonical meaning of the words that occur within them, but they do nonetheless have typical lexical realisations. Stefanowitsch and Gries task themselves with determining what 'typical' means in this context. They propose that a word in a construction be termed a 'collostruction'. For a given construction 'c' they find a list of words 'w' occurring in it, or collostructions. From a corpus they find the number of instances of 'c+w', the total instances of 'c' and the total of instances of 'w'. They use Fisher's Exact Test to compare these numbers and so derive the collostructional strength of each word. As with measures of collocation, what matters is not the frequency of co-occurrence itself but the relative frequency in comparison with the overall frequency of both items.

For example, they study the ditransitive construction ('give someone something', 'promise someone something', 'tell someone something'), arguing that the 'agent + recipient + theme' (Stefanowitsch and Gries 2003: 227) feature of the construction has a meaning that is not completely dependent on the meaning of the verbs in it. Thus, a transitive verb such as 'bake' has a ditransitive meaning when used in the 'two noun' construction ('bake someone something'). Collostruction analysis confirms that the verbs most intuitively connected with the ditransitive do indeed have the highest collostructional strength: 'give', 'tell', 'send', 'offer', 'show', 'cost', 'teach' and so on. They further observe that although the 'central' meaning of 'give' (or 'transfer possession') is present in some of these verbs (such as 'send'), verbs with other meanings, such as 'tell' or 'show', are also high up the list. This suggests that the construction has a range of meanings rather than one.

Stefanowitsch and Gries's work aligns with, though is not derived from, classic phase 2 approaches to language. For example, constructions such as 'accident waiting to happen' look very similar to Sinclair's Units of Meaning. Those such as 'predicative *as*' or 'causative *into*' are identifiable as grammar patterns (Frances et al 1996; Hunston and Francis 1999). The essential difference is that constructions are offered as a psycholinguistic theory, whereas Sinclair and Frances's work contributes to a model of language that may or may not match that in the brain. Other observations are very similar, such as the recognition that a pattern / construction may have meaning that is independent of the words found in them, or the hypothesis that language changes as words begin to be used in patterns / constructions by semantic analogy with the words formerly used in them.

Where Stefanowitsch and Gries's work is different is that it offers a robust quantitative approach to the question of collostruction, rather than relying on impressions of frequency. The numerical value given to collostructional strength is used to offer an insight into how constructions come to express meaning, as this is said to be derived from the meaning of the collostructions with the highest strength.

Example 3: Co-occurrence measurements as exploratory mechanisms

In this third example I include two research projects carried out recently, both of which involve exploiting quantitative measures to explore corpora. These are 'bottom-up' in the sense that there is no pre-emptive model of language at the outset, but unlike phase 2 studies they have numbers rather than words at their heart. For me, they have the genuine sense of exploring the unknown and of encountering unexpected insights that phase 2 studies also

have. The studies are somewhat controversial, in that they treat language as a 'bag of words', that is, without adding linguistic knowledge about structure, meaning and so on. Linguists normally look askance at 'bag of words' studies, but I do think they offer interesting new ways of carrying out 'corpus-driven' research.

In introducing these examples I have used the term 'co-occurrence', meaning that two words frequently co-occur in the same text. Co-occurrence of words within a short span (i.e. 'collocation') is a traditional concept in Corpus Linguistics, as noted at the beginning of this paper. Collocation, as Firth famously almost said, gives us a lot of information about a word: its denotational and connotational meanings, for example. It has been widely accepted with Corpus Linguistics that collocation needs to be measured within a fairly short span: + / - 5 words from the node is common. Beyond this, the influence of a given word is negligible. In the studies described in this section, a rather different view of co-occurrence is taken. There is no node word and no directional influence, and the purpose is not to find out more about an individual word. Rather one aim is to gain novel insights into a set of texts by observing the co-occurrence of words within them. The second aim is to gain novel insights into those words by organising them into groups according to the strength of their co-occurrence in given texts.

The first is a project initiated by Neil Millar (Millar and Hunston, 2015). The corpus studied was compiled from the website Rate My Professors (www.ratemyprofessors.com), where students are invited to grade their professors with numerical scores and also to leave free-text comments about those same individuals. Millar compiled a corpus of about half a million individual comment texts amounting to about 25 million tokens. From that corpus he extracted all adjectives occurring after a copular verb (e.g. 'Prof x is wonderful / awful'), giving a list of 1856 adjectives. He then divided the corpus, assigning texts based on who they described and whether the overall score was high, medium or low. This resulted in a spreadsheet with 1856 adjectives and 41633 individual instructors as the axes. The final step was to use Principle Component Analysis to assign co-occurrence strength to the adjectives, measuring each adjective against all the others. The outcome is a re-arrangement of the adjectives in the corpus. They are grouped according to the likelihood of their co-occurrence. The identified groups are:

Helpful, willing, sweet, caring, available, understanding, approachable. ..

