Cuadernos
*del Instituto Historia
de la Lengua*

Aproximación a la Lingüística de Corpus y su contribución en la elaboración de diccionarios

Cristina Martín Herrero
Centro de Investigaciones Lingüísticas (Universidad de Salamanca)

Breve CV: Dr. D. Wolfgang Teubert es Catedrático de Lingüística de Corpus en la Universidad de Birmingham (Reino Unido). Fundador de la revista *International Journal of Corpus Linguistics*, es co-autor, junto con Anna Cermáková, del libro *Corpus Linguistics: a Short Introduction*, publicado en 2006 por Continuum, y también editor, junto a Ramesh Krishnamurthy, de la obra en seis volúmenes *Corpus Linguistics. Critical Concepts in Linguistics*, publicada por la editorial Routledge en el año 2007.

Resumen: En esta entrevista, entre otras cuestiones, el Doctor D. Wolfgang Teubert sitúa en su contexto la aparición y edición de la revista *Internacional Journal of Corpus Linguistics*, que fundó en 1996 y de la que ha sido editor durante diez años. A lo largo de la entrevista, ofrece una panorámica de distintas vertientes y aplicaciones de la Lingüística de Corpus, desde sus orígenes hasta la actualidad, e incluso señala posibles proyecciones de esta disciplina lingüística. Asimismo, el Profesor Teubert comenta algunos principios de la Lingüística de Corpus, como la manera particular en que ésta se acerca al lenguaje o su idea de *significado*. Finalmente, se aproxima a las coordenadas en las que esta corriente lingüística resulta especialmente útil en los estudios diacrónicos de la lengua, en concreto en los ámbitos de la lexicografía y lexicología históricas.

CRISTINA MARTÍN: *In 1996, you founded the* International Journal of Corpus Linguistics. *What is this international journal about?*

DR. D. WOLFGANG TEUBERT: I am not anymore the editor of the journal; I have passed this role on to a young colleague of mine in Liverpool, to Michaela Mahlberg. After I had been the editor of the journal for over ten years, I felt I had to pass on the helm to someone more dynamic and proactive. Times have changed. When I set up the journal twelve years ago, there was already a growing interest in this new paradigm of empirical language research, but there was only a rather small community of corpus linguists. They were mostly working in applied linguistics, concerned with the training of future language teachers and a few lexicographers. In the meantime, corpus linguistics has spread to more fields. It is the preferred methodology in much of critical discourse analysis (CDA). While originally focussing on a synchronic view of language, it has now begun to develop tools for the exploration of the diachronic dimension of the discourse. It is also finding its feet in (descriptive) translation studies. While these fields are all part of applied linguistics, corpus linguistics has also become a common label in other linguistic paradigms, particularly in cognitive linguistics. As a result, it has now become a bit difficult to define the essence of it. There are now several journals carrying the word *corpus* in their name, and plenty of papers in other journals and conference proceedings have it in their title. Computational linguists and those working in artificial intelligence, too, are using the label. Today, there is an abundance of views of the essence of corpus linguistics.

The *International Journal of Corpus Linguistics* caters to those linguists who endorse a bottom-up view of language, as opposed to those who try to relate language data to top-down categorisations and conceptualisations, categories and concepts which precede the analysis of the data and thus are language-external. They could be seen, for instance, as logical or as mental universals, or as elements of a «language-independent» conceptual ontology. «Bottom-up» means that corpus linguists prefer to deal with categories derived directly from the analysis of language data and tend to be sceptical of the received wisdom which has accumulated over centuries. They do not use their corpora to find examples for the rules we find in grammars, but to question these rules. They analyse the entirety of the language data of their corpora by searching for statistically significant co-occurrences between linguistic phenomena. Keeping an open mind, they often explore what was uncharted territory for traditional linguistics. Without a corpus, it was impossible to look into the role of collocation. It was corpus linguistics that has put an end to the cherished belief that the single word in isolation is the key carrier of meaning.

