Evidence contra Ignorance

Leila Behrens, Universität zu Köln

> "If we engage the issue with evidence, we give the question credibility. If we dismiss it, we provide evidence of our arrogance."
>
> (anonymous blogger)

1 Epistemological Questions of the Seventies and Today

According to Thomas Kuhn (1976), there is typically no or only very little interest in epistemological issues during periods of "normal science". In such times, scientists are said to be interested in "research as usual", i.e. in what Kuhn calls "puzzle solving" within the theoretical and methodological boundaries of the respective "disciplinary matrices" or "paradigms". Instead of scrutinizing the key elements of the paradigm at hand (i.e. its basic theoretical notions and assumptions, its methods and instruments, and its general values) on a daily basis, scientists normally take them for granted and routinely apply them to those unsolved problems ("puzzles") which are considered relevant at the moment in the scientific community one belongs to. Kuhn suggests that asking the big questions of philosophy of science, e.g. concerning knowledge acquisition and progress in science in general or the necessity of a revision of a higher-order methodology in a discipline is usually considered waste of time in such "normal" or "quiet" periods.

It is not my intention to enter a critical discussion on the question of whether Kuhn's well-known stance is generalizable across all scientific disciplines and whether the inverse conclusion (high interest in epistemological issues is a sign of a paradigm shift) is a valid or (at least) useful description in the history of science. However, we do indeed observe a sort of alternation between different phases in the history of linguistics. Periods in which discussing methodology (i.e. the "how" of knowledge acquisition, instead of presenting research results which are supposed to increase existing knowledge) is rather unpopular alternate with periods in which methodology becomes a central topic. Typically, the latter are somewhat shorter and accompanied by hot debates, not only about the "right" sort of data acquisition or analysis but, in general, about the self-conception of linguistics and its status and place within other scientific disciplines.

We are currently experiencing such a period in time, characterized by a higher degree of openness for epistemological topics, and the willingness to discuss the basis questions: where does linguistics stand today? How useful are its methods and in which direction will or can it move? This attitude can be witnessed in several papers, workshops, and volumes which have titles such as "Linguistic Evidence" (cf. Kepser/Reis 2005a) or "What counts as evidence in linguistics?" (cf. Penke/Rosenbach 2004a). Those who believed that, long ago and on both (Chomskyan and anti-Chomskyan) sides, everything has been said about competence/performance (or I-and E-Language), stimulus (in)dependence, the psychological reality of theoretical constructs, the risk and merits of representationalism, etc., might be surprised at the considerable response to Devitt's (2006) new book "Ignorance of Language". While Devitt challenges the Chomskyan line of argumentation from the perspective of philosophy of science, Dan Everett (2007) does so by presenting linguistic data from one single language, i.e. Pirahã. He argues that Pirahã lacks recursion (a feature which is still assumed as being an indispensable part of UG (Universal Grammar)), and therefore it "presents problems for the notion of Universal Grammar" and as such ultimately provides arguments against "the methods traditionally used in the Generative Grammar tradition" (Everett 2007).

The really interesting thing is not the content of this claim (i.e. the lack of recursive structures and its consequences) but the question how it is possible that this kind of language-specifically supported argument could have caused such a sensation that even the non-academic press (e.g. "The New Yorker", April 16, 2007) devoted a significant amount of space to this issue. Why just this piece of counterevidence against UG, why not other ones presented earlier (cf. Tomasello 2004)? I guess this is, among other things, because the time, as in the seventies, is ripe once again for such debates. Indeed, the epistemological debates fought out at that time bear a striking resemblance to the current ones. However, besides this feeling of déjà vu, there are also considerable differences between the seventies and the present. In this paper, I will try to work out some crucial differences between these two periods, and, at the same time, I would like to draw attention to some areas of linguistic methodology which are still misunderstood or simply ignored.

