Göran Hammarström

Comment on Klaas Willems' contribution

Klaas Willems begins (p. 1) with some thoughts about linguistics: "in the last two decades, the use of quantitative methods in synchronic and diachronic linguistics has been booming". And he says further: "Anyone who has taken the trouble to study large amounts of data will readily admit that [...] corpus-based research reveals more than any intuition-based or introspection-based focus can provide." These are strange statements. For more than 200 years linguistics has been almost completely corpus based and it still is. The studies of the early historical linguists and those who were usually called philologists or neogrammarians were practically always corpus based. I think they can even be critiqued for often using more examples than was really needed to prove their point. In the first half of the twentieth century many universities had departments of phonetics where the amount of data was enormous. (One may regret that many of these departments have become departments of linguistics or communication or they have disappeared.) The structuralists were corpus orientated but had limited interest in variants. From about 1950, when the easy to handle tape recorders became generally used, the study of spoken language with enormous amounts of data has boomed. This kind of research has been called corpus linguistics, an inappropriate name because it does not distinguish this research from almost all other kinds of linguistics, which are also corpus based. Variationists have done a great amount of important work. There have been centres of quantitative linguistics. Quantitative methods have always been used but they have not boomed in the last two decades as Willems believes. I cannot see what Willems refers to. Some of Chomsky's followers may have recognised that the data one produces sitting at one's desk are not sufficient but I do not believe that this is what Willems considers to be a change in linguistics. Willems says (p. 1) that what one finds "in large corpora [...] is often at variance with what traditional grammars and dictionaries of German claim". This is what no competent linguist would doubt: grammars and dictionaries are generally successful in including the most important facts. Since there is already a high number of these facts, there is only limited room for variants. On the other hand books and articles based on large corpora which began being published more than one hundred years ago are full of variants and types of facts which can find no place in grammars and dictionaries. There are thousands of such publications that Willems seems to ignore. It is unbelievable that he thinks that it is appropriate "to demonstrate the importance of the issue", i.e. the fact that large corpora contain variants that are not included in grammars and dictionaries. Then he proceeds to his demonstration which is an account of the variants in his present study of certain German verbs. Both his strange idea and his demonstration are based on an unbelievable misconception of what most linguists find normal. It would be meaningless to discuss here the details of a demonstration which is superfluous in the present context because it is just carried out in the well known way of all similar investigations. However, I cannot help mentioning the significance tests of the differences he has found. The result is that it is extremely improbable that these differences would be non-existent or reversed if all the examples in the corpus (or even more examples)

had been included. Just looking at the differences he shows in his statistical tables is enough to say that they are absolutely certain not to change even if more examples were included. This certainty is better than the high degree of probability of his significance testing. The examples which are different from the mainly used construction are so small in number, 1 and 10 %, that one suspects that some speakers never use them. Some of these examples could also be errors. Willems says (p. 3) about his results: "Findings such as these have potentially implications for the general debate about the relationship between grammar and usage." No linguist would doubt that language use possesses variants that have not been included in a grammar book and in the dictionary. This is not a new finding that can give rise to new conclusions. Willems does not show any of the potential implications and one cannot see that there is anything that can be debated.

At the level of syntax to which Willems' demonstration belongs I do not believe that frequency has a place because it has no function that is described as part of syntax. There is, however, an interesting although little known aspect of language where frequency is crucial. Despite my deep respect for corpus based data, I will provide comments on an example from my memory. However, I am certain that it is correct. Most students would in conversation say *uni* while speakers who less often say the word use *university*. Imagine a student who most mornings leaves for university. He may say *I am off to uni* or similar. Why should he say *university* with five mainly unnecessary syllables when the interlocutors have understood when he has uttered the first phoneme /j/ of the word *uni*. Persons present have perhaps even known that the word was coming even before it was pronounced. Even *uni* is somewhat longer than it had to be but that is generally true for words. Such considerations hold for the whole vocabulary of a language. Frequent words are shorter because they are easier to be perceived by the hearer or reader and less frequent words have more cues because they are less easy to be perceived. The former are used in contexts that have often been used, which is not the case when rare words are used. One problem with proper names is that in spoken and written form they have usually no context that can help the listener and the reader to perceive them. If one studied the pronunciation of speech segments in different positions in a sentence (which nobody has done as far as I know), certain segments would be found to be clearer pronounced because they are more difficult to predict by the hearer and other segments would be less clearly pronounced because they are more predictable, more frequent in some position. This problem can be studied at the level of phonemes, morphemes, lexemes and syntagmemes. In a more detailed and technical study one will use notions such as frequency, transitional probability, redundancy, predictability and amount of information. I have used these notions in Sound change and information theory (in H. Eichner et al., eds, *Fremd und Eigen* [...] *in memoriam Hartmut Katz*. Wien: Editions Praesens, p. 33–37). It is a study of the disappearance of phonemes in words between Latin and Modern French. I show that the more predictable a phoneme is in a word, the more probable it is that it disappears from the language. Willems does not see the existence and the importance of the ideas explained from the beginning of the present paragraph. (I have also mentioned these problems in my Comment on Johannes Kabatek's Comment on Göran Hammarström's contribution.) He is interested in the problem of frequency and does not see the most important frequency problem in a language. He says (p. 5, footnote 1) "a form which is very likely to occur in the next sentence {...} may be infrequent in the language". While this remark is correct, it is uninteresting because there