Since 1996, when this journal was founded, proper corpus linguistics has rapidly developed. Today, it has become the accepted paradigm of language research in applied linguistics all over the world. But theoretical linguists or linguists working in the framework of cognitive semantics or cognitive linguistics or other kind of cognitive studies, too, use the methods developed by corpus linguists for providing an empirical basis for the claims they make. What these linguists overlook, though, is that corpus linguistics, as a different and new way to look at language, is embedded in its own framework, and can make fully sense only within this theory. When we look at the contributions published over the years, we find that the pragmatism of applied linguistics assumes these theoretical foundations often in a rather implicit way and shies away from a confrontation with the stances of other linguistic paradigms. If corpus linguistics is to survive as a distinct and new way to look at language, such a discussion must take place, sooner rather than later.

C. M.: *What does the corpus mean for corpus linguistics?*

W. T.: Corpus linguistics is empirical. It deals with data, real language data. Corpus linguists are not the only empirical linguists. In the 19th century, it was the philologists who also analysed real language, in particular texts of dead languages, like Old English, Latin, Greek and Hebrew, to find out about their grammar and lexis. They believed in a rule-based language system. There are rules telling us which constructions can be used with which words. So there are for instance verbs in Latin which have an A.c.I. (an accusative with an infinitive) as a complement, e. g. the *verba dicendi*, the verbs of saying. Some of them, for instance *admonere* («admonish»), can also take an *ut*-clause. Things are similar in English: we find «she admonishes him to pay his bills» and «she admonishes him that he should pay his bills». So a philologist would describe the local grammar of *admonere* by pointing out the two options. The corpus linguist would not stop there. She would analyse all occurrences of *admonere* in a corpus representing the Latin language according to the specifications she has made (such as period, keyness of authors, etc.). She would not look only at the verb in question and the grammatical constructions coming with them, but also at the lexical context in which the verb forms are embedded. For her, these forms are the node of possible collocations, and she will check out the hypothesis that the one construction is more frequent with certain collocates, while the other construction is mostly found with other collocates. While philologists would generally assume that «open-choice principle» (as John Sinclair, one of the founding fathers of corpus linguistics, has called it) prevails so that it does not matter which words are slotted in, corpus linguists know that often the «idiom principle» is at work: certain grammatical constructions are equivalent to certain lexically realised patterns. A corpus linguist would like to find out if that is the case for *admonere*. You can only do it if you have the full evidence, and not just a few hand-picked examples.

For the philologists, in the absence of a corpus there was no alternative to working with those examples which drew their (informed) attention. But corpus linguists have learnt that our linguistic intuition can often mislead us. This is why they insist on dealing with the full evidence, with all the occurrences of *admonere* in their corpus.

C. M.: *What are the guiding principles of corpus linguistics?*

W. T.: Once linguists had begun to realise that there is a lot more of lexical patterning going on than lexicographers and grammarians were aware of, corpus linguistics waited to be invented. This happened in the sixties of last century when computers were in the reach of academics and it became possible to work with electronic texts. It was John Sinclair's project on collocation in spoken language. It is truly amazing to see that all the foundational principles of it were already there in surprising clarity in this very first groundbreaking study. The final project report, which had for a long time been hidden from the community, *The OSTI-Report*, was finally published in 2004 as *English Collocation Studies*. But this project has left innumerable traces.

Over the last forty years it has become generally accepted by linguists of all feathers that meaning is not normally to be found in single words in isolation. Only if we look at words in their contexts we will find out about meaning, not so much the meaning of the single word itself but of the word in conjunction with its collocates, those words that co-occur with the node word in significant frequency. This is the principle of collocation. A corpus will tell us that the adjective *friendly* co-occurs more frequently with the noun *fire* than we would expect with a random distribution. If we analyse the occurrences of *friendly fire*, we notice that all of them have the same meaning, and that this meaning cannot be reduced to what the dictionary has to say about *friendly* and about *fire*. It is this co-occurrence of words that generates meaning.

