33

Corpus and systemic functional linguistics

Serge Sharoff

Introduction

Linguistic research requires empirical evidence to give satisfactory answers to questions such as: to what extent a phenomenon X is present in the system of language? Or what is the difference between choices X and Y? Such evidence can be provided by a corpus, 'a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language' (Sinclair 1996: 2). A modern computer-based corpus comes with an interface for retrieving appropriate linguistic constructions, such as sequences of word forms, also often lemmas (that is, dictionary headwords) and generic part-of-speech (POS) tags.

With the proliferation of texts in electronic form, linguistic research now has potentially unlimited access to evidence from manually collected corpora, such as the British National Corpus (BNC) (Aston and Burnard 1998), very large generic corpora derived from the Internet (Sharoff 2006b; Baroni et al. 2009) and smaller corpora for specific domains (Baroni and Bernardini 2004) or genres (Nesi and Gardner 2012).

This chapter will explore the field of corpus linguistics, with greater emphasis put on topics specifically related to systemic functional linguistics (SFL). The major topics covered in this chapter are:

1 the historical and genetic links between SFL and corpora;
2 the principles underlying the use of corpora in SFL-based studies;
3 studying language from the lexical end and from the grammatical end using corpora;
4 corpus composition and SFL; and
5 recommendations for corpus use and for enriching SFL theory and descriptions.

Historical perspectives

From the historical perspective, many of the early developments in corpus linguistics have close links with the systemic studies. First, the very notion of collocations was coined by

J.R. Firth (1957 [1951]), with his statement 'you shall know a word by the company it keeps' being arguably one of the best-known quotes in corpus research. Another important milestone was Halliday's PhD thesis completed in Cambridge in 1955 (Halliday 1959), which outlines 'descriptive grammar' of a single text in Chinese – namely, *The Secret History of the Mongols*. That study offered a corpus-based description of word classes, syntactic constructions, collocations and colligations – that is, the relations between words and grammatical patterns of their use. It was also instrumental in developing the first systemic account of grammatical analysis (Halliday 1957). Another historical link comes from the corpus studies performed by Halliday's PhD student, John Sinclair, who started with the English lexical studies project in 1963 and, over the course of the 1970s, moved to development of large corpora for lexicographic investigations. This research provided important contributions to the principles of corpus development and corpus-based lexicography, including the design of the Bank of English and of the *Collins-COBUILD* dictionary (Sinclair 1991). The latter was the first thoroughly corpus-based dictionary, which provided information on meaning distinctions, grammatical patterns of words and their collocations on the basis of their frequencies in a generic corpus. Since then, texts in electronic form have become much more common and this has enormously simplified corpus research. First, large, manually collected corpora were developed in the 1990s, such as the BNC (Aston and Burnard 1998); then, with the proliferation of Internet resources, ever-larger web-based corpora have been created by crawling the web for a number of languages (Sharoff 2006b; Baroni et al. 2009). Some linguistic studies also use interfaces to search engines as a source of corpus evidence. This situation led to numerous corpus studies in SFL research.[1]

The historical perspective also suggests a genetic link between corpus research and SFL – namely, that there are common strands naturally linking SFL research and corpus studies. One of the sources for this link comes from the emphasis made in SFL on social semiotics: language is a social phenomenon that is manifested in production of speech and text (Halliday 1978). This view leads to the system–process–product approach advocated in SFL: each text (this chapter, for example) and each linguistic unit in this text is a product of various processes of text production, an instantiation of a number of choices possible in the system of the English language conditioned by their use in the register of academic chapters. Unlike other kinds of linguistic analysis that emphasise the autonomy of syntax, a socio-semiotic approach to language has to deal with linguistic phenomena instantiated in their context of use, thus naturally promoting corpus-based analysis. As noted by Halliday (1992: 79): 'It seemed clear to me in 1960 that useful theoretical work in grammar was seriously hampered by lack of data.' In this view, the system of language is observable through instantiation, which needs to be quantified through corpus evidence.

Another genetic link between corpus studies and SFL is present in attention to variation. The speakers use language to achieve their communicative goals. The lexicogrammatical properties of their speech vary according to the context of its use, which leads to the concept of register as variation according to use. Another way in which the lexicogrammatical patterns vary comes from variation by the user. For example, it is possible to consider temporal, social or geographical variation (Gregory 1988). Similarly, variation is also at the heart of corpus linguistics: frequencies and patterns can be important by themselves, but the contrast between two kinds of uses is often more informative. For example, the absolute frequency of active vs passive constructions in a large representative corpus provides information about the preferred options in grammar, while variations in their frequency across different registers reveal the possible reasons for choosing between these options.

