Sinclair, J./Hanks, P./Fox, G./Moon, R./Stock, P. (eds.) (1987), *Collins COBUILD English Language Dictionary.* London: Collins.

Teubert, W./Čermáková, A. (2004), Directions in Corpus Linguistics. In: Halliday, M. A. K./Teubert, W./Yallop, C./Čermáková, A. (eds.), *Lexicology and Corpus Linguistics*. London: Continuum, 113−166.

Trier, J. (1931), *Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes. Bd. 1: Von den Anfängen bis zum Beginn des 13. Jahrhunderts.* Heidelberg: Winter. Reprinted in Lee, A. van der/Reichmann, O. (eds.) (1973), *Aussätze und Vorträge zur Wortfeldtheorie*. The Hague: Mouton, 40−65.

Vendler, Z. (1967), Verbs and Times. In: Vendler, Z. (ed.), *Linguistics in Philosophy.* Ithaca, NY: Cornell University Press, 97−121.

Michael Hoey, Liverpool (UK)

46. Theory-driven corpus research: Using corpora to inform aspect theory

1. Introduction

The theory-driven versus data-driven distinction in linguistics is a manifestation of the conflict between rationalism and empiricism in philosophy. The extremist views of these two approaches to linguistics are vividly illustrated by Fillmore's (1992) cartoon figures of the armchair linguist and the corpus linguist. The armchair linguist thinks that what the corpus linguist is doing is uninteresting while the corpus linguist believes that what the armchair linguist is doing is untrue. It is hardly surprising that the divorce of theory and empirical data results in either untrue or uninteresting theories because any theory that cannot account for authentic data is a false theory while data without a theory is just a meaningless pile of data. As such, with exceptions of a few extremists from either camp who argue that "Corpus linguistics doesn't mean anything" (see Andor 2004, 97), or that nothing meaningful can be done without a corpus (see Murison-Bowie 1996, 182), the majority of linguists (e. g. Leech 1992; Meyer 2002) are aware that the two approaches are complementary to each other. In Fillmore's (1992, 35) words, "the two kinds of linguists need each other. Or better, […] the two kinds of linguists, wherever possible, should exist in the same body".

This article discusses the use of corpus data in developing linguistic theory (section 2) and presents an effort to achieve a marriage between theory-driven and corpus-based approaches to linguistics via a series of case studies of aspect (section 3), which has long been studied, but rarely with recourse to corpus data.

2. Can corpora contribute to linguistic theory?

To answer this question, we must first of all find out what linguistics is about. We will then discuss the use of intuitions and corpora as evidence in linguistic theorizing and explore how corpus data can contribute to linguistic theory.

2.1. What linguistics is about

It has been argued that linguistics is "the study of abstract systems of knowledge idealized out of language as actually experienced", i. e. "idealized internalized I-language" (Widdowson 2000, 6). If linguistics is defined as such, we must admit that any linguistic analysis involving performance data (i. e. "E-language") has nothing to do with "linguistics" and should claim no place in "linguistics" at all (cf. Leech 2000, 685). The assumption underlying Widdowson's definition is Chomsky's (1965; 1986) claim that competence can be separated from performance to be studied alone. But can the two be separated?

The competence vs. performance divide is rooted in the hypothesis that grammar is autonomous within the human mind. Generative grammarians argue that our use of language (performance, E-language) cannot reflect our internal knowledge of language (competence, I-language), because of the constraints in naturally occurring language. Performance errors have been likened to abnormal conditions like tiredness and drunkenness in human communication (e. g. Radford 1997, 2). Only the internal grammar, which is based on native intuitions and not polluted by performance constraints, is said to be part of competence. The corollary of this argument is the sharp distinction between langue and parole (Saussure 1916 [1966]), between performance and competence (Chomsky 1965), and between grammar and usage (Newmeyer 2003). Nevertheless, this dichotomy is arguably over-stated. Evidence from recent research in psycholinguistics, neurolinguistics, and biology shows that the hypothesis of an autonomous grammar, which underlies the sharp distinction between competence and performance, is unsustainable (see Shei 2004 for a review). Rather, grammar is constantly shaped by culture (or environmental factors) and interpersonal interactions. In de Beaugrande's (1997, 302) words, "performance can crucially determine the development and quality of competence". On the other hand, performance does not spring from nowhere − it is a natural and actual product of competence. Clearly, as Leech (1992, 108) observes, "the putative gulf between competence and performance has been overemphasized".

Given the nature of this interdependence, the Chomskyan linguists' practice of forcing a sharp distinction between competence and performance is simply misleading in that it is in essence merely an "idealization of language for the sake of simplicity" (Abney 1996, 11). In doing so, real language is replaced by idealized language which does not exist but which purports to sustain an explanation of language (cf. de Beaugrande 1997) while attested language data is denied a place in theory building. In the dialectic view of the relationship between competence and performance, therefore, the assertion is simply unsustainable that performance "cannot constitute the subject-matter of linguistics" (Chomsky 1965, 20), because competence is not directly accessible and our only gateway to it is through performance (cf. Meyer/Nelson 2006). Linguistics is in fact concerned with what language really is − as reflected by our knowledge, as well as use, of language. Just as Kennedy (1998, 270) argues:

"Furthermore, description of the system we use is not the only legitimate goal of the study of language. The linguistic system is both derived from and instantiated by specific instances of use. It is thus perfectly legitimate to describe language both in terms of the system we use and our use of this system and for the description thus to encompass language as possibility as well as probability of use".

2.2. Intuitions and corpus data in theory building

Intuitions and corpus data are two important types of evidence in linguistic theory. Linguistic intuitions can be used in introspection to invent (grammatical, ungrammatical, or questionable) example sentences for linguistic analysis, or make judgments about the acceptability/grammaticality or meaning of an expression. They are always useful in linguistics as the linguist can invent purer examples instantly for analysis. This is so because intuitions are readily available and invented examples are free from language-external influences existing in naturally occurring language. Intuitions are even in a sense indispensable in linguistic theorizing because categorization, which usually involves intuitive judgments, is unavoidable in theory building (see section 3). Nevertheless, intuitions should be applied with caution (cf. Seuren 1998, 260−262). Firstly, it is possible to be influenced by one's dialect or sociolect (cf. also Krishnamurthy 2000b, 172). Consequently, what appears grammatically unacceptable to one speaker may be perfectly felicitous to another (cf. also Wasow/Arnold 2005, 1482; see Schütze 1996 for further discussion of grammaticality judgments). Secondly, when one invents an example to support or disprove an argument, one is consciously monitoring one's language production. Therefore, even if one's intuitions are correct, the example may not represent typical language use. Thirdly, introspective data is decontextualized because it exists in the analyst's mind rather than in any real linguistic context. Context is particularly relevant to acceptability and grammaticality judgments. With proper contexts, what might appear ungrammatical or unacceptable out of context can become grammatical and acceptable while "our imagination is powerful (so we can conceive of possible contexts for the most implausible utterances)" (Krishnamurthy 2000a, 32−33). Fourthly, results based on intuitions alone are difficult to verify as introspection is not observable. Fifthly, excessive reliance on intuitions blinds the analyst to the realities of language usage (cf. Meyer/Nelson 2006). For example, we tend to notice the unusual but overlook the commonplace because of the psychological salience of rare words and usages (Sinclair 1997, 33; Krishnamurthy 2000b, 170−171). Finally, there are areas in linguistics where intuitions cannot be used reliably but must rely upon corpus data, e. g. language variation, historical linguistics, register and style, first and second language acquisition (Meyer 2002; Léon 2005, 36).

With that said, we must hasten to add that we do not mean that linguistic intuitions are useless and should be abandoned in theory building as corpus-driven linguists have advocated (see section 2.3.). Absolutely not. We use our intuitions in the first place to decide what to examine and then to interpret what we find (cf. Krishnamurthy 2000b, 172). Without intuitions, it is also impossible to make judgments of the acceptability or meaning of an expression, or to categorize data − even though contextual clues are sometimes available. Our point is simply that intuition as a type of linguistic evidence should be used in conjunction with empirical evidence collected from other sources rather than being relied upon solely as the basis of linguistic theory. However, there has been an overreliance upon introspective evidence in theoretical linguistics (cf. also Gast 2006). One of the symptoms is that, as Adger (2002, no page) notes, "[s]ome of the data of core linguistics is actually generated by the theory, in that experimental and intuition-based data are collected to test theory, and the collection techniques are designed with this aim in mind". That explains why such data can be "biased towards the point that is to be proved, i. e. linguists may see what they want to see" (Johansson 1995; cited in Krishnamurthy 2000b, 170). It is simply a vicious circle to develop a linguistic hypothesis on the basis of an analyst's introspective data, which is used in turn to verify the same hypothesis. As such, Labov (1972, 199) argued that "linguists cannot continue to produce theory and data at the same time".