Funny, hilarious, entertaining, fun, sarcastic, crazy. ..

Horrible, unclear, terrible, confusing, unorganized, boring. ..

Rude, condescending, arrogant, unhelpful, mean. ..

Passionate, brilliant, intelligent, knowledgeable, inspiring, engaging, interesting, smart. ..

Tough, hard, difficult, fair, intimidating, clear. ..

Hot, gorgeous, young, beautiful, attractive. ..

What Millar has essentially done is to replace (1) an impressionistic or intuitive categorisation of adjectives according to their meaning as perceived by the analyst with (2) a taxonomy of evaluative meaning based on the collective language behaviour of an on-line community (i.e. the people who express their opinions on the RMP website), with the taxonomy derived through bottom-up quantitative measures. It reveals community-specific meaning sets, such

as ‘tough but fair’. It also suggests a consumer-oriented view of university teaching staff. The existence of the ‘helpful’ and ‘approachable’ group is one indication of this. Another is the occurrence of ‘engaging’ and ‘interesting’ as part of the ‘intelligent’ group. In short, a methodology that presupposes no linguistic knowledge other than the identification of adjectives – a ‘bag of selected words’ – has given information about categories of meaning.

My second example under this heading is another study undertaken as part of the IDRD project outlined above and used the same corpus. One of our aims in that project was to find bottom-up and novel ways to explore a corpus whose contents were diverse but not in known ways. To do this we used topic modelling (Murakami et al., in press). Topic modelling measures the strength of likelihood of co-occurrence of words in texts. The researcher specifies what is meant by ‘a text’ and also, crucially, specifies the number of groups of words (known as ‘topics’) to be identified. For the IDRD project, each journal article was divided into ‘texts’ of approximately 300 words, and the number of topics was set at 60 (after some trial and error). A further level of pre-processing applied a stop list of some grammatical words, and stemmed the words so that all word-forms in a lemma were treated as the same item. The topicmodels package (Grün & Hornik, 2011) in R was used to build the topics.

The output from topic modelling is a set of lists of words that are grouped according to the probability of their co-occurrence within the specified texts. What the lists give us, first of all, is a sense of the ‘aboutness’ of the corpus. Here are a few of the themes that can be identified (from Murakami et al., in press):

- Kinds of natural environment e.g. [forest, carbon, deforest, topic, land, area, cover, conserve, forestri, timber = Topic 19]; [flood, sea, rise, coastal, area, level, protect, impact, loss, sealevel = Topic 23]; [speci, biodivers, conserve, area, ecosystem, plant, divers, protect, veget, site = Topic 39]
- Geographical locations e.g. [local, scale, level, region, differ, spatial, nation, these, across, which = Topic 32]; [country, develop, nation, world, intern, their, india, global, industri, most = Topic 35]; [region, afrida, south, southern, europ, area, central, north, most asia = Topic 60]
- Kinds of human economic activity e.g. [crop, product, agriculture, soil, food, yield, increase, fertile, use, plant = Topic 4]; [energi, use, fuel, technolog, power, sector, transport, consumpt, industry = Topic 5]; [product, sector, trade, import, increase, export, consumpt, fish, market, economy = Topic 34]
- Political institutions and actions e.g. [govern, institute, actor, state, network, power, polit,author, their, role = Topic 6]; [polici, polit, this, issu, maker, question, decis, make, what, which = Topic 36]; [program, state, it, us, govern, agenc, nation, committee, offici, support = Topic 52]
- Aspects of risk e.g. [adapt, vulner, capac, or, sensit, social, cope, exposur, measure, abil = Topic 9]; [environment, global, problem, environ, economy, concern, issu, chang, secur, polit = Topic 15]; [risk, health, disast, effect, hazard, diseas, people, affect, reduc, potenti = Topic 20]
- Research actions e.g. [group, respond, particip, interview, survey, their, question, they, respons, inform = Topic 26]; [studi, this, analysi, paper, approach, section, discuss, case, how, present = Topic 38]; [indic, variabl, measur, eqsym, valu, signific, index, effect, correl, relationship = Topic 44]
- Groups of people e.g. [individu, their, public, respons, action, people, they, behaviour, perceive, percept = Topci 16]; [group, respond, particip, interview, survey, their, question, they, respons, inform = Topic 26]