Critical issues and topics

Frequency and instantiation

When we talk about quantification of evidence, we need to distinguish data coming from different regions on the following cline of instantiation (Matthiessen 2006: 104):

> **Text level** logogenetic patterns with texts as processes, they unfold as meaning is made within individual texts or a group of texts produced in the same context.
>
> **Register level** more generic patterns, which are still specific to an individual register; for example, commands realised as imperatives are possible in instructional texts as well as in scientific articles, but with unequal frequency.
>
> **Language level** such patterns are inherent in the linguistic system, at least at a certain period in time.

In SFL, the natural objects of quantification are the choices in each system and their realisations through lexicogrammatical features. A corpus usually does not contain information beyond the level of linguistic forms; therefore the choices can be easily recognised on more delicate levels of the system. However, less delicate choices can be counted as well if they are recorded through annotation. For example, counting the collocations on a raw corpus is straightforward, while counting the types of process requires explicit annotation of the process types.

A system can have no marked choices – that is, the probabilities of choices are roughly similar – or it can have one unmarked choice, with considerably higher probability. Halliday (1992: 65) gives an example with marked probabilities (example (1)) and suggests that the distribution of the probabilities of marked and unmarked choices as 9:1 is relatively common in the system of language, since this allows a suitable balance of processing efforts in terms of the information-theoretic entropy principles.

(1) polarity: positive (0.9) / negative (0.1)

Expectations are at the heart of both language production and processing. The frequencies refer to indicative probabilities of the choices to be made within such systems. One can expect a certain ratio of choices of nominalisations in a research article vs a shop encounter dialogue. However, even at the level of an individual text, the expectations expressed in probabilities can vary as the text develops. In a research article, for example, such sections as 'Related work' or 'Research methodology' are likely to vary in their probabilities of choices: considerably more instances of past tense reporting can be expected in the former. In the most extreme case, the 'Bibliography' section will not contain major clauses at all.

Even within a more homogeneous section of a research article, some variations in the probabilities are still possible. Halliday (1992) illustrates logogenetic development of references to the concept of cracking within a scientific article. This starts from *how glass cracks* (*crack* Process; *glass* Actor), goes through *will make slow cracks grow* (*crack* as nominalised caused Actor; property *slow*) and, finally, develops to a fairly complex nominalisation *decrease the crack growth rate* (property nominalised as *rate*). If there is a suitable measure for grammatical complexity, a corpus investigation can test this hypothesis on a large collection of scientific articles.

However, testing such hypotheses is not straightforward. The quantum mechanics metaphor of the wave–particle dualism is quite common in systemic studies: many linguistic phenomena can be described from two viewpoints, which offer complementary perspectives. With respect to corpora, we can notice the duality of description between form and function. Most of the counting by the computer can be done on the basis of forms – that is, the actual word forms as typeset in a text. This can be extended towards linguistic descriptions on a slightly higher level by using information about POS tags and lemmas, so that a search for *left* can produce different results depending on its grammatical environment: *my left hand* vs *he's just left*.

Nevertheless, even this basic generalisation can lead to:

- possible processing errors – an automatic POS tagger might have a systematic bias towards certain interpretations of ambiguous forms as more or less probable, for example *spread* can be more often treated as a noun; and
- uncomfortable design decisions – a POS tagger might define all participial forms as verbal forms, *interested*, *limited*, *measured*, so that *interested* will not be present as an adjectival collocate to *parties*.

If we go further away from forms towards their function, we can, with some degree of reliability, detect the more common participant configurations, for example Actor–Process–Goal for material processes. To test the hypothesis concerning development of grammatical complexity within a text, we can use a POS tagger to measure the distribution of lengths of nominal groups. For example, we can detect them by a pattern such as Rule A:

Rule A: (Adj|Noun)* Noun (of (Det|Adj|Noun)* Noun)*

This states that a noun can be preceded by zero or more adjectives or nouns, and it can be followed by a number of other nominal groups after *of*.

In each case, when we generalise forms to their function, we also need to estimate the accuracy of our generalisation procedure manually on a small sample. The standard measures in this case are **precision** – that is, the proportion of *correctly* identified items out of the items a rule *detected* in a text – and **recall** – that is, the proportion of *correctly* identified items out of all items *present* in a text. Precision is inversely proportional to the amount of incorrect items in the output list: cleaner lists correspond to higher precision. Recall, on the other hand, is inversely proportional to the number of items left in a text unidentified: more complete lists correspond to higher recall.

In a corpus of Wikipedia articles on renewable energy (Sharoff 2012), precision of Rule A reaches 95 per cent, with the errors mostly caused by incorrect POS tagging. For example, in a sentence such as example (2), the underlined fragment is identified as a nominal group because *many* has been tagged as an adjective and *spread*, as a noun.