Another symptom of the overreliance is that intuitions are given such a privileged status in generative linguistics that "evidence other than intuitions is brought in only as supporting evidence" (Wasow/Arnold 2005, 1484) whilst "contradictory evidence from other sources [...] is simply ignored" (Wasow/Arnold 2005, 1486). Given that it "would simply be a waste of time and energy" to "devise operational and experimental procedures" (Chomsky 1969, 81) while usage data cannot seriously "constitute the actual subject matter of linguistics" (Chomsky 1965, 4), the overreliance in the generative tradition upon intuitions is quite unsurprising. Nevertheless, as "All of us, even native speakers or 'expert speakers', have only a partial knowledge of that language" (Krishnamurthy 2000b, 172), a linguistic theory will become more reliable and convincing if linguists care to check whether their intuitions are "in accord with what people actually say and write" (Wasow/Arnold 2005, 1486).

In contrast with introspective data which relies solely on one's own intuitions, a corpus pools together the linguistic intuitions of a range of speakers and/or writers. Corpora comprise samples of spoken/written language which has already occurred naturally in real linguistic context. As people speak and write on the basis of their intuitions in real contexts, corpus data is also intuition-based; but it is more natural than introspective data because it is not created specifically for linguistic analysis. In relation to data collected through introspection of an individual, corpus data typically reflects the intuitions of a much greater number of language users. Corpora can also provide frequencies readily, which cannot be predicated by intuition reliably (cf. McEnery/Wilson 2001, 15). As such, corpus data allows a linguist to avoid any potential bias in his/her own intuitions and distinguish what is statistically central and typical from what is statistically marginal in theory development. In short, a corpus typically provides data that is attested, contextualized and quantitative. It can also find differences that intuitions alone cannot perceive (cf. Francis/Hunston/Manning 1996; Kennedy 1998, 272). In addition, corpora have opened up or foregrounded a number of new areas of linguistic research that would not have been possible on the basis of intuitions alone, most notably register and variation studies (see section 3.3. for an example; see also article 38 in this volume).

While the corpus-based approach has won widespread popularity and has been used in nearly all branches of linguistics (see McEnery/Xiao/Tono 2006 for an overview), corpora have also become the target of a number of criticisms. For example, Chomsky (1957) has argued that a corpus only contains a finite number of sentences while language is "an infinite set of sentences" (though see de Beaugrande (2002, 105) for a counter-argument for this definition of language; and see article 2 in this volume for discussion of early generative linguistics). Since a corpus does not include each and every possible sentence of language, corpus data is by nature "skewed" – "[s]ome sentences won't occur because they are obvious, others because they are false, still others because they are impolite" (Chomsky 1962, 159, cited in Leech 1991, 8). These criticisms are certainly valid, especially when they were made in the 1950s. Corpora, especially those used in what McEnery/Wilson (2001) call "early corpus linguistics", were ready targets of such criticisms because of their small sizes and inadequate sampling. Chomsky can indeed be considered as the person who has helped to "shape the approach taken by the corpus today" (McEnery/Wilson 2001, 19), because these criticisms have led to such key concepts as balance, representativeness and sampling in corpus linguistics which, coupled with developments in technology, and especially the development of ever more powerful computers offering ever increasing processing power and massive storage at relatively low cost, have made corpora of today as large and balanced as practically possible to be maximally representative of the language or language variety under consideration (see article 9 in this volume). While it might be true that a 100-million word balanced corpus is still skewed to some extent, it is certainly less skewed than a dataset obtained through introspection on the basis of one analyst's intuitions. Intuitions can be skewed because "the process of introspection may not be systematic" (McEnery/Wilson 2001, 15) and because intuitions are discriminating – "[m]atters of wit, curiosity and love of the unusual, the absurd, etc., have a further impact on the intuition" (Sinclair 1997, 33). Corpora have been criticized for being skewed simply because they are observable and open to scrutiny whereas intuitions are not.

Chomsky (1965) argued against corpora also because corpus data is likely to contain performance errors which have nothing to do with one's knowledge of language. This criticism is true, but it is reasonable to assume that a corpus is generally composed of sentences which are grammatical (cf. also McEnery/Wilson 2001, 16). Corpus data at least provides evidence of what speakers believe to be grammatically acceptable in their language. Intuitions are not error-free either, though, not to mention the bias as noted earlier. Labov (1975) has shown that one's intuitions of grammaticality may not necessarily be a true reflection of one's internal grammar. Furthermore, as a corpus presents data in context, it allows for research into what types of performance errors occur under what conditions and are typically associated with what contexts. Theories of this type cannot be developed on the basis of decontextualized introspective data but they are of practical importance in linguistics. In our view, therefore, a "performance grammar" (Chomsky 1962, 537–538) that copes with regular and irregular language phenomena (including performance errors) is of greater importance than a "competence grammar" that has little bearing on "everyday production or comprehension of language" (Schütze 1996, xi).

However, while the corpus-based approach has some advantages over the intuition-based approach, it also has some known weaknesses. Firstly, as a corpus cannot possibly include all sentences in a language, sampling is unavoidable and the representativeness of the corpus becomes an issue. Nowadays, representativeness is still regarded as an "act of faith" (Leech 1991, 127) for lack of a reliable scientific measure of corpus balance, though the confidence about a corpus can be increased when the corpus increases to a respectable size and achieves a wide coverage (cf. article 9 in this volume). Secondly, statistical methods that are more sophisticated and rigorous are required to interpret corpus data. Quantitative analysis is equally important as qualitative analysis in corpus research. Many statistical measures which are commonly used in corpus linguistics assume that linguistic features are evenly distributed in language − in different corpora or in different samples in a corpus − which may not be the case. Hence, we support Gries's (2006) argument for "more rigorous corpus linguistics" (see article 36 for further discussion of statistical methods in corpus exploration). Thirdly, a corpus does not provide negative evidence (but see Stefanowitsch 2006 for a counter-argument). A corpus, however large and balanced it is, cannot be exhaustive except for highly specialized cases (e. g. the corpus of the Bible mentioned in article 9), because natural language is infinite. As such, corpora cannot tell us what is possible or not possible in language. If a construction does not exist in a corpus, you cannot say that it does not exist in the language (but according to Stefanowitsch 2006, it is possible to tell what is "significantly absent" from what is "accidentally absent" on the basis of a properly annotated corpus); neither can you say that a construction found in a corpus is necessarily grammatically acceptable because a corpus may contain "performance errors". Nevertheless, everything included in a corpus is what language users have actually produced − for good or ill. The emphasis of corpus research is on "the repetitive and routine nature of language use" (Stubbs 2001a, 152), though hapax legomena can also be of interest in some studies (e. g. article 41). Corpus data is useful in showing what is statistically central and typical in language. If a "performance error" is repeated sufficiently often by a sufficiently large group of native language users, the "error" might as well be approached from the perspective of language variation or language change, while in the case of learner corpora it is precisely repetitive patterns of such performance errors that make data of this kind useful in interlanguage studies. Finally, while the corpus-based approach is good at yielding interesting findings, it cannot explain what we find in corpora. The explanations must be developed using other methodologies and evidence from other sources, including intuitions.

In spite of the philosophical tension between theoretical linguists and corpus linguists, the intuition-based approach and corpus-based approach are not necessarily antagonistic. But rather the two approaches corroborate each other and can be "gainfully viewed as being complementary" (McEnery/Wilson 2001, 19). Given that both introspective data and corpus data have their own weaknesses as noted earlier, it is our view that the theory-driven and data-driven approaches to linguistics should be combined to take advantage of their strengths while circumventing their weaknesses. Broadly speaking, compared with the more traditional intuition-based approach, which rejected or ignored corpus data, the corpus-based approach can achieve improved reliability because it does not go to the extreme of rejecting intuitions while attaching importance to empirical data. The key to using corpus data is to find the balance between the use of corpus data and the use of one's intuitions. As Leech (1991, 14) observes:

> "Neither the corpus linguist of the 1950s, who rejected intuitions, nor the general linguist of the 1960s, who rejected corpus data, was able to achieve the interaction of data coverage and the insight that characterize the many successful corpus analyses of recent years".

