How does a Dictionary Describe a Language?

Pius ten Hacken
School of Arts (Translation), Swansea University
Singleton Park, Swansea, SA2 8PP, United Kingdom
p.ten-hacken@swan.ac.uk

Abstract
It is common for monolingual dictionaries to state that they are a dictionary of the English (French, etc.) language. This raises the question as to what interpretation of *language* is meant in this context. Reference is often made to the use of a corpus. However, a corpus can never constitute the language being described. The lexicographer has to decide which occurrences in a corpus are errors and which reflect playful use of expressions not literally used in the text. The issue can be clarified by using Chomsky's distinctions between competence and performance and between I-language and E-language. Neither competence/I-language nor performance offer a viable notion of language that can be described in a dictionary of the English language. Although Chomsky considers E-language a problematic concept, it is shown that it can be used in the context of lexicography if we consider lexicography an instance of applied science.

1 The Self-Perception of Dictionaries

It is common to refer to a dictionary such as the *Collins English Dictionary* (*CED*) or *Webster's Third* (*W3*) as describing the English language. On closer inspection, such a statement is highly problematic. In what sense of language is there an English language described in *CED* and in *W3*? The purpose of this paper is to find an answer to this question that is at the same time convergent with linguistic theory and with lexicographic practice.

2 The Self-Perception of Dictionaries

Dictionaries often describe themselves on their title page relative to a particular language. *CED* calls itself *Collins Dictionary of the English Language*. *W3* has as its full title *Webster's Third New International Dictionary of the English Language, Unabridged*. This suggests that there is an object in the world, called *the English language*, that *CED* and *W3* are dictionaries of. An explicit reference to an object described in this way in the title of a dictionary is also common for other languages, e.g. *van Dale*, *Petit Robert*, *Zingarelli*.

In the foreword to the second edition of *CED*, William T. McLeod describes the first edition as in (1) and the motivation and preparation process of the second edition in (2) and (3), respectively.

(1) "a dictionary based on a fresh survey of the contemporary language as it was actually being used in both its written and its spoken forms"

(2) "the language is changing faster than at any time in the last 300 years."
(3) "The preparation of this new edition to take account of this rate of language change has meant the scrutiny and assessment of the many thousand citations for new words, meanings, and idioms accumulated by our in-house reading programme since the publication of the first edition;"

The English language is presented in (1) as an entity that can be used and in (2) as one that changes. A dictionary is a description of this entity. This description is based on a corpus which constitutes a survey, as (1) tells us. In (3) we get a more precise description of the working method involved. It should be emphasized here that (1-3) characterize lexicography more than the *CED*. They have been singled out only because they provide a particularly clear view of the notion of language and the corpus-based method adopted in lexicography. Other dictionaries often have similar statements in their preface or assume similar views tacitly.

3 Problems with the Lexicographic Concept of Language

A concept of language that is suggested in particular by (2) is that of a language as an organism. This view developed in the 19th century under the influence of historical-comparative studies resulting in genealogical trees (Schleicher's *Stammbaumtheorie*, cf. Robins (1967:178ff.)) and was clearly influenced by Romanticism. Although it can be used as a metaphor in some contexts, the concept of a language as an organism is problematic for various reasons. Thus a dialect continuum of which the extremes are clearly different languages, e.g. between French and Italian, is difficult to account for in terms of a French language and an Italian language as separate organisms. Ten Hacken (2005) discusses the implications of this observation for dialect geography in more detail.

An alternative view of the concept of language suggested by (1) and (3) is that its main characteristic is its being represented by a corpus. This is reminiscent of mid-20th century American linguistics, whose attitude is formulated in (4).

(4) "The universe of discourse of any linguistic study is a set of utterances." [Hockett (1942:3)]

The problem with (4) is that it only characterizes linguistic study, not language. We can think of language as what underlies this set of utterances, but its exact nature remains vague.