(2) *there are <u>many fragmented spots of high intensity geothermal potential spread across</u> the continent*

However, recall of Rule A is about 60 per cent – significantly lower than its precision – because, in addition to POS tagging problems, a simple pattern like this fails to identify various other postnominal constructions, for example *single-bladed rotor with a teetering hub*, which includes a preposition, *with*, and a form tagged as a gerund, *teetering*. It is possible to expand

Rule A to cover such patterns too, but this should decrease its precision, since its updated version will also apply to such phrases as *to do an experiment with passing high-voltage electricity through rarefied air*.

Investigation from the lexical ends and from the grammatical end

The wave–particle metaphor also applies to the opposition of lexicon vs grammar. One view on the lexicon is to consider it as more delicate grammar (Hasan 1987). In this view, the grammatical systems specify more general configurations, such as the types of clause, their processes and participants, while the lexical choices are specified in more delicate systems following the grammatical choices. This approach has its own advantages, since it starts with small, well-defined oppositions reflecting grammatical choices.

An SFL study from the lexical end starts from lexical patterns and generalises them into more grammatical phenomena, such as transitivity or modality. Research by John Sinclair and his colleagues in the COBUILD project produced considerable evidence about the possibility of describing lexicogrammar from the lexical end. One example concerns detection of a new grammatical class of 'shell nouns' – that is, nouns that need to be expanded in the surrounding text, such as *belief*, *fact*, *question*, *theory of*, *that* (Hunston and Francis 2000).

Conducting a corpus investigation from the lexical end is considerably easier than one from the grammatical end, since the word forms are immediately accessible to a corpus query. This offers the possibility of creating concordances, frequency lists and collocation lists. Once this information is available, it is possible to classify any regular patterns of use. Even if POS tagging is not entirely reliable and often results in a very coarse model of grammatical functions, basic colligation patterns can be also checked by relatively simple queries. For example, it is easy to use POS patterns to count the types of reporting that follow certain verbs, such as *tell something to someone*, *tell someone that something*, *tell someone to do something*, etc. A statistical corpus-based investigation from the grammatical end is also potentially very fruitful, but its progress is hampered by the lack of large annotated corpora in which the grammatical phenomena have been tagged automatically.

Because the two perspectives are complementary, the interpretations often differ when the lexical and grammatical approaches are applied to the same phenomenon. As mentioned, the approach treating lexis as the most delicate grammar provides a useful interface between the two viewpoints. However, many lexical constructions and collocations cross the grammatical ranks, so it is often convenient to describe lexical constructions without a direct reference to their position in the systemic network for grammar. For example, a category such as 'shell nouns' has its meaning only as a part of a wider collocation, such as *arises/comes/follows/stems from the fact that* or *is the question of how/what/who/whether/where*. This combines elements of systemic choices at the levels of clause configuration and clause complex, as well as at the noun level.

Study of collocations is an example of natural lexical patterns that are more difficult to interpret from a purely grammatical viewpoint. The similarity in the meaning of lexically related expressions belonging to different grammatical categories leads to sharing their lexical context. For example, the constructions *lack of* and *to lack* share a lot of their collocates *complete/total/utter lack of ability/confidence/courage/credibility* vs *completely/totally/utterly lacking ability/confidence/courage/credibility*. If language is viewed as setting expectations, then the more obvious expectations are based on lexical collocates: a sentence starting from *Once upon a . . .* needs to be followed by the word *time* irrespectively of grammatical

choices. Also, while the grammatical phenomena are relatively parsimonious in terms of choices, by their nature the lexical phenomena require many more choices in the networks (Tucker 1998; Sharoff 2006a).

It is natural that the two views can enrich each other. Hunston (2010: 38) provides an example concerning the contexts of use of the expression *the naked eye*. One can expect that it is usually preceded by *to* or *with* and it occurs in the context of such words as *detect*, *look*, *visible*. What is less predictable is that this particular expression more commonly occurs in the context of negation. As mentioned above, the overall frequency of negation in the system of language is predicted to be around 10 per cent. However, the conditional probability of *invisible* in the four-word context of *to the naked eye* is much higher than expected, as illustrated in the top collocates for the 723 instances of *naked eye* from the British English Web Corpus (ukWaC) listed in Table 33.1. Even though *invisible* is nearly six times less frequent than *visible*, it is about 70 per cent as common as *visible* in this context, with many contexts of *visible* also including *barely*, *not*, *rarely*.

Corpus composition

A corpus provides evidence about the uses of linguistic constructions in a finite sample of documents that it contains. However, it is natural to assume that this evidence can be generalised to a wider body of uses. In addition to statistical requirements for such generalisations, which will be mentioned later, we need to make an important assumption that this wider body of uses is similar with respect to its communicative functions to texts from the corpus that was the source of our evidence. To be able to test this assumption, we have to operationalise the notion of communicative functions by providing an exhaustive list of their types, so that a comparison can be made between collections of different texts.