For me, this also means that meaning is less a phenomenon of what Ferdinand de Saussure called *la langue*, the abstract system of the French language, but of *la parole*, the language as it occurs in texts. There is no secret formula telling us what *friendly fire* means. It is the discourse community that negotiates the meaning of lexical items. The phrase appears to have first been used in the Vietnam War, as a hardly recognisable counterpart to *hostile fire*. In the beginning people had to be told by those using it what it meant («a term that means mistakenly shooting at your own side», Google tells us), and there will have been any amount of explicit or implicit negotiations among people if, when and how this new phrase should be used. Today, we find it also used outside the military context. In 2001, we read in the *Times*: «Blair's effort was under friendly fire. Labour rebels, disgruntled backbenchers, forced the first Commons vote on the conflict in Afghanistan». What has also changed that in this new usage is that *friendly fire* can be intentional, not occurring by mistake, but on purpose. Corpus linguistics makes us aware that meaning is not part of an immanent language system but that it can be renegotiated and changed whenever challenged in a discourse situation.

C. M.: *Are there other important principles in corpus linguistics?*

W. T.: Corpus linguistics is an empirical approach to language. It looks at what kind of language is being used by various people. There is, for corpus linguists, no «ideal native speaker», as posited by Noam Chomsky, the listless promoter of language universals, who has dominated much of theoretical linguistics in the second half of the 20th century. For him, differences between natural languages are but a surface phenomenon. He has said, time and again, that if a Martian linguist were to visit Earth, he would deduce from the evidence that there was only one language, with a number of local variants. For Chomsky, and for the language philosopher Jerry Fodor, a former student of Chomsky, the «real» language is not what someone speaks or writes, but the language of thought, the language each of us is born with. In their view, language acquisition is the same as learning to translate from the language of thought into one's native language. For them, the language system is situated in the mind, and it is universal. This is their object of research. Corpus linguistics takes in many ways the opposite view. It looks at language as a social phenomenon, as something taking place in form of exchanges between the members of a discourse community. It takes Saussure's concept of *parole* seriously, regarding the discourse as the only linguistic reality available for our analysis. It investigates what is been exchanged in terms of symbolic content between people and what is shared by them.

Corpus linguistics is not concerned with the kind of rules we find in traditional grammar or with the rules posited in Chomsky's universal grammar or with rules which are a consequence of the ways a mind works. You can disobey rules of the first kind; external things (e. g. *stammering*) can impair your linguistic performance and violate rules of the second kind, while as long as your mental language processing mechanism is working properly, you cannot volitionally misapply rules of the third kind.

Corpus linguistics does not presuppose a language system. It analyses nothing but real language data, as we find them in the discourse, or in a corpus representing a discourse. It is bottom-up. Corpus linguists use increasingly sophisticated statistical tools to probe into the co-occurrence of all kinds of language phenomena, thus revealing patternings of all sorts that were just not visible before the advent of corpora. These patternings can be interpreted and described as trends or even as regularities, and thus corpus linguists might in the end come up with an endorsement of the rule that subject noun and finite verb form have in principle to agree in number. But on the way there, they would be able to list up all those nouns having plural forms but often taken to be a singular (such as *United States*, *data*, *news*) and all those nouns that in spite of being in singular are treated as plurals (*police*, *staff*, *class*, etc.). Corpus linguistics presents, describes and interprets the data. The generalisations it aims at are based on corpus evidence; they do not involve anything outside the discourse under investigation. The rules I mentioned above all presuppose a discourse-external reality, be it the categories in the tradition of Latin grammaticography, be it the *parole*-independent system of the French language or the universal laws a Chomskyan language system, or be it a mental mechanism.