Let us for the moment accept the vagueness in underlying concept of (4) and consider to what extent it is possible to work with a "set of utterances" or, as (1) states it, a "survey of the contemporary language as it was actually being used". A number of methodological problems arise, among others that the corpus may contain phrases such as (5).

(5) "A full service of your toilets are at 18:30" [notice at some toilets at Swansea University]

It is clear that (5) contains an error. It is also clear that it represents actual use of the language. The corpus does not mark errors as such, so that we cannot rely on it as an authority.

Lexical errors are usually much more subtle than syntactic errors of the type represented in (5). The corpus we use is likely to contain them and it will not tell us where they occur.

A second type of problem is the non-occurrence of certain expressions. Corpora are notoriously bad as a tool for the study of idioms. The expression *kick the bucket* is likely to be more frequent in linguistic papers on idioms than in any other section of a corpus. As Moon (1998) shows, many idioms are more often used in playful, non-canonical forms than in their standard form. Dictionaries usually solve these problems, but they have to do so by going beyond the use of the corpus as in (4). The *Oxford Dictionary of English Idioms* (*ODEI*) actually gives (6) as an example of *fiddle while Rome burns*.

(6) The novel is rich in evidence of the trivial snobberies and hypocrisies which obsess our upper and upper-middle classes as they fiddle while London smoulders.

The idiom illustrated by (6) is only alluded to in this example. Nevertheless, ODEI takes it to be an example of the idiom in use. Without knowledge of the idiom, it is not possible to interpret *fiddle while London smoulders*.

In conclusion, the concepts of language that come to mind most immediately on reading (1-3) cannot be taken as a basis for the answer to the question in which sense of *language* a dictionary describes a language. The idea of a language as an organism can only be used as a metaphor. It may be useful in restricted contexts, but should not be taken literally. The concept of a language as represented in a corpus is incomplete, because a corpus generally contains errors and does not contain all expressions of the language.

4 Three Conceptions of Language

Developments in linguistics from the late 1950s onwards have led to a renewed and more intensive discussion of the question of what is a language, ultimately leading to a clarification of the concept. Two distinctions are essential. Although the terminology in which they are expressed here is based on work by Chomsky, the concepts referred to are adopted by most current approaches to linguistics in one form or another.

The first distinction is the one between competence and performance. According to Chomsky (1966:3), "A distinction must be made between what the speaker of a language knows implicitly (what we may call his *competence*) and what he does (his *performance*)." Competence and performance are entities of a very different nature. It is our competence which allows us to recognize an error in (5) and to categorize *fiddle while London smoulders* as a modified instance of *fiddle while Rome burns* in (6). As far as Chomsky's distinction has been challenged, the most contentious point has been whether grammatical competence can or should be distinguished from pragmatic competence. That performance is an entity of a different kind is beyond dispute.

A second distinction is the one between I-language and E-language, introduced by Chomsky (1986:19-22). I-language is internalized. It can be seen as what is in the competence of a speaker. As he states elsewhere, it "consists of a computational procedure and a lexicon" (1995:15). E-language is externalized. It is a collection of linguistic forms paired with meanings, understood independently of the speaker's mind/brain.

It should be noted that while I-language and competence refer to what is basically the same concept, E-language is something quite different from performance. Competence is a mental object in the real world. It is physically realized in the human brain although there is currently no way to relate individual physical processes (neurons firing) to individual elements of knowledge of language. Performance is a non-mental object in the real world. It can be realized acoustically in sound waves, visually in signs or characters, etc. By contrast, E-language is not an object in the real world. It consists of grammatical sentence types with their meanings. The lexicon of an E-language contains the correct meanings of words as used in grammatical sentences.

Let us now consider what is described in a dictionary in terms of these distinctions. We can see immediately that a dictionary does not describe performance. Performance is at the basis of the survey referred to in (1), but this is a survey *of* the language, not the language itself. This is also recognized in lexicology, when Sinclair (2003), for instance, warns against the uncritical use of corpus tools. At first sight it may seem attractive to state that a dictionary describes an I-language. We should keep in mind, however, that it is physically impossible for two people to share an I-language. They would have to share the same mind to do so.