Approaches to classifying a corpus can start from two different ends: we can follow text-external criteria or we can follow text-internal criteria. The text-external approach is based on parameters related to the context of its production, such as its author, the intended aims of the author or perception by the audience. The text-internal approach is based on parameters related to the lexicogrammatical choices made within texts, such as the use of connectives, conditional expressions or nominal phrases. These two views also need different terminology for describing text classes. Even though researchers often use these terms in various incompatible ways, in this chapter we will refer to 'genres' when considering the text-external view and to 'registers' when considering the text-internal one, unless another terminology is preferred by the author being mentioned (cf. Lee 2001; Biber and Conrad 2009; Santini et al. 2010).

Table 33.1 The most significant collocates of *to the naked eye* and their frequencies in ukWaC

Collocate	*In construction*	*Overall*
almost	13	336,594
barely	23	22,116
invisible	212	18,034
just	20	1,747,209
not	72	6,461,794
obvious	13	100,046
undetectable	6	1,157
visible	303	61,649

The wave–particle metaphor is appropriate here again: the two views are directed to the same phenomenon, which naturally combines form and function, but the perspectives are somewhat complementary. If our task is to design a corpus or to compare one corpus with another, we are more likely to use text-external parameters. For example, we can start by postulating the proportion of such genres as Reportage vs Editorials vs Newspaper Reviews (Categories A, B and C, respectively in the Brown corpus). Such categories are reflected in various guidelines for text producers and they are well recognised by the readers, unlike the more linguistically inspired categories such as Narrativity.

However, if our task is to understand the varieties of language use in a corpus, we are more likely to rely on text-internal parameters. For example, we can refer to the choices made in expression of narrative sequences, ways of expressing stance or persuasion and by relating them to the context of their use. Genre analysis of the text-external kind and register analysis of the text-internal kind have attracted considerable attention in SFL research (Bateman 2008; Martin and Rose 2008), as well as in SFL-related corpus studies (Sinclair 2003a).

From the text-external viewpoint, the problem is in designing categories to cover a wide range of parameters describing the variety of texts, which can be considered in a corpus. In his *Introduction to Functional Grammar*, Halliday (1985: xv) defined the aim of his grammatical analysis as being to say 'sensible and useful things about any text, spoken or written, in modern English'. In contrast to this, the existing SFL-based classifications of genres or registers do not cover the entirety of texts in a generic corpus such as the BNC or ukWaC. For example, Martin and Rose (2008) discuss a range of genres centred on different situations of the classroom use. They provide a systemic network that contrasts informing about things (descriptions and reports) vs informing about events (observations, recounts and narratives). Similarly, Bateman (2008) considers a range of visually rich genres, such as newspapers, instruction manuals or field guides, in a multimodal genre space. At the same time, a generic corpus contains many more text types, such as blog entries, annual reports, patents, product reviews, laws, contracts, etc., which also need to be described in some terms. Partly, this is explained by the fact that register analysis is conducted within the stratum of (discourse) semantics, which is not presently covered by an exhaustive list of choices.

Sinclair (1996) responded to the need to describe the composition of large generic corpora by suggesting several text-external dimensions for text classification, including the following six 'intended aims of text production':

- **Information** – reference compendia (Sinclair adds 'an unlikely outcome, because texts are very rarely created merely for this purpose');
- **Discussion** – polemic, position statements, argument;
- **Recommendation** – reports, advice, legal and regulatory documents;
- **Recreation** – fiction and non-fiction (biography, autobiography, etc.);
- **Religion** – holy books, prayer books, orders of service (which label does not refer to religion as a topic); and
- **Instruction** – academic works, textbooks, practical books.[2]

This typology is applicable to a large proportion of texts. However, some classes in this list are too generic: the vast majority of texts are aimed at discussing a state of affairs, so this makes no differentiation between reportage or editorial sections of the newspapers, personal blog entries, political speeches or hotel review forums.

Another important aspect of a text typology – or, indeed, of any annotation scheme – is that it needs to be reliable itself, in the sense that different people performing annotation

on the same set of texts are likely to produce the same result, while any disagreement in their annotations is less than what could have occurred by chance (Krippendorff 2004). In the suggested list of the aims of text production, we can expect considerable disagreement between annotators on what counts as, for example, a recommendation (reports), instruction (academic works) or discussion (argument).

One way of achieving reliable assessment of corpus composition is by positioning texts in a genre space (Lemke 1999), so that the genre of a text is assessed through several test questions, such as the Forsyth and Sharoff's (2014) functional text dimensions (FTDs):

> **A1: Argumentative** To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view?
>
> **A7: Instructive** To what extent does the text aim at teaching the reader how something works?
>
> **A11: Personal** To what extent does the text report from a first-person point of view?
>
> **A17: Evaluative** To what extent does the text evaluate something?