Unsurprisingly, a number of areas in modern linguistics have relied upon a fusion of corpus evidence and intuitions, ranging from the more practical aspects such as sociolinguistic studies (see article 6), language teaching (see article 7) and lexicography (see article 8), to more theory-driven research including syntax (see article 42) and grammar (see article 43).

Theory-driven corpus research has so far been confined largely to the distribution of forms rather than semantic aspects of language (cf. Kennedy 1998, 272). While meanings related to forms (e. g. semantic prosody, semantic preference, and pattern meaning) have also become a focus of recent corpus research (e. g. Louw 1993, 2000; Stubbs 1995; Partington 1998; 2004; Xiao/McEnery 2006a), core semantic notions such as aspect have rarely been approached from a corpus-based perspective. This is probably because the study of aspectual meaning involves much greater use of intuitions than lexical and grammatical studies and thus has been approached traditionally without recourse to corpus data. However, as we will see in the case studies presented in section 3, the theory-driven and data-driven approaches can be fruitfully combined in the development of aspect theory. But before we present the case studies, it is appropriate to discuss how corpora can contribute to linguistics.

2.3. Corpus-based versus corpus-driven linguistics

Whether corpora should be used at all in linguistics is one issue, and how corpora should be used is another. Having established that corpus data can indeed contribute to linguistic theory, this section discusses how corpora are used to achieve this goal. Even among those who advocate the use of corpus data, there are different opinions and different approaches. One further area where differences diverge in corpus linguistics is with regard to the question of corpus-based and corpus-driven approaches. While, as we will see shortly, the distinction between the two is overstated, what underlies the proposed distinction is highly relevant to the discussion of the present article − how pre-corpus theoretical premises and intuitions should be incorporated in corpus research. In a nutshell, corpus-driven linguists aim to build theory "from scratch" − claiming that they are completely free from pre-corpus theoretical premises − and base their theories exclusively on corpus data, assuming that "all the relevant information is contained in the corpus itself, and the linguist's task is to *extract* that information and make it visible" (Gast 2006, 114), whilst corpus-based linguists tend to approach corpus data "from the perspective of moderate 'corpus-external' premises" (ibid.) with the aim of testing and improving such theories.

In the corpus-based approach, it is said that corpora are used mainly to "expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini Bonelli 2001, 65). Corpus-based linguists are accused of not being fully and strictly committed to corpus data as a whole as they have been said to discard inconvenient evidence (i. e. data not fitting the pre-corpus theory) by "insulation", "standardization" and "instantiation", typically by means of annotating a corpus. In contrast, corpus-driven linguists are said to be strictly committed to "the integrity of the data as a whole" (ibid., 84) and therefore, in this latter approach, it is claimed that "[t]he theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus" (ibid., 85). Upon interrogating the available evidence, nevertheless, it is found that the proposed sharp distinction between the corpus-based vs. corpus-driven approaches is overstated and that this "'radically empiricist' way of doing corpus research" (Gast 2006, 114) is an idealized extreme. There are four basic differences between the corpus-based vs. corpus-driven approaches: types of corpora used, attitudes towards existing theories and intuitions, focuses of research, and paradigmatic claims. Let us discuss each in turn.

Regarding the type of corpus data used, there are three issues − representativeness, corpus size and annotation. Let us consider these one by one. According to corpus-driven linguists, there is no need to make any serious effort to achieve corpus balance and representativeness because the corpus is said to balance itself when it grows to be big enough, as the corpus achieves so-called cumulative representativeness. This initial assumption of self-balancing via cumulative representativeness, nonetheless, is arguably unwarranted. For example, one such cumulatively representative corpus is a corpus of Zimbabwean English that Louw (1991) used in his contrastive study of collocations of in British English and Zimbabwean English. This study shows that the collocates of *wash* and *washing*, etc. in British English are *machine*, *powder* and *spin* whereas in Zimbabwean English the more likely collocates are *women*, *river*, *earth* and *stone*. The different collocational behavior was attributed to the fact that the Zimbabwean corpus has a prominent element of literary texts such as Charles Mungoshi's novel *Waiting for the Rain*, "where women washing in the river are a recurrent theme across the novel" (Tognini Bonelli 2001, 88). One could therefore reasonably argue that this so-called cumulatively balanced corpus was skewed. Especially where whole texts are included, a practice corpus-driven linguists advocate, it is nearly unavoidable that a small number of texts may seriously affect, either by theme or in style, the balance of a corpus. Findings on the basis of such cumulatively representative corpora may not be generalizable beyond the corpora themselves as their balance is easily affected by the availability of electronic text of different types.

The corpus-driven approach also argues for very large corpora. While it is true that the corpora used by corpus-driven linguists are very large (for example, the Bank of English has grown to 524 million words), size is not all-important, as Leech (1991, 8−29) and McCarthy/Carter (2001) note (but see Krishnamurthy 2000b for a counter-argument). Another problem for the corpus-driven approach relates to frequency. While it has been claimed that in the corpus-driven approach corpus evidence is exploited fully, in reality frequency may be used as a filter to allow the analyst to exclude some data from their analysis. For example, a researcher may set the minimum frequency of occurrence for a pattern which it must reach before it merits attention, e. g. it must occur at least twice − in separate documents (Tognini Bonelli 2001, 89). Even with such a filter, a corpus-driven grammar would consist of thousands of patterns which would bewilder the learner. It is presumably to avoid such bewilderment that the patterns reported in the *Grammar Patterns* series (Francis/Hunston/Manning 1996, 1998), which are considered as the first results of the corpus-driven approach, are not even that exhaustive. Indeed, faced with the great number of concordances, corpus-driven linguists are often found to analyze only the nth occurrence from a total of X instances. This is in reality currently the most practical way of exploring a very large corpus which is unannotated. Yet if a large corpus is reduced to a small dataset in this way, there is little advantage in using very large corpora and it can hardly be claimed that corpus data is exploited fully and the integrity of the data is respected. It appears, then, that the corpus-driven approach is not so different from the corpus-based approach − while the latter allegedly insulates theory from data or standardizes data to fit theory, the former filters the data via apparently scientific random sampling, though there is no guarantee that the corpus is not explored selectively to avoid inconvenient evidence.

The corpus-driven linguists have strong objections to corpus annotation. This is closely associated with the second difference between the two approaches − different attitudes towards existing theories and intuitions. It is claimed that the corpus-driven linguists come to a corpus with no preconceived theory, with the aim of postulating linguistic categories entirely on the basis of corpus data, though corpus-driven linguists do concede that pre-corpus theories are insights cumulated over centuries which should not be discarded readily and that intuitions are essential in analyzing data. This claim is a little surprising, as traditional categories such as nouns, verbs, prepositions, subjects, objects, clauses, and passives are not uncommon in so-called corpus-driven studies. When these terms occur they are used without a definition and are accepted as given. Also, linguistic intuitions typically come as a result of accumulated education in preconceived theory. So applying intuitions when classifying concordances may simply be an implicit annotation process, which unconsciously makes use of preconceived theory. As implicit annotation is not open to scrutiny, it is to all intents and purposes unrecoverable and thus more unreliable than explicit annotation. Like the purely rationalist approach to linguistics that rejects corpus data, corpus-driven linguists take a radically empiricist approach that aims to reject everything outside a corpus in spite of the known weaknesses of corpus data (see section 2.2.). In contrast, corpus-based linguists do not have such a hostile attitude towards existing theory. The corpus-based approach typically has existing theory as a starting point and corrects and revises such theory in the light of corpus evidence. As part of this process, corpus annotation is common. Annotating a corpus, most notably part-of-speech tagging, inevitably involves developing a tagset on the basis of an existing theory, which is then tested and revised constantly to mirror the attested language use. In spite of the usefulness of corpus annotation as a result, which greatly facilitates corpus exploration, annotation as a process is also important. As Aarts (2002, 122) observes, as part of the annotation process the task of the linguist becomes "to examine where the annotation fits the data and where it does not, and to make changes in the description and annotation scheme where it does not". The claimed independence of preconception on the part of corpus-driven linguists is clearly an overstatement. A truly corpus-driven approach, if defined in this way, would require something such as someone who has never received any education related to language use and therefore is free from preconceived theory, for as Sampson (2001, 135) observes, schooling plays an important role in forming one's intuitions. Given that preconceived theory is difficult to totally reject and dismiss, and intuitions are indeed called upon in corpus-driven linguistics, we cannot see any real difference between the corpus-driven demand to re-examine pre-corpus theories in the new framework and corpus-based linguists' practice of testing and revising such theories. Furthermore, if the so-called proven corpus-driven categories in corpus-driven linguistics, which are supposed to be already fully consistent with and directly reflect corpus evidence, also need refinement in the light of different corpus data, the original corpus data is arguably not representative enough. The endless refinement will result in inconsistent language descriptions which will place an unwelcome burden on the learner. In this sense, the corpus-driven approach is no better than the corpus-based approach.