The bottom-up approach of corpus linguistics does not lead to rules comparable to those of traditional linguistics or other system-focussed linguistic paradigms, but to generalisations, based on the statistical analysis of large amounts of real language data. Statistical tools tell us which co-occurrences of language phenomena are «significant». It is important to remind ourselves that this concept does not entail that something which is statistically significant is also relevant. It just tells us that according to a particular mathematical calculation two (or more) language phenomena, for instance two words like *friendly* and *fire*, co-occur in a corpus significantly more often (or less often) than a random distribution of these phenomena would make expect us. Now we have seen that *friendly fire* cannot easily be decomposed and therefore has to be learned as a unit of meaning. But actually in the British National Corpus of 100 million words, we find that *friendly atmosphere* is more frequent than *friendly fire*. Should we also describe it as a unit of meaning? Or is this a phrase that can be decomposed, and will we understand it if we know how the words *friendly* and *atmosphere* are commonly used? Indeed for a language learner it might be important to know that you can talk about a *friendly*, a *homely* and a *relaxed* but rarely about an *amiable atmosphere*. It can be useful to know that *friendly*, but not *amiable* is a significant collocate of *atmosphere*. For traditional linguists there is no way to tell us that the phrase *amiable atmosphere* is to be avoided. They are at a loss when they are asked if a given phrase that doesn't violate fixed rules may be used, and they have to rely on their intuition: «It's not what I would say». The corpus linguist can demonstrate that so far hardly anyone has been using it. Corpus linguistics tells us what is safe to say. But nothing should keep us from being innovative. A new bestselling novel with the title *An Amiable Atmosphere* might motivate many people to use this phrase.

C. M.: *How does corpus linguistics deal with language change?*

W. T.: So far, there is not a strong tradition to analyse language change. The reason is that in the past most corpus linguists have been analysing synchronic corpora, corpora that took no notice of the date when a given text was said, written or published. Lexicography has been one of the key areas of corpus research. What you enter into a general language dictionary is what can be generalised about lexical items or units of meaning. So far diachronic corpus linguistics, to the extent it exists, has mostly been comparing the language of two periods by comparing the generalised findings of two corpora representing them.

In the future, I hope, corpus linguistics will also explore the meaning of a single occurrence of a lexical item or a phrase. It will compare one particular occurrence of *friendly fire* to all the previous ones and the subsequent ones found in a truly diachronic corpus and try to find out in which way it may differ from what has been said before and how it may have impacted on what has been said later. This is, I believe, a necessary step if we want to move beyond the remit of making generalisations, if we want to develop corpus linguistics as a kind of *parole*-linguistics, a linguistics that can help us to make sense of a particular text or text segment within a discourse. For language is dynamic. The synchronic perspective which has dominated so much of 20th century linguistics does not make us aware of the importance of the diachronic dimension of language. For no text starts at point zero. Each new text is a reaction to what has been said before. It can endorse it, it can vary or modify it, it can reject it, or it can reflect it. In doing so, it will recombine, permute and rephrase words and phrases in slightly new ways. To understand a text (segment) fully, we have to contextualise it, we have to contrast it with the diachronic continuum of the texts of a given discourse from which it evolved.

What is needed if we want to explore the diachronic dimension of the discourse is a method of detecting and visualising intertextuality. What a given unit of meaning means depends, as I have said in the beginning, very much on the context in which we find it embedded. This is the context of the window of four or five words to the left and to the right of our node, the only context that our statistical tools can deal with, because once you enlarge this window it becomes sheer impossible to separate noise from what is significant. But what this unit may mean also depends on a much wider context, indeed on everything that precedes it in the text to which it belongs. We know that *a hostile atmosphere* means not poisonous gasses surrounding a planet but the mood or tone of a place or setting if the text in question has been talking about symbolic interactions that are taking place or mentions localities such as a court room or xenophobic provincial town. Intertextuality tells us that we can safely infer that a hostile atmosphere can have threatening aspects, even if *threat* or *threaten* does not occur in the context of our exemplar, as half of the citations in the discourse contain the word. Everything that has been said about hostile atmospheres anterior to the citation we are analysing can have made an impact on it. If our text has mentioned *a hostile atmosphere of antagonism*, we might be surprised to find that Google lists 72 occurrences of this phrase, and it would be useful to know how our citation compares to them and if it refers, at least implicitly, to something that has been said before. If corpus linguistics has over the last forty years concentrated on the business of generalisation, the time has come, I believe, to point our attention to what makes a given occurrence of a lexical item, a phrase or a text segment unique, and how it deals with what has been said before.