A text receives a score (for example from 0 to 1) for each dimension and more than one dimension can be principal – that is, its value close to 1. If 'a register is a syndrome of lexicogrammatical probabilities' (Halliday 1992: 68), a genre can be considered as a syndrome of functional text dimensions.

Annotation studies show that the level of agreement for such dimensions is greater than for atomic genre labels. Also, some of the commonly used genre classes exhibit considerable internal variation: the generic genre class of blog entries, for example, can have fairly different instances along the Argumentative, Instructive, Personal or Evaluative principal dimensions.

Main research methods

The main methods for corpus research involve corpus collection, annotation and querying. Given the abundance of texts in electronic form and availability of large, ready-made corpora nowadays, the collection of ad hoc corpora for a specific task is reasonably straightforward.

Corpus annotation

As for manual annotation, one of the tools commonly used in SFL research is the UAM CorpusTool,[3] which provides ways of designing an annotation schema in the form of systemic networks, for example for annotating clauses for the process types or for polarity. Annotation can be done on several levels, for example for the document, clause and group levels. The tool can search for examples with a particular combination of features and can run statistical analysis comparing data subsets, for example for contrastive analysis of expressions of appraisal across different genres. In addition to purely manual analysis, the UAM CorpusTool can offer some automation in the form of word lists or rules, for example if a clause contains a word-level feature such as passive, the clause will be preselected to be passive-clause at the clause-level annotation stage.

If the annotation procedure uses an untested annotation schema, it is advisable to estimate its inter-annotator reliability by requesting two or more independent annotations for the same collection of texts. Otherwise, any corpus results concerning the difference between the categories of the scheme, for example description vs report in the framework of Martin

and Rose (2008), cannot be treated as statistically significant. One of the commonly used reliability measures is Krippendorff's α, which estimates the difference between the actual disagreement and the disagreement by chance.[4] The level of $\alpha \geq .667$ indicates an acceptable reliability limit (Krippendorff 2004).

Manual annotation of a large corpus is not feasible. Most often, a basic level of annotation can be achieved by using POS taggers. Some of the commonly used tools for this task are TreeTagger[5] and the Stanford Tagger.[6] TreeTagger is particularly handy, since it comes with a lemmatisation tool and tagging models for a large number of languages. At the same time, the Stanford Tagger can be used together with the Stanford Parser,[7] which can produce basic dependency relations (subject, direct object, prepositional phrases, etc.) for English, as well as for Arabic, Chinese and German. These relations can be used and interpreted by the UAM Corpus Tool in terms of the SFL model to provide a crude first-pass automatic annotation.

Corpus querying

For corpus querying, there is a distinction between stand-alone concordancers, which work directly on user-provided texts, for example AntConc[8] or WordSmith,[9] and client–server interfaces, which usually operate via a web interface with large corpora stored on a server, for example CQPWeb (Evert and Hardie 2011),[10] IntelliText (Wilson et al. 2010)[11] or SketchEngine (Kilgarriff et al. 2004).[12]

In its simple form, making a corpus query is nearly the same as making a search query on the web. However, searching often becomes more challenging because of the need to specify various conditions, such as the possibility to restrict search for lemmas or POS tags, and to indicate possible gaps between words. Many client–server interfaces, including those mentioned above, are based on the Corpus Query Language (CQL), which was developed in the IMS Corpus Workbench.[13] For example, the CQL query to get all examples of the phrasal verb *leave behind* with a possible gap from zero to two words is as follows:

```
[lemma='leave' & pos='V.*'] []{0,2} [lemma='behind']
```

With the output of concordancing tools, the next task is to observe the concordance lines such as those in Table 33.2. There are several good introductions to the procedure (Sinclair 2003b; Hunston 2010). The main task is to detect a strong pattern in data, which could be explained by a hypothesis about meanings associated with the query. Usually, hypothesis formulation is accompanied with several more probing queries to test and generalise the pattern observed in the original data sample. This process can be further extended to multilingual contrastive analysis – that is, when we examine the patterns in two languages for potential translation candidates, often close cognates in related languages, and identify the differences between them. For example, even though *absolutely* in English and *assolutamente* in Italian are fairly good translation equivalents, *assolutamente* is often used in negative contexts, which are likely to be translated into English by using *not entirely* (Partington 1998).

When starting any corpus investigation, it is important to remember that corpus linguistics by itself is a tool; it does not provide a theory for investigation. Irrespectively of how 'corpus-driven' (that is, theory-neutral) an investigation is, observing the instances leads to a theory of the underlying system, which in its turn creates the possibility of interpreting the instances.