The third important difference between the corpus-driven and corpus-based approaches is their different research foci. As the corpus-driven approach makes no distinction between lexis, syntax, pragmatics, semantics and discourse (because all of these are pre-corpus concepts and they combine to create meaning), the holistic approach provides, unsurprisingly, only one level of language description, namely, that of functionally complete units of meaning or language patterning. In studying patterning, corpus-driven linguists concede that while collocation can be easily identified in KWIC concordances of unannotated data, colligation is less obvious unless a corpus is grammatically tagged. Yet a tagged corpus is the last thing the corpus-driven linguists should turn to, as grammatical tagging is based on preconceived theory, and consequently results in a loss of information, in their view. To overcome this problem, Firth's definition of colligation is often applied in a loose sense − in spite of the claim that corpus-driven linguistics is deeply rooted in Firth's work − because studying colligation in Firth's original sense necessitates a tagged or even a parsed corpus. According to Firth (1968, 181), colligation refers to the relations between words at the grammatical level, i. e. the relations of "word and sentence classes or of similar categories" instead of "between words as such". But nowadays the term *colligation* has been used to refer not only to significant co-occurrence of a word with grammatical classes or categories (e. g. Hoey 1997, 2000; Stubbs 2001b, 112) but also to significant co-occurrence of a word with grammatical words (e. g. Krishnamurthy 2000a). The patterning with grammatical words, of course, can be observed and computed even using a raw corpus.

A final contrast one can note between corpus-based and corpus-driven approaches is that the corpus-based approach is not as ambitious as the corpus-driven approach. The corpus-driven approach claims to be a paradigm within which a whole language can be described. No such claim is entailed in the corpus-based approach. Yet the corpus-based approach, as a methodology which makes use of corpus data and intuitions, has been applied in nearly all branches of linguistics.

The discussion in this section shows that the sharp distinction forced between the corpus-based vs. corpus-driven approaches to linguistics is in reality fuzzy. If the purely intuition-based rationalist approach discussed in the previous section and the radically empiricist corpus-driven approach characterized in Tognini Bonelli (2001) are viewed as the two ends of a scalar rationalism-empiricism continuum (i. e. "the armchair linguist" and "the corpus linguist" in Fillmore's (1992) humourous account), it can be said that just as the former goes to one extreme by rejecting corpus data, the latter goes to the other extreme by rejecting everything outside a corpus. The corpus-based approach lies in between the extremes, seeking to strike a balance between the use of corpora and the use of intuitions. As both intuitions and corpus data have known weaknesses which can be avoided when the two types of complementary and corroborating evidence are taken into account (see section 2.2.), the corpus-based approach is arguably more reliable than the extremist methods that reject either corpora or intuitions.

Having established that corpus data can contribute to linguistic theory and that in so doing the corpus-based approach is more appropriate, we will present a case study of aspect in the remainder of the article, which seeks to achieve a marriage between theory-driven and corpus-based approaches to linguistics.

3. Using corpora to inform aspect theory

Aspect is a linguistic phenomenon that is related to the temporal properties of linguistically described situations in the world (situation aspect) and how these situations are presented (viewpoint aspect). Situation aspect is composed of inherent features whereas viewpoint aspect is composed of non-inherent features of aspect. The two components of aspect interplay to determine the aspectual meaning of an utterance (cf. Smith 1997). As the temporal notion denoted by aspect is essential to human languages, aspect has long been the subject of intensive studies by both semanticists and grammarians. However, while corpora have been used extensively in a wide range of areas in linguistics, research on aspect has rarely used corpus data. Yet corpora have a role to play both in developing and testing such theories.

With a few exceptions, most studies on aspect published to date have been based on a handful of examples invented through introspection (e. g. Verkuyl 1993; Smith 1997; Klein/Li/Hendriks 2000), some of which are, if not intuitively unacceptable, unnatural and atypical of attested language use (see Xiao/McEnery 2004a; 2004b for further discussion). Furthermore, those proposals have not been tested with corpus data, which can serve as a test-bed for the linguistic theory proposed as well as for the intuitions on which the theory is based (cf. section 2.2.). This section reports on the corpus-based research in aspect which we have recently undertaken.

3.1. Situation aspect: A corpus-based two-level model

Situation aspect is concerned with the aspectual classification of verbs and situations according to their temporal features such as dynamicity, durativity and telicity. While the earliest literature on aspectual classification dates as far back as Aristotle, modern approaches to aspect are normally considered to start with Vendler (1967), who classified verbs into four classes: state, activity, accomplishment, and achievement, as shown in Table 46.1.

Tab. 46.1: Vendler’s four verb classes

Classes	[±dynamic]	[±durative]	[±telic]	Examples
State	–	+	–	know, love, believe, possess
Activity	+	+	–	run, walk, swim, push a cart
Accomplishment	+	+	+	run a mile, walk to school, paint a picture
Achievement	+	–	+	recognize, spot, find, lose, reach, win

As can be seen in the table, Vendler’s analysis basically works at the lexical level (cf. Verkuyl 1993, 33), though it also involves predicates rather than simply verbs alone. As such, Vendler has to put *run* and *walk* under the category of activity and put *run a mile* and *walk to school* under the category of accomplishment. Ever since Vendler (1967), a number of theories have been proposed to account for the compositional nature of situation aspect. The most important models include Verkuyl (1993) and Smith (1997). However, all of the models are deeply flawed. For example, Verkuyl incorrectly argues that durativity is linguistically irrelevant and that external arguments also contribute to situation aspect, while Smith’s model only works at the sentential level.

Xiao/McEnery (2004a) developed a two-level model of situation aspect on the basis of an investigation of the English and Chinese languages using a fusion of native speaker intuitions and evidence from corpora. The new model of situation aspect consists of three components: a lexicon, a layered clause structure and a set of rules mapping verb classes onto situation types.

In this new theory, situation aspect is modeled as verb classes at the lexical level and as situation types at the sentential level. Using a newly established five-way classification system, situation aspect is classified into six verb classes at the lexical level, namely, individual-level state (ILS), stage-level state (SLS), activity, semelfactive, accomplishment, and achievement (see Table 46.2). The verb classes at the lexical level constitute the lexicon of our model.

Tab. 46.2: Feature matrix system of verb classes

Classes	[±dynamic]	[±durative]	[±bounded]	[±telic]	[±result]
Activity	+	+	−	−	−
Semelfactive	+	−	±	−	−
Accomplishment	+	+	+	+	−
Achievement	+	−	+	+	+
ILS	−	+	−	−	−
SLS	±	+	−	−	−

At the sentential level, situation aspect is classified into the same six basic situation types and five derived situation types. Situation types are the composite result of the rule-based interaction between verb classes and complements, arguments, peripheral adjuncts and viewpoint aspect at three layers of the clause structure: the nucleus, core, and clause levels.

Our two-level approach to modeling situation aspect was motivated by the deficiencies of Vendler (1967) and Smith (1997). The Vendlerian approach works well at the lexical level, but not at the sentential level. Conversely the approach of Smith (1997) works well at the sentential level but not at the lexical level. The two-level approach to situation aspect has sought to bridge this gap, operating at both lexical and sentential levels. While the two-level approach to modeling situation aspect has given a better account of the compositional nature of situation aspect by proposing a set of rules mapping verb classes at the lexical level onto situation types at the sentential level, it has also provided a more refined classification of situation aspect, most notably by distinguishing between two types of states. As the new model of aspect is based on and verified by corpus data from English and Chinese, it is more explanatory of attested language usages in the two distinctly unrelated languages. Indeed, as Xiao/McEnery (2002) observe, situation aspect is language independent. Our two-level model of situation aspect represents an extension of Smith's (1997) two-component aspect theory.