is not much to be studied about frequency relationships if one considers relationships between two sentences. However, there are inside every sentence many interesting problems of predictability based on frequency. Willems quotes (p. 5) Haspelmath who says that "frequency of use implies short coding because frequent items are more predictable". Willems finds this unacceptable because Haspelmath does not spell out the implication of "predictable". This is why Willems criticises it. The implication is, however, according to my explanations above, that when something is said or written, the hearer or reader can be provided with a shorter "coding" when there is high predictability because then he can perceive the meaning easily. The cue can be short. He may not even have to hear some phoneme or see some typeme because it is completely predictable inside the word. It can be added that common words are often said in frequent contexts, and this increases their predictability. On the other hand, if the predictability of a segment is low, it suits the hearer and the reader that it is longer.

Willems discusses the opinions of other authors (p. 3–7). Most of this discussion concerns topics which do not fall under "frequency" mentioned in his heading. Saussure's dichotomies *langue – parole* and *synchrony – diachrony* are discussed in some length. He agrees with the authors, among whom is Coseriu, who believe that Saussure's rigorous separation of the elements of the two dichotomies can and should be overcome. I believe that this is a misconception. Against all kinds of arguments, I am convinced that the separation is absolute. The most basic facts are as follows: An individual has a langue, i.e. knowledge of how to produce spoken and written texts. Whatever happens when he uses this knowledge, he can only use it as it is in the moment it is used. If a linguist studies data of parole, he cannot be allowed to add to the speaker's langue aspects that are not there. The *langue* of a speaker is a knowledge stored in his memory and it is different from the activity of producing *parole*. One can add to this knowledge the knowledge of how to use it. The langue cannot change while it is being used. It is what it is. From the speaker's viewpoint it is purely synchronic. Linguists can add all kinds of viewpoints but they do not falsify Saussure's correct ideas of how speakers produce parole. If a linguist studies diachrony and considers language in a more general way than considering just one speaker, he still has the problem that if he mixes results from one synchronic description into the synchronic description of some other time, he only produces confusion. A diachronic study must be based on at least two pure synchronic descriptions from different times. I have argued that if langue is a knowledge, and parole an activity, one can add a third main part which is missing, i.e. the spoken or written text. I have often discussed these things (see, for instance, "Two basic problems: Static synchrony and causes for change", *Folia Linguistica Historica 19/1–2*, 1998, p. 3–6, and *Fundamentals of Diachronic Linguisics*. München: Lincom Europa 2012. See further details in my Comment on Johannes Kabatek's Comment on Göran Hammarström's contribution). It can be added that in a text written by hand, some letters are more and others less clearly written. It seems, however, that such differences are less clear and less common than in spoken language.

Willems finishes his paper (p. 7) with the following words: "To the extent that multifactorial quantitative analyses are able to capture subtle interactions between a diverse array of factors, one may be confident that their empirical findings will greatly benefit the ensuing grammatical analyses without falling prey to a naive positivism which confuses descriptive detail with a comprehensive understanding of language in discourse." I find

these impressive words extraordinarily empty. Willems has presented a piece of research which he has carried out exactly as the positivists did 100 years ago. They tried to explain in different ways what they found in their statistically treated data. Willems has not drawn any of the conclusions mentioned by him on the basis of his statistical data. Where are "the multifactorial quantitative analyses" and "the subtle interactions" which "will greatly benefit the ensuing grammatical analyses"? Even if he tried to find the impressive results he imagines, I am afraid he will not achieve much because frequencies and statistics are not the kind of data which contain anything on which his imagined results in grammar can be based. The first fifty years of the last century would have been the most positivistic in linguistics. Statistics based on hundreds or even thousands of examples were common in papers and books. The advantage was that one looked closely at all the examples and found many interesting variants and drew the conclusions one was capable of drawing.