C. M.: *Isn't this kind of interpretation rather subjective?*

W. T.: Absolutely. The question has always been where we place linguistics. Does it belong to the «hard» sciences, like chemistry, or does it belong to the *Geisteswissenschaften*, the human sciences, also called the interpretive sciences? While the philologists of the 19th century contented themselves with making sense of the texts they were analysing, it was Ferdinand de Saussure whose aim was to establish linguistics as a «hard» science. The «hard» or natural sciences are about facts and truth and reality. They are searching for laws or rules, formulae, that explain what is happening and that predict what we are going to find. Much of mainstream linguistics in the 20th century was about modelling a system that would tell us which sentences are possible and which are not.

While it makes sense to say that meaning, for instance the meaning of a phrase like *a hostile atmosphere of antagonism* is in the discourse and nowhere else, in the ways this phrase has been used and in the contexts in which we find it embedded, and in the usages and contexts of the elements of which this phrase is made up, it needs an act of interpretation to produce a paraphrase that sums this meaning up. Interpretations, however, are of necessity contingent and therefore subjective. Based on the same evidence, I may paraphrase the meaning of *a hostile atmosphere of antagonism* differently from the way you do. There is no perfect, final paraphrase of it. Indeed, each new paraphrase itself contributes something new to the meaning of our phrase, thus adding a novel touch. Language is dynamic, and to deal with this aspect is, I believe, the new challenge for corpus linguistics.

C. M.: *How could corpus linguistics contribute to the historical dictionaries making?*

W. T.: Traditionally lexicographers first take a look at what has been said in existing dictionaries. They try to reinterpret what they find on the basis of their knowledge or intuition and they will adduce perhaps some examples they have picked out from the texts they have read. Corpus linguistics does not pick individual examples but analyses all the occurrences of an item in question. It makes an effort to order the results not according to the established categories of, say, traditional grammar, but in a data-driven bottom-up approach on the basis of the patternings that the analysis has documented. This is what distinguishes the *Cobuild* dictionary, designed by John Sinclair, the first completely corpus-based dictionary, from all the previous learners' dictionaries. But the *Cobuild* dictionary is synchronic. For a diachronic dictionary we need a suitable corpus of historical depth that is ordered by the date of the texts of which it consists. A diachronic dictionary has to record language change. We want it to tell us when *friendly fire* came up, how this new phrase was discussed at the time, how it was used then and how the way it is used has changed over time. The discourse is full of paraphrases, particularly when new lexical items are introduced and have to be explained to a wider audience, or when there is disagreement in how they should be used. Whenever we find paraphrases for a word (for instance for *globalisation*) or a phrase like *friendly fire*, we can be sure that we are monitoring language change. Therefore I would expect a corpus-driven historical dictionary to give its users access to such paraphrastic content. Of course, it would be impossible to compile a corpus that would document in the way sketched here the history of a whole language such as Spanish. Our selection of texts would only ever contain a tiny fraction of what has been written, and thus we would find only random intertextual links, and only now and then a relevant paraphrase. This is why historical dictionaries such as the *Oxford English Dictionary* tell us about the all the different senses in which a word has been used and what the first record of a new sense is in a given corpus, electronic or not. But they do not tell us about the discussions accompanying the emergence of a new sense that alone would give us an insight into the dynamics of language.

Yet if we think of a more focussed discourse, for instance the discourse of the catholic social doctrine in the 19th century, we could easily compile a corpus containing all the more relevant texts. It is a rather limited community of people who were in a position to contribute to this discourse, and they were mostly aware of what the others had said. In such a corpus there would be an abundance of intertextual links in the form of overt and covert references, and there would also be an abundance of paraphrastic content, detailing how concepts such as «work», «property», «justice» and «rights» changed the way they were seen over time.