Our model of situation aspect has drawn evidence from both corpus data and intuitions. In theory-driven corpus research of this kind, both types of evidence are indispensable because they interact with each other in theory building. On the one hand, classifying verbs at the lexical level and situations at the clause level on the basis of semantic features is a task that is virtually impossible without recourse to one's intuitions, because the feature values such as telicity cannot be determined reliably using linguistic co-occurrence tests (see Xiao/McEnery 2006b for a discussion of using completive and durative temporal adverbials such as *in/for an hour* as telicity tests). The indispensable role of intuitions in theorization as demonstrated in this example shows that such a purely empiricist approach as taken by corpus-driven linguists, that rejects everything outside a corpus, is merely wishful thinking (cf. section 2.2.). On the other hand, corpora are not only a valuable resource that helps to test old hypotheses and formulate new ones, they are also a touchstone for our intuitions.

3.2. Aspect in Mandarin Chinese: A corpus-based model

Mandarin Chinese as an aspect language has played an important role in the development of aspect theory. Nearly all of the major works on aspect theory make reference to Chinese (e. g. Comrie 1976; Smith 1997). Nevertheless, while a few aspect markers have been studied intensively in Chinese linguistics for decades, little attention has been paid to date to the question of systematically describing the linguistic devices that the language employs to express aspectual meanings. Still less attention has been paid to the inherent temporality of situations denoted by utterances in Chinese. But aspect markers that signal different perspectives from which a situation can be presented are only one component of aspect. Worse still, there has been no generally agreed account even of the three most frequently studied aspect markers -*le*, -*zhe*, and -*guo*. For example:

Should the verb-final -*le* be distinguished from the sentence-final *le* and the modal particle *le*? Does the verb-final -*le* indicate the termination or completion of a situation? Does the verb-final -*le* interact with stative and atelic situations? Should the sentence-final *le* be covered in a study of aspect in Chinese? If so, what is its aspectual meaning? Is it necessary to distinguish between the experiential -*guo* and the RVC (resulative verb complement) *guo*? Does the imperfective -*zhe* indicate resultativeness or durativeness? How should the interchange between -*le* and *guo* be accounted for? Under what conditions is the perfective -*le* interchangeable with the imperfective -*zhe*? While intuitions are essential for answers to questions like these, proposals based on introspective evidence alone cannot account for the complexities existing in authentic language data.

Xiao/McEnery (2004b) presents a corpus-based study of aspect in Chinese, which demonstrates how corpora and linguistic theory can interact. All of the above issues have been addressed in this book. More importantly, the book explores aspect at both the semantic and grammatical levels. The two levels correspond to the two components of aspect, namely, situation aspect and viewpoint aspect. Situation aspect operates at the semantic level while viewpoint aspect operates at the grammatical level, but the two also interact with each other, thus explaining why some aspect markers are incompatible with some situation types while other aspect markers show a preference for other situation types. The corpus-based model of aspect in Chinese represents a systematic and structured exploration of linguistic devices which Chinese employs to express aspectual meanings. In addition to situation aspect, which is inherent in linguistic expressions of situations in human languages, this book has identified, on the basis of corpus data, four perfective viewpoints (the actual aspect marked by -*le*, the experiential aspect marked by -*guo*, the delimitative aspect marked by verb reduplication, and the completive aspect marked by RVCs) and four imperfective ones (the durative aspect marked by -*zhe*, the progressive aspect marked by *zai*, the inceptive aspect marked by -*qilai*, and the continuative aspect marked by -*xiaqu*) in Chinese, and has discussed the characteristic features of each of them in exhaustive detail on the basis of their behavior in attested language use. Barring the three most studied aspect markers mentioned earlier, the aspectual values of the others have been overlooked in most research to date. For example, while RVCs were found in this book to be the most productive perfective markers indicating the completiveness of a situation, their aspectual meanings have rarely been discussed elsewhere. While cursory discussions of some of these markers can be found scattered around a number of studies, they have mostly been misunderstood. Kang (1999, 223−243), for example, correctly treats -*qilai* as an aspect marker, yet she conflates its resultative and completive meanings together with its inceptive meaning. The current work has overcome these problems and defined the meaning and form of each aspect marker, thus giving a consistent account of viewpoint aspect in Mandarin Chinese. In addition, the book has corrected many intuition-based misconceptions and associated misleading conclusions readily found in the literature (see below).

Of particular importance is that the model of Chinese aspect focuses on the interaction between situation aspect and viewpoint aspect, which can only be explored reliably using a corpus-based approach because of the gradient nature of this interaction. For example, quite contrary to many intuition-based proposals in the literature (e. g. Pan 1993; Smith 1997; Li 1999), the perfective -*le* in Chinese is not sensitive to the features [±dynamic] or [±telic]. Rather, as our corpus data shows, -*le* can interact with all situation types in Chinese but it strongly prefers spatially or temporally bounded situations, which account for about 90 percent of the situations co-occurring with -*le*. With unbounded states, -*le* demonstrates the feature of ingressive dynamicity and coerces these situations into derived activities at the clause level. As a perfective marker, -*le* only indicates the actualization and focuses on the entirety of a situation but does not provide any final endpoint as the English simple aspect does. The interaction between the progressive *zai* and achievements is also not as simple as has traditionally been assumed. It has been asserted (e. g. Smith 1997; Yang 1995; Kang 1999) that achievements never occur with the progressive marker *zai*. Nevertheless, our corpus data shows that achievements of different types demonstrate different degrees of compatibility with the progressive aspect in Chinese. While simplex achievements and complex achievements with completive RVCs are strictly incompatible with the progressive, those with result-state and directional RVCs show some tolerance to the progressive aspect. These examples not only demonstrate that corpus data can correct biased intuitions, they also show that quantitative data readily available from corpora "can decide issues that less empirically minded researchers could debate endlessly without ever reaching a conclusion" (Stefanowitsch 2006, 98).

Xiao/McEnery (2004b) has sought to achieve a marriage between theory-driven and corpus-based approaches to linguistics through a study of aspect in Chinese. The use of corpus data as an input to the semantic analysis of aspect represents something new. Previous approaches to the semantics of aspect have rarely used corpus data. Yet the marriage of the corpus-based approach and traditional intuition-based semantic analysis has enabled this book to produce a more realistic account of situation aspect and viewpoint aspect in Chinese in a way that has not been attempted previously. As such, we believe that the book is a powerful demonstration of the way in which corpus data may lead to more accurate linguistic descriptions and hence theories.

3.3. Aspect marking: Contrastive and translation studies

The corpus-based aspect model established in Xiao/McEnery (2004b), which was first developed in Xiao (2002), also demonstrates its value as a unified language-independent framework not only for analyzing a single language, but also for contrasting two or more languages and explaining shifts of situation and viewpoint aspect which often occur in translations, hence helping us to explore the process of translation.

McEnery/Xiao/Mo (2003), for example, used the aspect model to contrast aspect marking in Chinese, British and American English on the basis of three comparable corpora, namely, the Lancaster Corpus of Mandarin Chinese (LCMC) and its matches for British and American English FLOB and Frown (see article 20). The study shows that while Chinese and English are distinctly different, aspect markers in the two languages show a strikingly similar distribution pattern, especially across the two broad categories of narrative and expository texts, as shown in Figure 46.1. In both LCMC and FLOB/Frown, the text categories where the frequency of aspect markers is above the average are the five fiction categories (text categories L, M, N, P, and K) plus humor (R), biography (G), and press reportage (A). The text categories where aspect markers occur least frequently include reports/official documents (H), academic prose (J), skills/trades/hobbies (E), press reviews (C), press editorials (B), religion (D), and popular lore (F). In both Chinese and British/American English, there is a great difference in usage between the first and second groups of texts, which indicates that the two are basically different. Text types like fiction, humor, and biography are narrative whereas reports/official documents, academic prose, and skills/trades/hobbies are expository. Press reportage appears to be a transitory category which is more akin to narrative texts. Statistic tests show that in both Chinese and the two varieties of English, the differences between the distribution of aspect markers in narrative and expository texts are significant. According to Hopper (1979), among many others, the discourse functions of aspect