When comparing the seventies with the first decade of 21st Century, we can discover a certain shift in the focus of interest. The main question at those days was whether or not linguistics is an empirical science at all (cf. Ballmer 1976; Perry 1980), whereas the question of evidence was

more or less subordinate to this issue. In contrast, linguists' concentration nowadays revolves around the question of how to define "evidence" under different goals and in the different subfields of linguistics. This seems to presuppose a latent agreement that linguistics belongs to the class of empirical sciences. What is still open is how empirical status or quality should be measured in the case of linguistics, i.e. by comparison to natural sciences such as physics and chemistry as it was usually done according to the standards of philosophy of science in the 20th century, which made a binary division between empirical sciences (physics and chemistry) and non-empirical ones (formal logic, mathematics) but ignored the possibility of further divisions between empirical sciences? The rapid development and emancipation of biology encourages a rethinking of this traditional picture; Mayr (2004), Caramazine et. al. (2001) and others argue that biological systems are basically different from physical systems (i.e. the former being characterized by self-organization, greater diversity, higher complexity of interactions, etc.), and that consequently, biology requires basically different strategies of research and different measurements of empiricalness as well. Some linguists (cf. Ackerman 2007) adopt a similar view for linguistics, arguing that linguistics has a considerable similarity to biology but not to physics or chemistry.

A further difference in comparison to the situation in the seventies concerns the growing diversification of "linguistic models", and, in connection with this, the growing diversification of attitudes to data acquisition. Retrospectively, the seventies appear as a heyday of a cold war between generative and non-generative linguistics. According to the "official claims" on both sides, the study of linguistics was described by a mutually exclusive characterization of a) the subject of investigation (language capacity in general vs. languages as observable in their use), b) the types of explanation (innate constraints vs. functional considerations), c) the "right" type of data sources (introspections vs. corpus data), and, perhaps, also by d) the type of presentation ("formalized" i.e. stated in a mathematically precise vocabulary vs. expressed in ordinary language). As noted by Kepser and Reis (2005b: 2), this debate over explanatory and descriptive goals and appropriate data sources "died down after the seventies without virtually any consequences on linguistic practice". However, in my view this does not mean that, as suggested by Kepser and Reis, generativists and non-generativists "continued relying" on complementary data sources (the former only on introspective data, the latter only on corpus data). In actual fact, both of them have always worked with mixed types of data

acquisition, typically depending on rather trivial factors such as the (non-)availability of prior knowledge and resources, a particularly important point in poorly documented languages, which were as a rule not the native languages of the investigators.

However, I agree with Kepser and Reis that the necessity of systematically controlling introspective data, just like the necessity of systematically comparing and complementing data from different sources, was not generally acknowledged until the early to mid nineties. In that sense, early critical suggestions against the standard positions basically fell on deaf ears, both the suggestion that introspection is not a direct key to competence, but itself a particular kind of performance whose result has to be analyzed in order to reconstruct competence, and the suggestion that "observability" is an illusion in the context of linguistics (perhaps with the exception of phonetics/phonology), based on a crude, structuralistic simplification which ignores the pervasive ambiguity of language and the fact that linguistic utterances cannot be measured in the same manner as physical objects but have to be interpreted beforehand.[1]

The reasons for the turning point in the nineties can be traced back to several background events. First, we should mention the dynamic development in "computational corpus linguistics" (CCL), which made very large corpora in many languages electronically accessible and searchable. In this way, some of the old objections against the structuralist concept of "corpus" became obsolete and working with corpora more and more attractive. In addition, it was research in CCL (cf., e.g., Justeson/Katz 1991) that, by systematically comparing corpus-based evidence with experimental evidence (e.g. frequency results with acceptability or goodness ratings), helped to get the idea accepted that "every method counts" (cf. Arppe/Järvikivi forth.; Featherston 2005). Another important step toward stronger methodological sensibility was the publication of Schütze's (1996) book that provides a very careful and well-balanced summary of the entire linguistic and psycholinguistic literature dealing with theoretical and methodological problems involved in grammaticality judgments. His conclusion that all grammatical judgments are in fact acceptability judgments (cf. Schütze

[1] Behrens (1998) provides a detailed evaluation of this old methodological debate, tracing back how certain linguistic myths (such as those about "structuralist corpus", "observability of linguistic phenomena", etc.) have evolved and then been polemically employed in this debate. At the same time, I have tried to collect early self-critical voices and proposals for optimally combining all available methods and data sources.

1996: 26) initiated a process of rethinking across the generative/non-generative boundary, leading to the following claim with which I fully agree: "even so-called primary data from introspection as well as authentic language production are complex performance data".