The way I would visualise such a dictionary is as a users' electronic workbench. They would type in the phrase they are interested in, and they would be given, in temporal order, the relevant corpus citations, and they could search for intertextual links, similar phrases occurring in previous and subsequent texts. Thus such an electronic dictionary, rather a workbench, would be a sophisticated query system enabling its users to pursue the clues they are interested in, in as much detail as they like. Their queries would link them up with all the relevant corpus citations, not just with the limited digest of them that lexicographers have to come up with in printed dictionaries, due to space restrictions. Of course, we are still a long way away from such a utopian tool. For the immediate future, I am afraid, we are still stuck with printed dictionaries. That means we are forced to accept compromises due to their limitations in size. Instead of the whole corpus evidence we have to accept the lexicographer's digest of it plus perhaps a few subjectively selected citations.

C. M.: *What would be the community of the Spanish discourse on science and technology of the 16th century Spanish, for which we are now engaged in a historical dictionary project?*

W. T.: That is very difficult to say really. Traditionally, I believe, many scientists in the 16th century, particularly those dealing with the more philosophical aspects of their disciplines, still used Latin among themselves for communication. It was only when these scientists needed to communicate with the craftsmen designing and constructing the technical devices that they would communicate with them in the vernacular, that is in Spanish. But I may be wrong. What is necessary is to sample the libraries archives and see how these texts were preserved. Those considered at the time more important will have been published by learned societies or by university presses, in book form or in academic journals to the extent that these existed at the time. There were any number of disputations in which different opinions were weighed against each other and a lot of paraphrases were used to define the discourse objects the scientists were discussing. For the communication between them and the early engineers, if we can apply this term to the craftsmen, I am afraid we have to compile letters exchanged between them that we may still find in archives. There we would find, as I see it, extremely valuable material, for detailed explanations had to be given for the ideas that the scientists wanted the craftsmen to work on.

To some extent, the scientists then would also have wanted to attract a larger audience of educated people who did not understand enough Latin and not enough about the disciplines to read the texts written for discussion within the scientific community. The humanists certainly wanted to educate a wider public, and therefore often used the vernacular besides Latin. There will also be translations of Latin texts and popular versions of them in Spanish. These would certainly contain a broad range of paraphrases, as the terminology used needs to be explained, but not so many intertextual links. All this could go into an apposite corpus. It would be exceedingly expensive to compile, because these texts have to be keyboarded by hand. Perhaps it would be easier and less expensive in the short run to compile different corpora for different disciplines.

C. M.: *Would all that has been said about science and technology in the 16th century contribute to the meaning of the relevant terms and phrases?*

W. T.: There is, at least in theory, quite a principled distinction between terms and lexical items, as the terminologists have taught us. Terms are supposed to denote concepts which are strictly defined in their essence, preferably using a language-independent definition, such as H2O for «water». The idea is that if I do not abide by the decreed definition I misuse the term. How lexical items can be used and what they mean is a matter of negotiation by the discourse community. In practice, however, such a clear distinction between terms and lexical items is rarely maintained. Perhaps we could say that when it comes to terms in the 16th century, the community of those to be in a position to negotiate their meaning would be much narrower than that for lexical items as they are used in the general language. It was for instance the theologians and the *literati* who kept discussing the concept of «conscience», taking the first cautious steps towards internalising it inside the individual mind of a person. The educated public had no part in the definition of *conscience*.

As to the meaning of the dictionary entry items, it really would depend on what your envisaged user should expect. If they are interested in how a word was used by the general educated public in the 16th century, you would have to focus on texts written for such a public. If your users want to find out how the meaning of a lexical item like *conscience* was discussed and how it gradually underwent change, then ideally they should be able to compare all the relevant paraphrases offered by various schools of thought at the time, and not just the lexicographer's digest of the evidence. If you want to go even further, you would have to also include the Latin original texts and find out which Spanish translations equivalents were chosen for which Latin expressions. For a printed dictionary, it might be a good idea to select primarily those (paraphrastic) citations that are intertextually linked with other citations, for instance of the kind «It is wrong to say that X is Y because in reality X is Z».