This leaves E-language. However, according to Chomsky (1986:26), "languages in this sense are not real-world objects but are artificial, somewhat arbitrary, and perhaps not very interesting concepts." A crucial problem is that because an E-language is not a real-world entity, there is no obvious way to determine whether a sentence or a word in a particular meaning are part of it.

This problem can be clarified by considering the status of (5) and (6). Example (5) is verifiably part of performance. We can individually verify that it is not grammatical according to our competence. The sense of shared knowledge, however, which seems to emerge from our individual verification that (5) is ungrammatical, is not realized as an object. Yet it is this shared knowledge that lexicographers aim to describe. Example (6) highlights another aspect of the problem. Lexicographers at *ODEI* analysed it as the creative use of an idiom. If a pedantic person P would object to this and maintain that the use is incorrect, how could *ODEI* lexicographers defend their decision? Not by an appeal to occurrence in the corpus, because (5) also occurs. Not by an appeal to competence, because they do not know the details of P's competence. This is what I will call the *authority problem* associated with E-language.

5 Empirical versus Applied Science

Linguistics of the type for which Chomsky proposed the distinctions between competence and performance and between I-language and E-language differs in several respects from lexicography. One aspect that is essential in the present context concerns the type of goal they pursue. Science can be divided into different types according to the role they attribute to explanation. The most important difference in this context is the one between empirical and applied science.

An empirical science has as its purpose to describe and explain phenomena that can be observed in the world. Typically the explanation is in terms of a model of the underlying system for the observations. An example of an empirical science is astronomy. Astronomy

explains the position where planets can be observed on the basis of the orbits of the Earth and the individual planets around the Sun. These orbits are in term explained in terms of the interaction of gravity with other factors (e.g. mass).

Linguistics is also, in its most common form, an empirical science. It tries to account for observations concerning language by describing the underlying system of competence. Chomskyan linguistics analyses competence as a mental module interacting with various other mental modules and tries to explain it in terms of the language faculty, deemed responsible for its development in an individual. Other approaches to linguistics analyse the role of competence differently. They have in common, however, that they attempt to formulate a theory that describes a real-life entity which explains observations about language.

If lexicography is taken as an empirical science, we have to see the dictionary as a theory of an empirical object. This is problematic, because performance and competence, though empirical, are not what is described and E-language is not empirical. It is also questionable whether a dictionary should be seen as a theory at all.

Applied sciences differ from empirical science in that they incorporate a problem-solving component. A prototypical example of an applied science is medicine. Medicine encompasses various branches, including curative medicine, preventive medicine, and palliative medicine. They are all concerned with illness, but rather than just explaining the observations of illnesses in patients they also have a practical goal: curing the patient, preventing illness, or alleviating suffering. Compared to the goals of astronomy, they are of an essentially different type.

If lexicography is taken as an applied science, we have to identify a problem that the dictionary is supposed to solve. This perspective of a dictionary as a tool for solving problems is much closer to how many lexicographers understand their product than the perspective of a dictionary as a theory. Dictionary users turn to a dictionary in order to get answers to their questions. They find an answer to the extent that the lexicographer has foreseen their question. The nature of the answer depends not only on the input material used by the lexicographer, but also, and crucially so, on the lexicographer's interpretation of this material. The lexicographic problem can therefore be formulated broadly as describing an E-language in a way that satisfies the users.

Applied science is not only a matter of solving problems, but also of doing so in a scientific way. In empirical science, theories are typically supported by testing their predictions in experiments. In applied science, the corresponding support for a problem solution can be given in two ways. The first is to consider the solution as a prediction, the second to explain the solution in terms of underlying knowledge. In the case of curative medicine, a problem solution is a cure. The proof that a cure works corresponds to these two methods. First, support for the validity of the data showing that it works is collected in terms of large, double-blind experiments. Second, the way the cure works is explained in terms of biochemical processes in an anatomical model.