Another important consideration is corpus and sample size, because, statistically speaking, language consists of a large number of rare events (Baayen 2008). Mega-word corpora

Table 33.2 Concordance lines for *naked eye* in the BNC

oceans seems devoid of plants – at least	**to the naked eye**	; and if weed grows conspicuously on co
reveal patterns not immediately obvious	**to the naked eye**	; an example of this is shown in exerci
pared to buy houses with flaws invisible	**to the naked eye**	, but now we'd fallen for one wi
elaborate colonies, often quite visible	**to the naked eye**	, in which different individuals perfor
e so faint, not a single one is visible	**to the naked eye**	. Another 15 per cent of stars are oran
les being indistinguishable individually	**to the naked eye**	. Clays, the product of chemical weath
or viruses, although still not visible	**to the naked eye**	. Other species of animal are affected
tars the size of Earth and all invisible	**to the naked eye**	. So the Sun is hardly average. And ne
the internal organs looked pretty normal	**to the naked eye**	. There were some granulomatous area on
ible, and the human egg is just visible	**to the naked eye**	. The three key cell structures are the
are small, they are not always visible	**to the naked eye**	. They are revealed in their millions,
limetre in diameter and are just visible	**to the naked eye**	. Unlike the chick they contain no yolk
estern Europe. The mite is just visible	**to the naked eye**	and feeds on honey bees and their grubs
looks as smooth and featureless as glass	**to the naked eye**	but, unlike glass, it is a crystallin

are needed to capture the linguistic properties of even moderately frequent words such as *malicious*, which has 2 instances in the 1 million words of the Brown corpus, 337 in the 100 million words of the BNC and 6,182 instances in the 2 billion words of ukWaC. Data becomes sparser when we consider the frequencies of collocates. A corpus of the size of ukWaC provides some evidence about the patterns of use of *to the naked eye* (see Table 33.1), but even this evidence is not statistically reliable. Statistically significant evidence is not achievable with the 50 instances of this construction in the BNC.

A related issue concerns evidence obtained through corpus use. Corpora do not provide negative evidence – that is, a confirmation that a linguistic phenomenon does not occur. Negative evidence can be inferred only from a non-significant number of examples if positive evidence on a large corpus is overwhelming. Positive evidence suggests that a phenomenon exists, but more attention is needed to the conditions in which it is used, as well as to a comparison against other contexts of use. The number of positive examples is also not a measure in itself; it becomes meaningful only when compared against another value, for example an alternative construction or a different context of use. For example, there are 325 instances of the lemma *acanthus* in ukWaC, but there is no instance of any of the two potentially possible plural forms: *acanthuses* or *acanthi*. However, their absence does not declare their impossibility in the system. Similarly, the presence of 683 instances of the form *advices* in ukWaC indicates that the plural form is possible, in spite of the fact that the dictionaries

declare it to be a mass noun. However, this use is much rarer than 452,138 uses of the non-count form and, in the majority of cases, the nominal form *advices* was used in texts from the nineteenth century, language teaching materials as a suggestion to avoid the plural form, or in the specific sense of informing documents, such as *remittance advices*.

For smaller corpora, both positive and negative evidence is less likely to be significant: the difference between three instances of one construction vs six instances of another one is not going to be significant. A conservative estimate for the smallest number of instances for statistically significant results is 30 (Upton and Cook 2001), even if, in many interesting linguistic cases, this is not achievable. However, even greater figures can be misleading for the purposes of generalisation, if the result includes 'confounding' variables, for example most of the instances coming from the same source, either the same text or the same author.

Future directions

As mentioned earlier, corpus linguistics does not provide a linguistic theory in itself; its primary contribution is a set of methods, which can help in developing linguistic theories. Therefore future directions in corpus research have primarily indirect impact on SFL studies by providing tools, rather than research hypotheses.

One of the techniques that has been driving research in automatic natural-language processing (NLP) is statistical machine learning (Manning and Schütze 1999). It has already made considerable impact on the way in which the linguistic resources are produced. An incomplete list of success stories includes POS tagging, syntactic parsing, text classification, information retrieval and machine translation. The machine learning approach starts from some amount of data, which is used to produce automatic annotation of sufficient quality. Sometimes, this data contains a desired annotation level, which serves as an example for annotating more data of the same kind. Technically, this is called 'supervised machine learning'. For example, if some sentences have been annotated with POS categories, the probabilities of the POS sequences detected in them can be used to process any new sentences without annotation to resolve the ambiguities such as *spread* as a noun or as a verb (Manning and Schütze 1999). Similarly, a number of examples of *think* in the sense of 'believe' vs *think* in the sense of 'contemplate' encoded as respectively Mental or Behavioural processes can help in determining correlations with features of the surrounding contexts. In the end, this can annotate verbs in a corpus with their more likely reading, for example Mental or behavioural. Sometimes, we do not have texts at a desired annotation level, while we can still use some more basic linguistic features for inferring statistical regularities at the desired annotation level. Technically, this is called 'unsupervised machine learning'. Biber's (1988) multidimensional analysis is an early example of unsupervised machine learning at the text-classification level, for example making a reliable association of surface-level features such as past-tense verbs or first-person pronouns with narration. Another possible application of unsupervised learning concerns automatic detection of the most common patterns of use for the main processes in the clause, which can lead to a data-driven way of describing the process configurations.