If you have to make a selection of texts for your corpus, you obviously have to choose the more relevant texts. These are those texts which have an impact on other texts. Again this is a matter of intertextuality. Some foundational texts may be frequently referred to, while other texts leave only little or no traces in subsequent texts. These more peripheral texts contribute almost nothing to the meaning of a lexical item. They can easily be left out.

C. M.: *Would a historical dictionary as you have sketched it here not give up the distinction between encyclopaedic and lexical knowledge?*

W. T.: Very much so, and quite on purpose. The idea that we can distinguish between the two is prevalent among lexicographers, but they themselves do hardly abide by this rule. A concept like «conscience» is an object of the discourse, not of a discourse-external reality. As a discourse object, it does not exist outside language. The lexical item *conscience* and the discourse object «conscience» are entirely co-referential. This is also true of other scientific concepts, for instance *gold*. It does not matter what gold is in the world outside the discourse. We can only communicate within the discourse, and we can only describe the concept of 'gold' and the meaning of the word *gold* by relating what has been said about it.

C. M.: *How do you see the future of corpus linguistics?*

W. T.: It's difficult to say, at the point where we are today, how things will look in ten years' time. The term *corpus linguistics* is now used in so many ways that it can mean almost anything. So far I do not see that those working in the Sinclairean tradition of corpus linguistics are really making an effort to develop it into a paradigm with its own theoretical foundation. A reason may be that most serious work in corpus linguistics, for instance that of Michael Stubbs or Susan Hunston or Michael Hoey or Douglas Biber, is taking place in what is called applied linguistics, encompassing language teaching, the analysis of languages for special purposes, lexicography and other useful areas. Corpus linguists often seem too polite to clash with theoretical linguists.

If corpus linguistics is there to stay we have to present it as a novel way to look at language, to discuss language as a social phenomenon, and to establish it as the kind of *parole*-linguistics that Saussure failed to deliver. Then it will become an invaluable instrument in our aim to make sense of the many discourses with which we are confronted. For corpus linguists, language is not a mirror of a reality out there. Our reality, the reality that we refer to when we communicate with each other is a reality that has been constructed in the discourse. Corpus linguistics can make visible the patternings that we are not trained to recognise; it will show the contexts in which words and phrases occur, and it will reveal the hidden links between different texts. Corpus linguistics gives us access to the meaning of what has been said. It provides a shareable foundation on which we can collaboratively interpret the meaning of words, phrases, text segments and complete texts.

This view situates corpus linguistics within the philosophical paradigm of hermeneutics. Corpus linguistics, as I see it, should be recognised as the cornerstone of an *ars interpretandi* that allows us to collaborate in our endeavour to make sense of what is said and what is known without forcing us to agree to any particular reading.

C. M.: *As Gadamer says in the title of his book* Truth and Method, *can this new method be a way to find the truth?*

W. T.: Hans-Georg Gadamer was a great ironist, and I am pretty sure he accepted this title of his *magnum opus*, suggested to him by his publisher, because for him hermeneutics was exactly the opposite, namely the absence of method and truth. As he sees it, hermeneutics is an art and therefore cannot be reduced to a method, and no interpretation would ever reveal the true meaning of a text but add only another reading to all the readings already there. The best we can come up with is, he tells us, an interpretation that we can accept as the meaning a given text (segment) has for us in this moment. In this sense, hermeneutics and corpus linguistics are cognate, I believe. But different from corpus linguistics, hermeneutics has always investigated the diachronic dimension of the discourse. It tells us that in order to understand what *conscience* has meant in the 16th century we must first of all make sure what *conscience* means today. We have to lay open, step by step, the intertextual links that connect today's meaning to the meaning the word had then. Gadamer calls this the fusion of the horizons of understanding. It is this diachronic dimension that is still absent from corpus linguistics as it is practiced today. I am convinced it is a challenge worth taking.