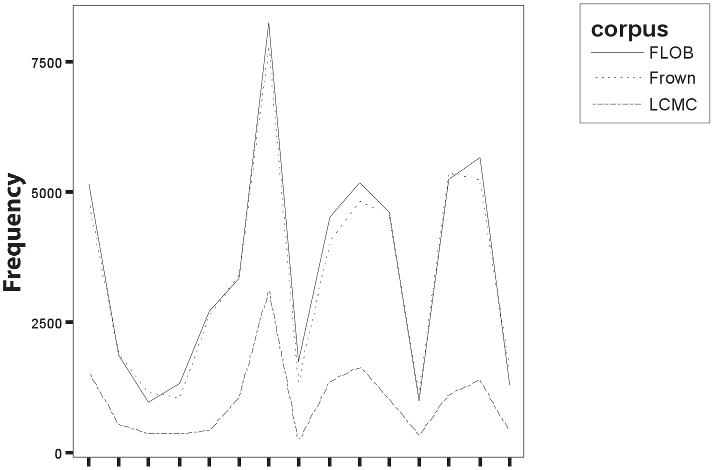

Fig. 46.1: Distribution of aspect markers (frequency)

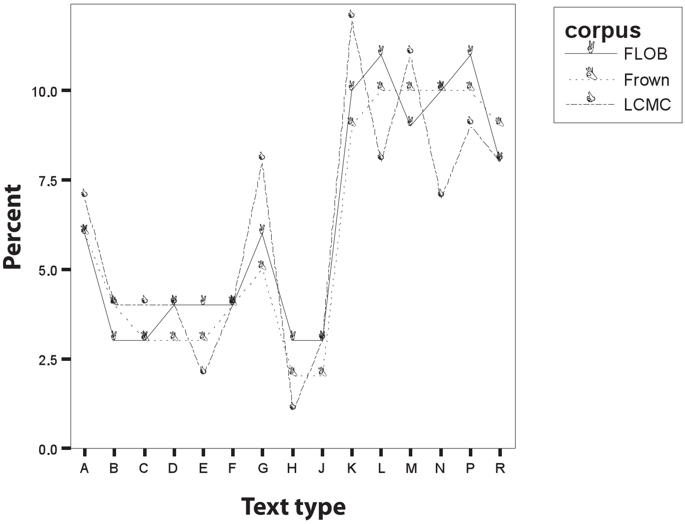

Fig. 46.2: Distribution of aspect markers (percentage)

marking are theoretically linked to foregrounding/backgrounding in narration. Consequently, it is hardly surprising to find that cross-linguistically, aspect markers are significantly more common in the "active, event-oriented discourse" than the "more static, descriptive or expository types of discourse" (Biber 1988, 109). Note that variation studies of this kind are important in linguistic theorizing because variation is inherent in language (cf. Meyer 2002, 3). As Elliot/Legum/Thompson (1969, 52) observed, there are facts about linguistic theory "whose existence will be obscured unless variation is taken into account". While intuitions are particularly useful in a semantic domain such as aspect, they are of little use in language variation research, which relies heavily upon corpus data (cf. section 2.2.).

The contrastive study also reveals some important differences in the distribution of aspect markers in Chinese vs. English and British vs. American English across fifteen text categories, as shown in Figure 46.2. The figure shows the distribution of aspect markers (expressed as percentages) across the fifteen text categories in the three corpora. As can be seen, by comparison to the two major varieties of English, aspect markers in Chinese occur more frequently in text categories G (biography) and K (general fiction) but less frequently in N (adventure fiction), L (mystery fiction), H (official documents), and E (hobby/skill/trade). British English and American English also differ in that the latter variety does not show such a marked fluctuation in aspect marking in narrative texts, notably in biography and the five types of fiction.

Further analysis of the corpus data reveals that in expository texts, perfective aspect markers in Chinese generally occur more frequently than those in English whereas in narrative texts, perfective markers in English are generally more frequent than those in Chinese. The relatively high frequency of perfective markers in narrative texts and their lower frequency in expository texts in English can be accounted for by the fact that aspect markers in English express both temporal and aspectual meanings because the aspect and tense markers in English combine morphologically. For example, over 80 percent of perfective markers in FLOB and Frown are simple past forms. Given that narrative texts usually relate to what happened in the past whilst expository texts are typically non-past, the relatively high frequency of perfective markers in narrative as opposed to expository texts in English is hardly surprising.

In marked contrast, imperfective aspect markers show a totally different distribution pattern from perfective markers. In expository texts, imperfective markers in both varieties of English typically occur more frequently than those in Chinese, whereas imperfective markers in Chinese are generally more frequent than those in English. This phenomenon can be explained as follows. First, the Chinese progressive marked by *zai* can only signal progressiveness literally. In contrast, "the progressive in English has a number of other specific uses that do not seem to fit under the general definition of progressiveness" (Comrie 1976, 37). Although the different uses of the progressive in English and Chinese account for the slightly higher frequency of the English imperfective markers in expository texts, this cannot explain the relatively low frequency of these markers in narrative texts. Nevertheless, we can find an answer in the Chinese imperfective marker -*zhe*, which accounts for 88 percent of the imperfective markers in the Chinese corpus. This marker has three basic functions: to signal the durative nature of a situation, to serve with a verb as an adverbial modifier to provide background information, and to occur in locative inversion to indicate existential status (Xiao/McEnery 2004b). Of the three functions of -*zhe*, only the first is used in expository texts. Hence, in spite of the high overall frequency of -*zhe* in LCMC, only about 20 percent of all examples of -*zhe* occur in expository texts. In contrast, all of the three functions of -*zhe* apply to narrative texts. Furthermore, in addition to inducing a background effect, -*zhe* can also be used in an apparently "foregrounded" situation to move narration forward (see Xiao/McEnery 2004b). As such, it is only natural to find that Chinese imperfective markers occur more frequently in narrative texts than English imperfective markers.

Using the same analytic framework, McEnery/Xiao (2002) and Xiao/McEnery (2005) explore, on the basis of aligned parallel corpora, how aspectual meanings in English are translated into Chinese. It is found that in English-Chinese translation, most progressives in English (over 58 percent) do not undergo a shift in viewpoint aspect, though some of the translations (about 15 percent) may take the unmarked form. Whether a viewpoint aspect shift occurs in translation depends largely on the specific use of the progressive in the English source data, and on the interaction between situation aspect and viewpoint aspect in the Chinese target language. This means that on the one hand, when progressives in the English source data that indicate habitual situations or anticipated happenings are translated into Chinese, they necessarily undergo a viewpoint aspect shift, because the progressive in Chinese does not indicate habituality or futurity. On the other hand, when a translation triggers a situation type shift into individual-level states (ILSs) or achievements in the Chinese translations, a viewpoint aspect shift is expected, because these two types of situations do not normally take the prototypical progressive.

When English perfect constructions are translated into Chinese, they more often than not depend on context to indicate the perfect meaning. This is because Chinese does not have a grammatical aspect marker for the perfect. In this case, however, aspect markers such as -*le* and -*guo* could be used to mark the perfect meaning. Whether the translations take overt aspect markers or imply the perfect meaning contextually depends largely on the type of perfect, i. e. the perfect of result, the perfect of experience, the perfect of recent past, and the perfect of persistent situation (see Comrie 1976) in the English source texts.

The perfect progressive is an interaction between the perfect and the progressive. Chinese translations of the perfect progressive may shift towards the progressive or the perfect meaning, depending on the situation type involved and the translator's choice of viewpoint. But in most cases both perfect and progressive meanings can be retained, with the perfect being lexicalized by temporal adverbs and the progressive being signalled by the progressive aspect marker *zai* or implied by the context. The pluperfect progressive is similar to the perfect progressive with the exception that it signals progressiveness with a relatively past time reference. Situations referred to by the English pluperfect can be translated into Chinese with the progressive or the durative aspect unless the translator chooses to present them perfectively or there is a shift in situation type which prohibits them taking the progressive or the durative aspect.

Situations marked by the English simple aspect are mainly presented perfectively and most of them take the covert form in Chinese translations. The high frequency of perfectives in translations of the simple aspect can be accounted for by the fact that the simple forms in English, the simple past in particular, are basically perfective in nature (cf. Brinton 1988). Translations of the simple past show a marked/unmarked ratio twice as high as that in translations of the simple present. A natural explanation for this contrast is that the simple present typically denotes states, which do not have to be marked aspectually. Translations of the simple future frequently take modal auxiliaries or adverbs that lexicalize future time references. This is because modal and future meanings are closely related (cf. Comrie 1976).