In lexicography, the same two types of evidence can be used to raise the level of scientificness. In practice, the method that has been applied is the reference to a corpus. Invoking the corpus, as in (1), as an authority and the careful analysis of the corpus, as in (3), can be interpreted as a way of supporting the validity of the data. These data do not have the exactly parallel role to the data collected by double-blind testing in medicine, however. A closer par-

allel is the study of dictionary use of the type described by Bogaards (2003). The second type of data, corresponding to explaining the efficiency of a cure, would explain the correlation between the description of the lexicon as an E-language in a dictionary and the usefulness of the dictionary for particular purposes in terms of a theory of language.

6 Conclusion

I started this paper with the question in what sense of *language* the *CED* and *W3* can be said to be dictionaries of the English language. If the relation *be a dictionary of* is taken to be parallel to "describe", there is no adequate answer to this question. "The English language" only exists as an E-language, not as an empirical entity. The crucial property that makes it unfit for empirical study is the lack of a naturally occurring authority to decide, for instance, what its boundaries are. A much more promising approach is a conception of lexicography as an applied science and dictionaries not as descriptions but as solutions to practical problems. In this view of lexicography, we can understand the lexicographer as providing the missing authority that the concept of E-language suffers from in the context of empirical science.

References

A. Dictionaries

CED: *Collins Dictionary of the English Language*, second edition, ed. Patrick Hanks, Glasgow: Collins, 1986.

van Dale: *van Dale Groot Woordenboek der Nederlandse Taal*, 12th edition, ed. G. Geerts & H. Heestermans, Utrecht/Antwerpen: Van Dale Lexicografie, 1993.

ODEI: *Oxford Dictionary of English Idioms*, by Cowie, A.P.; Mackin, R. & McCaig, I.R., Oxford: Oxford University Press, 1993.

Petit Robert: *Le Nouveau Petit Robert: Dictionnaire alphabétique et analogique de la langue française*, ed. Josette Rey-Debove & Alain Rey, Paris: Dictionnaires Le Robert, 2003.

W3: *Webster's Third New International Dictionary of the English Language, Unabridged*, ed. Philip B. Gove, Merriam, Springfield (Mass.), 1961.

Zingarelli: *Il Nuovo Zingarelli: Vocabulario della lingua italiana*, 11th edition, ed. Miro Dogliotti & Luigi Rosello, Bologna: Zanichelli, 1988.

B. Other literature

Bogaards, P. (2003), 'Uses and users of dictionaries', in van Sterkenburg, P. (ed.), *A Practical Guide to Lexicography*, Amsterdam, Benjamins, pp. 26-33.

Chomsky, N. (1966), 'Topics in the Theory of Generative Grammar', in Sebeok, T. A. (ed.), *Current Trends in Linguistics Volume III: Theoretical Foundations*, Den Haag, Mouton, pp. 1-60.

Chomsky, N. (1986), *Knowledge of Language: Its Nature, Origin, and Use*, Westport (Conn.), Praeger.

Chomsky, N. (1995), 'Language and Nature', *Mind* 104, pp. 1-61.

ten Hacken, P. (2005), 'The Disappearance of the Geographical Dimension of Language in American Linguistics', in Spurr, D. & Tschichold, C. (eds.), *The Space of English*, Tübingen, Narr, pp. 249-264.

Hockett, C. F. (1942), 'A System of Descriptive Phonology', *Language* 18, pp. 3-21.

Moon, R. (1998), 'Frequencies and Forms of Phrasal Lexemes in English', in Cowie, A.P. (ed.), *Phraseology: Theory, Analysis, and Applications*, Oxford: Clarendon, pp. 79-100.

Robins, R. H. (1967), *A Short History of Linguistics*, London: Longman, 2nd edition, 1979.

Sinclair, J. (2003), 'Corpus processing', in van Sterkenburg, P. (ed.), *A Practical Guide to Lexicography*, Amsterdam, Benjamins, pp. 179-193.