- Modelling the future e.g. [model, use, simul, base, paramet, each, which, result, repress, function = Topic 1]; [will, futur, may, this, can, if, more, like, current, need = Topic 11]; [would, could, not, if, might, or, this, but, ani, should = Topic 30]

Each 'text' and each article in the corpus can then be described in terms of the probability of occurrence of the various topics, and a topic profile of each 300-word text and each article can be derived. Various studies are then done, for example identifying those topics which are linked to the beginnings or ends of articles, or those topics which have become more or less frequent over time. For example, Topic 45 (pollution, control, air, ozone, environment, waste, effect, deplete, which, problem) and Topic 49 (global, warm, increase, atmosphere, change, climate, temperature, effect, concentration, level) have a higher level of probability in the early years of the journal (1990-1995) whereas Topic 18 (communities, people, local, their, tradition, live, indigenous, which, knowledge) and Topic 56 (farmer, household, their, income, farm, village, migration, livelihood, food, rural) have a higher level of probability in the later years (2005-2010). As the journal, and the field, progresses, its focus changes from agenda-setting and global pollution concerns, to empirical research and a greater concern for the 'local' as a contributor to the 'global'. It is also possible to identify papers that focus on a single topic and those that incorporate a number of topics. In this interdisciplinary journal, then, some articles are tightly focused whereas others are more wide-ranging.

There are, of course, other methods of tracking topics through time, or of identifying the focus of a given paper. Simply reading the paper is an obvious method! What is significant about using topic modelling is precisely that it does not rely on pre-existing ideas about what a topic might consist of. It throws open the notion of 'aboutness' and uses a data-driven way of organising the content of a large corpus.

Conclusion

In this paper I have distinguished between a number of approaches to quantification, or phases, in Corpus Linguistics. I have suggested that there tends to be a tension between statistical rigour and a desired objective of using data-driven methods to drive theoretical innovation. However, methods of identifying word co-occurrence provide a way of organising a corpus to lead to new insights.

One of the key questions for Corpus Linguistics is how a corpus might satisfactorily be 'analysed'. In general, the investigation is top-down, in the sense that a question is asked of the corpus and means devised to find the answer to the question. The project relating to 'Rate My Professors', described above, is one such investigation, where the question: 'what categories of individual qualities are discernible from the comments made' is answered using the strength of co-occurrence of adjectives as the research method. A guiding principle in Corpus Linguistics, however, is that one should 'trust the text' (Sinclair 2004). This means conducting research in which the question, as well as the answer, is not precisely known. For example, Sinclair undertook word by word analysis (leading to the Cobuild dictionary of English) in the belief that in language lexis is more important than grammar and that the behaviour of individual words underpins grammatical systems. This general belief and the subsequent investigation led to the theory of language known as 'units of meaning'. In my opinion, Topic Modelling offers a means, not to analyse a corpus but to offer a sketch of it in a way that optimally trusts the text.

References

Biber D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

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

Goldberg A. (2006). *Constructions at Work: The nature of generalization in language*. Oxford University Press.

Grün B. and Hornik K. (2011). Topicmodels: An R Package for fitting topic models. *Journal of Statistical Software* 40(13). Retrieved from http://www.jstatsoft.org/v40/i13.

Hunston S. (2003). Lexis, wordform and complementation pattern: a corpus study. *Functions of Language* 10(1): 31-60.

Hunston S., Murakami A., Thompson P. and Vajn D., submitted. 'Multi-dimensional analysis, text constellations, and interdisciplinary discourse.

Leech G. and Smith N. (2006). Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English. In Renouf, A. and Kehoe, A. editors, *The Changing Face of Corpus Linguistics*. Rodopi. Pages 185- 204.

McEnery T. and Hardie A. (2012). *Corpus Linguistics*. Cambridge University Press.

Matthiessen C.M.I.M. (2006). Frequency profiles of some basic grammatical systems: An interim report. In Thompson G. and Hunston S. *System and Corpus: Exploring connections*. Equinox. Pages 103-142.

Millar N. and Hunston S. (2015). Adjectives, communities, and taxonomies of evaluative meaning. *Functions of Language* 22(3): 297-331.

Murakami A., Thompson P., Hunston S. and Vajn D., in press. What is this corpus about? Using topic modelling to explore a specialized corpus. *Corpora*.

O'Donnell, M. UAM Corpus Tool. www.corpustool.com

Sinclair, J. (2004). *Trust the Text: Language, corpus and discourse*. London: Routledge.

Stefanowitsch A. and Gries S. (2003). Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-243.