In conclusion, the research on situation aspect, on aspect in Chinese, and the contrastive and translation studies reported in this section have demonstrated that corpus data can indeed be used to inform aspect theory. Our corpus-based aspect model has not only provided an explanatorily adequate account of aspect in Chinese, it is also a useful framework for contrastive language study and translation research. Methodologically, the case studies presented in this section show that intuitions and corpora are complementary rather than antagonistic. The two types of data must complement each other so as to circumvent their weaknesses if as broad a range of research questions as possible are to be addressed by linguists (cf. section 2.2.).

4. Conclusions

This article explored theory-driven corpus research, as exemplified by the case studies in section 3 as well as articles 42 and 43 in this volume. The discussion shows that if linguistics is defined as the study of language as reflected by our knowledge as well as use of language − which it should be − instead of as the study of "idealized language", corpus data can indeed contribute to linguistic theory, because corpora can provide attested, contextualized and quantitative language data. As noted in section 2.2., intuitions and corpora have their own strengths and weaknesses, and the two are not mutually exclusive. Different research questions require different kinds of data. The key is to find the balance between the use of corpus data and the use of intuitions to suit the needs of the research question under consideration. It is also clear from the discussion that the sharp distinction between the "corpus-based" and "corpus-driven" approaches is in reality overstated and that the theory-free corpus-driven linguistics is at best an idealized extreme. The different approaches may be more appropriate in different areas of studies (consider, for example, the roles played by intuitions and corpora in collocation and aspect studies; see article 58 for discussion of collocations), but corpora are what they are. They can be used to verify and revise existing linguistic theories, and they can also be used to provide what intuitions alone cannot discern, on the basis of which entirely new linguistic theories can be developed. The corpus-based research of aspect reported in this article demonstrates that the theory-driven and data-driven approaches can, and indeed should, complement each other in linguistic analysis to make linguistic theory true and interesting at the same time. As Fillmore (1992) expects, the best practice is for the armchair linguist and the corpus linguist to exist in the same body.

5. Literature

Aarts, J. (2002), Review of *Corpus Linguistics at Work*. In: *International Journal of Corpus Linguistics* 7(1), 118−123.

Abney, S. (1996), Statistical Methods and Linguistics. In: Klavans, J./Resnik, P. (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: MIT Press, 1−26.

Adger, D. (2002), Why Theory is Essential: The Relationship between Theory, Analysis and Data. In: Gallagher-Brett, A./Dickens, A./Canning, J. (eds.), *Guide to Good Practice for Learning and Teaching in Languages, Linguistics and Area Studies*. Southampton: The British Higher Education Academy Subject Centre for Languages, Linguistics and Area Studies. Available at: http://www.lang.ltsn.ac.uk/resources/goodpractice.aspx?resourceid=405.

Andor, J. (2004), The Master and his Performance: An Interview with Noam Chomsky. In: *Intercultural Pragmatics* 1(1), 93−111.

de Beaugrande, R. (1997), Theory and Practice in Applied Linguistics: Disconnection, Conflict, or Dialectic? In: *Applied Linguistics* 18(3), 279−313.

de Beaugrande, R. (2002), Descriptive Linguistics at the Millennium: Corpus Data as Authentic Language. In: *Journal of Language and Linguistics* 1(2), 91−131.

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

Brinton, L. (1988), *The Development of English Aspectual System*. Cambridge: Cambridge University Press.

Chomsky, N. (1957), *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1962), Explanatory Models in Linguistics. In: Nagel, E./Suppes, P./Tarski, A. (eds.), *Logic, Methodology, and Philosophy of Science*. Stanford: Stanford University Press, 528−550.

Chomsky N. (1965), *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1969), Language and Philosophy. In: Hook, S. (ed.), *Language and Philosophy: A Symposium*. New York: New York University Press, 51−94.

Chomsky, N. (1986), *Knowledge of Language*. New York: Praeger.

Comrie, B. (1976), *Aspect*. Cambridge: Cambridge University Press.

Elliot, D./Legum, S./Thompson, S. (1969), Syntactic Variation as Linguistic Data. In: Binnick, R./Davison, A./Green, G./Morgan, J. (eds.), *Papers from the Fifth Regiongal Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 52−59.

Fillmore, C. (1992), "Corpus Linguistics" or "Computer-aided Armchair Linguistics". In: Svartvik, J. (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4−8 August 1991*. Berlin/New York: Mouton de Gruyter, 35−60.

Firth, J. (1968), A Synopsis of Linguistic Theory. In: Palmer, F. (ed.), *Selected Papers of J.R. Firth 1952−59*. London: Longman, 168−205.

Francis, G./Hunston, S./Manning, E. (1996), *Collins Cobuild Grammar Patterns 1: Verbs*. London: HarperCollins.

Francis, G./Hunston, S./Manning, E. (1998), *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

Gast, V. (2006), Introduction. In: *Zeitschrift für Anglistik und Amerikanistik. Special Issue on the Scope and Limits of Corpus Linguistics* 54(2), 13−20.

Gries, S. (2006), Some Proposals towards More Rigorous Corpus Linguistics. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), 191−202.

Hoey, M. (1997), From Concordance to Text Structure: New Uses for Computer Corpora. In: Lewandowska-Tomaszczyk, B./Melia, J. (eds.), *PALC '97*: *Proceedings of Practical Applications in Linguistic Corpora Conference*. Łódź: University of Łódź, 2−23.

Hoey, M. (2000), A World beyond Collocation: New Perspectives on Vocabulary Teaching. In: Lewis, M. (ed.), *Teaching Collocations*. Hove: Language Teaching Publications, 224−243.

Hopper, P. (1979), Aspect and Foregrounding in Discourse. In: Givon, T. (ed.), *Syntax and Semantics* (Volume 12) − *Discourse and Syntax*. New York: Academic Press, 213−241.

Johansson, S. (1995), Mens sana in corpore sano: On the Role of Corpora in Linguistic Research. In: *ESSE Messenger* IV(2), 19−25.

Kang, J. (1999), The Composition of the Perfective Aspect in Mandarin Chinese. PhD thesis, Boston University.

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*. London: Longman.

Klein, W./Li, P./Hendriks, H. (2000), Aspect and Assertion in Mandarin Chinese. In: *Natural Language and Linguistics Theory* 18, 723−770.

Krishnamurthy, R. (2000a), Collocation: From *silly ass* to Lexical Sets. In: Heffer, C./Sauntson, H./Fox, G. (eds.), *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham, 31−47.

Krishnamurthy, R. (2000b), Size Matters: Creating Dictionaries from the World's Largest Corpus. In: *Proceedings of KOTESOL 2000 − Casting the Net: Diversity in Language Learning*. Taegu, Korea, 169−180.

Labov, W. (1972), *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1975), *What is a Linguistic Fact?* Lisse: Peter de Ridder Press.

Leech, G. (1991), The State of the Art in Corpus Linguistics. In: Aijmer, K./Altenberg, B. (eds.), *English Corpus Linguistics*. London: Longman, 8−29.

Leech, G. (1992), Corpora and Theories of Linguistic Performance. In: Svartvik, J. (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4−8 August 1991*. Berlin: Mouton de Gruyter, 105−122.

Leech, G. (2000), Grammars of Spoken English: New Outcomes of Corpus-oriented Research. In: *Language Learning* 50(4), 675−724.

Léon, J. (2005), Claimed and Unclaimed Sources of *Corpus Linguistics*. In: *Henry Sweet Society Bulletin* 44, 36−50.

Li, M. (1999), Negation in Chinese. PhD thesis, University of Manchester.

Louw, B. (1991), Classroom Concordancing of Delexical Forms and the Case for Integrating Language and Literature. In: Johns, T./King, P. (eds.), *Classroom Concordancing*. *ELR Journal* 4. Birmingham: CELS University of Birmingham, 151−178.

Louw, B. (1993), Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In: Baker, M./Francis, G./Tognini Bonelli, E. (eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 157−176.

Louw, B. (2000), Contextual Prosodic Theory: Bringing Semantic Prosodies to Life. In: Heffer, C./Sauntson, H./Fox, G. (eds.), *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham, 48−94.