Either approach (or their combination in the form of *semi*-supervised machine learning) has a potential to change the situation with annotated corpora in SFL, improving availability of texts annotated at linguistic levels finer than the POS tags and lemmas as traditional in corpus studies. This should help in providing statistically significant evidence for the distribution of choices and their realisation, such as analysis of process–participant configurations, investigation of appraisal or thematic development. For example, if there is a

corpus that has been manually annotated with the types of appraisal group, as well as with their targets and modifiers (Bednarek 2009), and there is a reliable dependency parser, which can link the targets and modifiers automatically in a text (Nivre et al. 2006), then machine learning methods can help in training a classifier that should detect the types of appraisal group in any text. Given that a considerable number of manually annotated resources have been produced within the systemic community since 2000 and given that a large amount of raw text is readily available in electronic form, annotated resources can be utilised to bootstrap systemic research and to provide new kinds of evidence.

Apart from wider availability of texts overall, the web also helps in bringing more *kinds* of texts for linguistic analysis. Among other things, this includes the language of social media, such as Facebook, or traditional news sources enhanced with user-provided comments, as well as user-contributed content on websites such as Tripadvisor or collaborative editing in the form of Wikipedia. Potentially, this provides a far greater amount of data for investigation of linguistic interaction than what was available in the past. However, harnessing this sea of data needs advanced computational methods such as those suggested above.

Another recent development concerns eye-tracking technologies, which are becoming more widely used to link corpus research on the product – that is, the text – to perception research on the process of its interpretation. Eye tracking can detect the amount of time the eye spends on a particular word, as well as possible regressions – that is, cases in which the expected flow of reading is interrupted and the eye returns to a fragment read earlier. The wider availability of eye tracking gives another kind of evidence about what is treated as expected and what is unexpected by the readers, thus reflecting on the marked and unmarked choices.

Notes

1 For an overview, see Thompson and Hunston (2006).
2 A more elaborate exposition of the text-external and text-internal classification criteria is available in Sinclair (1996).
3 http://www.wagsoft.com/CorpusTool/.
4 http://dfreelon.org/utils/recalfront/ is a convenient online interface.
5 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
6 http://nlp.stanford.edu/software/tagger.shtml.
7 http://nlp.stanford.edu/software/stanford-dependencies.shtml.
8 http://www.antlab.sci.waseda.ac.jp/software.html.
9 http://www.lexically.net/wordsmith/.
10 https://cqpweb.lancs.ac.uk/.
11 http://corpus.leeds.ac.uk/it/.
12 http://the.sketchengine.co.uk/.
13 http://cwb.sourceforge.net/documentation.php.

References

Aston, G., and L. Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baayen, H. 2008. *Analysing Linguistic Data, Vol. 505*. Cambridge: Cambridge University Press.

Baroni, M., and S. Bernardini, 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, Lisbon (LREC 2004)*. Paris: ELRA, pp. 1313–16.

Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3): 209–26.

Bateman, J. 2008. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. Basingstoke: Palgrave Macmillan.

Bednarek, M. 2009. Language patterns and ATTITUDE. *Functions of Language* 16(2): 165–92.

Biber, D. 1988. *Variations across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., and S. Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Evert, S., and A. Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. Paper presented at the Corpus Linguistics Conference, University of Birmingham, 20–22 July.

Firth, J.R. 1957 [1951]. Modes of meaning. In J.R Firth (ed.) *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press, pp. 190–215.

Forsyth, R., and S. Sharoff. 2014. Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing* 29: 6–22.

Gregory, M. 1988. Generic situation and register: A functional view of communication. In J.D. Benson, M. Cummings and W.S. Greaves (eds) *Linguistics in a Systemic Perspective*. Amsterdam: John Benjamins, pp. 301–30.

Halliday, M.A.K. 1957. *Some Aspects of Systematic Description and Comparison in Grammatical Analysis* (Studies in Linguistic Analysis). Oxford: Blackwell.

Halliday, M.A.K. 1959. *The Language of the Chinese: Secret History of the Mongols*. Oxford: Blackwell.

Halliday, M.A.K. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Oxford: Blackwell.

Halliday, M.A.K. 1985. *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M.A.K. 1992. Language as system and language as instance: The corpus as a theoretical construct. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82* 65: 61–77.

Hasan, R. 1987. The grammarian's dream: Lexis as most delicate grammar. In M.A.K. Halliday and R.P. Fawcett (eds) *New Developments in Systemic Linguistics*. London: Pinter, pp. 184–211.