McCarthy, M./Carter, R. (2001), Size isn't Everything: Spoken English, Corpus and the Classroom. In: *TESOL Quarterly* 35(2), 337−340.

McEnery, A./Wilson, A. (2001), *Corpus Linguistics* (1st ed. 1996). Edinburgh: Edinburgh University Press.

McEnery, A./Xiao, R. (2002), Domains, Text Types, Aspect Marking and English-Chinese Translation. In: *Languages in Contrast* 2(2), 51−69.

McEnery, A./Xiao, R./Mo, L. (2003), Aspect Marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for Contrastive Language Study. In: *Literary and Linguistic Computing* 18(4), 361−378.

McEnery, A./Xiao, R./Tono, Y. (2006), *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Meyer, C. (2002), *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Meyer, C./Nelson, G. (2006), Data Collection. In: Aarts, B./McMahon, A. (eds.), *The Handbook of English Linguistics*. Oxford: Blackwell, 93−113.

Murison-Bowie, S. (1996), Linguistic Corpora and Language Teaching. In: *Annual Review of Applied Linguistics* 16, 182−199.

Newmeyer, F. (2003), Grammar is Grammar and Usage is Usage. In: *Language* 79(4), 682−707.

Pan, H. (1993), Interaction between Adverbial Quantification and Perfective Aspect. In: Stvan, L./Ryberg, S./Olsen, M. B./Macfarland, T./DiDesidero, L./Bertram, A./Adams, L. (eds.), *Proceedings of the Third Annual Formal Linguistics Society of Mid-America Conference*. Northwestern University, Bloomington, IN: Indiana University Linguistics Club Publications, 188−204.

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

Partington, A. (2004), "Utterly content in each other's company": Semantic Prosody and Semantic Preference. In: *International Journal of Corpus Linguistics* 9(1), 131−156.

Radford, A. (1997), *Syntax: A Minimalist Introduction*. Cambridge: Cambridge University Press.

Sampson, G. (2001), *Empirical Linguistics*. London: Continuum.

Saussure, F. (1916 [1966]), *Course in General Linguistics*. New York: McGraw-Hill.

Schütze, C. (1996), *The Empirical Base of Linguistics*. Chicago: University of Chicago Press.

Seuren, P. (1998), *Western Linguistics: A Historical Introduction*. Oxford: Blackwell.

Shei, C. (2004), Corpus and Grammar: What it isn't. In: *Concentric: Studies in Linguistics* 30(1), 1−18.

Sinclair, J. (1997), Corpus Evidence in Language Description. In: Wichmann, A./Fligelstone, S./McEnery, T./Knowles, G. (eds.), *Teaching and Language Corpora*. London: Longman, 27−39.

Smith, C. (1997), *The Parameter of Aspect* (1st ed. 1991). Dordrecht: Kluwer.

Stefanowitsch, A. (2006), Negative Evidence and the Raw Frequency Fallacy. In: *Corpus Linguistics and Linguistic Theory* 2(1), 61−77.

Stubbs, M. (1995), Collocation and Semantic Profiles: On the Cause of the Trouble with Quantitative Methods. In: *Function of Language* 2(1), 23−55.

Stubbs, M. (2001a), Texts, Corpora, and Problems of Interpretation: A Response to Widdowson. In: *Applied Linguistics* 22(2), 149−172.

Stubbs, M. (2001b), *Words and Phrases*. Oxford: Blackwell.

Tognini Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Vendler, Z. (1967), *Linguistics in Philosophy*. New York: Cornell University Press.

Verkuyl, H. (1993), *A Theory of Aspectuality*. Cambridge: Cambridge University Press.

Wasow, T./Arnold, J. (2005), Intuitions in Linguistic Argumentation. In: *Lingua* 115, 1481−1496.

Widdowson, H. (2000), The Limitations Of Linguistics Applied. In: *Applied Linguistics* 21(1), 3−25.

Xiao, R. (2002), A Corpus-based Study of Aspect in Mandarin Chinese. PhD thesis, Lancaster University.

Xiao, R./McEnery, A. (2002), Situation Aspect as a Universal Aspect: Implications for Artificial Languages. In: *Journal of Universal Language* 3(2), 139−177.

Xiao, R./McEnery, A. (2004a), A Corpus-based Two-level Model of Situation Aspect. In: *Journal of Linguistics* 40(2), 325−363.

Xiao, R./McEnery, A. (2004b), *Aspect in Mandarin Chinese: A Corpus-based Study*. Amsterdam: John Benjamins.

Xiao, R./McEnery, A. (2005), A Corpus-based Approach to Tense and Aspect in English-Chinese Translation. In: Pan, W./Fu, H./Luo, X./Chase, M./Walls, J. (eds.), *Translation and Contrastive Studies*. Shanghai: Shanghai Foreign Language Education Press, 114−157.

Xiao, R./McEnery, A. (2006a), Collocation, Semantic Prosody and Near Synonymy: A Cross-linguistic Perspective. In: *Applied Linguistics* 27(1), 103−129.

Xiao, R./McEnery, A. (2006b), Can Completive and Durative Adverbials Function as Tests for Telicity? Evidence from English and Chinese. In: *Corpus Linguistics and Linguistic Theory* 2(1), 1−21.

Yang, S. (1995), The Aspectual System of Chinese. PhD thesis, University of Victoria.

Richard Xiao, Ormskirk (UK)

47. Corpora and spoken language

1. Introduction: Evolution of spoken corpora

Spoken corpora have evolved over the last four decades from early attempts at corpus-building for the purposes of better understanding such phenomena as first-language acquisition, social variation and conversational structure, to the large, general spoken corpora of today, which have found applications in a variety of contexts from speech recognition, lexicography, sociolinguistics and first and second language acquisition. In this article we focus on spoken corpora and their applications in linguistics and applied linguistics, rather than on ‘speech corpora’, which are typically collected for the purposes of improving technology, a distinction discussed at greater length by Wichmann in article 11; see also article 30.

Some of the earliest spoken corpora were developed within the field of child language acquisition, an example of which was the child-language word-frequency analyses described in Beier/Starkweather/Miller (1967). Another example, which included informal spoken language by adults, as well as by selected age groups of children from six years upwards in a corpus of some 84,000 words, is described in Carterette/Jones (1974). A

Xiao, R./McEnery, A. (2004a), A Corpus-based Two-level Model of Situation Aspect. In: *Journal of Linguistics* 40(2), 325−363.

Xiao, R./McEnery, A. (2004b), *Aspect in Mandarin Chinese: A Corpus-based Study.* Amsterdam: John Benjamins.

Xiao, R./McEnery, A. (2005), A Corpus-based Approach to Tense and Aspect in English-Chinese Translation. In: Pan, W./Fu, H./Luo, X./Chase, M./Walls, J. (eds.), *Translation and Contrastive Studies*. Shanghai: Shanghai Foreign Language Education Press, 114−157.

Xiao, R./McEnery, A. (2006a), Collocation, Semantic Prosody and Near Synonymy: A Cross-linguistic Perspective. In: *Applied Linguistics* 27(1), 103−129.

Xiao, R./McEnery, A. (2006b), Can Completive and Durative Adverbials Function as Tests for Telicity? Evidence from English and Chinese. In: *Corpus Linguistics and Linguistic Theory* 2(1), 1−21.

Yang, S. (1995), The Aspectual System of Chinese. PhD thesis, University of Victoria.

Richard Xiao, Ormskirk (UK)

47. Corpora and spoken language

1. Introduction: Evolution of spoken corpora

Spoken corpora have evolved over the last four decades from early attempts at corpus-building for the purposes of better understanding such phenomena as first-language acquisition, social variation and conversational structure, to the large, general spoken corpora of today, which have found applications in a variety of contexts from speech recognition, lexicography, sociolinguistics and first and second language acquisition. In this article we focus on spoken corpora and their applications in linguistics and applied linguistics, rather than on 'speech corpora', which are typically collected for the purposes of improving technology, a distinction discussed at greater length by Wichmann in article 11; see also article 30.

Some of the earliest spoken corpora were developed within the field of child language acquisition, an example of which was the child-language word-frequency analyses described in Beier/Starkweather/Miller (1967). Another example, which included informal spoken language by adults, as well as by selected age groups of children from six years upwards in a corpus of some 84,000 words, is described in Carterette/Jones (1974). A