Hunston, S. 2010. How can a corpus be used to explore patterns. In A. O'Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp. 152–66.

Hunston, S., and G. Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell. 2004. The sketch engine. In *Proceedings of Euralex 2004*, Lorient: Université de Bretagne-Sud, pp. 105–16.

Krippendorff, K. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research* 30(3): 411–33.

Lee, D. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3): 37–72.

Lemke, J.L. 1999. Typology, topology, topography: genre semantics. Unpublished manuscript. University of Michigan, Ann Arbor, MI.

Manning, C., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Martin, J., and D. Rose. 2008. *Genre Relations: Mapping Culture*. London: Equinox.

Matthiessen, C.M.I.M. 2006. Frequency profiles of some basic grammatical systems: An interim report. In G. Thompson and S. Hunston (eds) *System and Corpus: Exploring Connections*. London: Equinox, pp. 103–42.

Nesi, H., and S. Gardner. 2012. *Genres across the Disciplines: Student Writing in Higher Education*. Cambridge: Cambridge University Press.

Nivre, J., J. Hall and J. Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, Lisbon (LREC 2004)*. Paris: ELRA, pp. 2216–19.

Partington, A. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Santini, M., A. Mehler and S. Sharoff. 2010. Riding the rough waves of genre on the web. In A. Mehler, S. Sharoff and M. Santini (eds) *Genres on the Web: Computational Models and Empirical Studies*. Berlin and New York: Springer, pp. 3–30.

Sharoff, S. 2006a. How to handle lexical semantics in SFL: A corpus study of purposes for using size adjectives. In G. Thompson and S. Hunston (eds) *System and Corpus: Exploring Connections*. London: Equinox, pp. 184–205.

Sharoff, S. 2006b. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics* 11(4): 435–62.

Sharoff, S. 2012. Beyond translation memories: Finding similar documents in comparable corpora. In *Proceedings of the Translating and the Computer Conference*, London.

Sinclair, J. 1991. *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1996. *Preliminary Recommendations on Text Typology*. Expert Advisory Group on Language Engineering Standards (EAGLES) Technical Report EAG-TCWG-TTYP/P, June.

Sinclair, J. 2003a. Corpora for lexicography. In P.V. Sterkenberg (ed.) *A Practical Guide to Lexicography*. Amsterdam: John Benjamins, pp. 167–78.

Sinclair, J. 2003b. *Reading Concordances: An Introduction*. Harlow: Longman.

Thompson, G., and S. Hunston (eds). 2006. *System and Corpus: Exploring Connections*. London: Equinox.

Tucker, G. 1998. *The Lexicogrammar of Adjectives: A Systemic Functional Approach to Lexis*. London: Cassell Academic.

Upton, G., and I. Cook. 2001. *Introducing Statistics*. 2nd edn. Oxford: Oxford University Press.

Wilson, J., A. Hartley, S. Sharoff and P. Stephenson. 2010. Advanced corpus solutions for humanities researchers. In *Proceedings of the Workshop on Advanced Corpus Solutions (PACLIC 24)*. Sendai: Tohoku University, pp. 36–43.

34

Translation studies

Kerstin Kunz and Elke Teich

Introduction

Translation studies, or translatology, is the discipline concerned with the theory, modelling and description of translation. Commonly, translation is considered from two (complementary) perspectives (cf. Bell 1991): translation as *product* and translation as *process*. Thus a comprehensive theory of translation will have to account for and explain both of these aspects: the human activity of translation involves interpretation – of a source language (SL) text – and production – of a target language (TL) text. For a linguistic theory to be useful for theorising translation, it needs to be able to describe and model linguistic products – that is, texts in a translation relation – and explain how they are processed (by human or machine). Essentially, then, a linguistic theory suitable for application to translation must be concerned with *language use*. Systemic functional linguistics (SFL) is such a theory. The conception of language use being dependent on context is embodied in SFL in the concept of *register* and SFL provides methods for modelling the context–language relation, as well as for describing single registers in terms of configurations of lexicogrammatical patterns (see Bowcher, this volume; Moore, this volume).

This chapter provides an overview of (a) how translatology has harnessed SFL as a framework for theorising and modelling translation, and (b) the various ways in which SFL has engaged in translation research. In the next section, it provides a brief history of SFL's involvement in translation studies, mainly discussing Catford's seminal work and Halliday's expositions on translation. It goes on to look at SFL's notion of register as a suitable basis for theorising and modelling translation. The chapter then discusses recent research on translation applying SFL for different research objectives such as register variation and regarding grammatical metaphor, cohesion and appraisal. A brief overview of the methods employed in the analysis of translation(s) follows and then we briefly describe the current practices in SFL-based translation research, with a focus on corpus-based studies. Finally, we briefly sketch possible future directions in SFL-based translation research.