Linguists are used to thinking of themselves in terms of antagonistic blocks such as "generative" vs. "non-generative" or "formal" vs. "functional" linguists. Such a simplification by reducing many controversies to two "clearly counterposed positions" could be useful in some contexts, as argued by Newmeyer (1998: 7). Concerning basic methodological questions, however, it turns out to be less and less helpful. As emphasized by Newmeyer himself, both orientations in question are quite heterogeneous (cf. also Penke/Rosenbach 2004b). This is particularly true of what he calls "functionalism", which comprises a variety of rather different approaches and therefore has aptly been compared to Protestantism by Elizabeth Bates, being like "a group of warring sects which agree only on the rejection of the authority of the Pope" (first cited in Van Valin 1990: 171, see also Newmeyer 1998: 13). In addition, most of the differentiating terms are misleading, because they are either ambiguous (and as such allow for an orthogonal classification in their different senses) or are on the way to becoming obsolete in their original sense in some approaches, or they represent loaded terms. The term "formal", for instance, is ambiguous between 'form-oriented' (pertaining to form rather than to meaning or form-meaning relations) and 'formalized'; however, as frequently noted, there exist formalized functional models (e.g. Functional Grammar) as well as form-oriented approaches that are not formalized. The expression "structuralist" (shortly considered by Newmeyer for replacing "formal" or "generative") is a historically loaded term. Not even the term "generative" necessarily applies in its original sense (i.e. in the sense of being able to "generate" all well-formed sentences of a language and specify their structures) to all models commonly classified as "generative". This depends on the question of how the goal of descriptive adequacy (on this interpretation) is considered in relation to explanatory goals (e.g. the specification of universal principles): is the former considered a prerequisite of the latter, as usually taught in textbooks (cf. Radford 1988: 27-30), or is it considered more or less irrelevant, as claimed by Chomsky in a famous quote?

> The class [of well-formed (grammatical) expressions of L] has no significance. The concepts 'well-formed' and 'grammatical' remain without characterization or known empirical justification;
>
> (Chomsky 1993: 44-45)

According to this view, the minimalist program is not any more a "generative" theory in the original sense of the term.

Under such circumstances of diversification I do not find it very surprising when different positions about the status of evidence "cannot be delineated along the formal-functional opposition", as noted by Penke and Rosenbach (2004b: 508) in the introductory paper to the special issue about evidence they edited. (Asked on the status of typological evidence in linguistic theory in general and in the construction of UG in particular, the authors of this volume uttered opinions that were clearly cross-cutting the traditional boundaries of "formal" vs. "functional" orientation.) The question is, of course: what do linguists exactly mean when they use the word "evidence"?

2 Different Types of Evidence and Some Misconceptions

Questions such as "What counts as linguistic evidence?" are ambiguous. In addition to a normative reading ('what should be counted as evidence according to some scientific norms?'), it may also have a descriptive interpretation: what kinds of entities linguists typically refer to when they talk about "evidence"? Under what circumstances do they say that they "have presented evidence" for something?

The polysemy of the English term "evidence"

Drawing some attention to the second reading might be interesting because it reveals some crucial discrepancies between norms defined in terms of philosophy of sciences and actual linguistic practice. Asking how reference is made to "evidence" in linguistic works can only be done with respect to particular languages. Because English is the current *lingua franca* in science in general and in linguistics in particular, I will consider the different uses of the English word *evidence*. There are some language-internal and cross-linguistic indications that the English word *evidence* is polysemous, even when we leave its uses in scientific contexts and in the court context aside. In everyday language, it can be understood either as 1) 'a hint, an indication, a sign' or as 2) 'a proof'. In a number of languages (e.g. German, Spanish, Hungarian), these two senses correspond with different translation equivalents (German: *Indiz*,

Zeichen for 1, and *Beweis* for 2). In such languages, the distinct translation equivalents of evidence typically represent so-called "contextual antonyms", i.e. they can be set in direct contrast in the right context, asserting that something is only a (weak) hint (German: *ein Indiz*) but not a (hard) proof (*ein Beweis*).[2] Obviously, the first sense has an inductive (or abductive) flavor, the second one a deductive flavor. These observations fit the collocational behavior of the English word itself:[3] among the most frequent collocations (attributive adjectives) of *evidence*, some (e.g. *supportive* or *convincing*) bias the weaker, inductive- or abductive-like sense, while others (e.g. *conclusive*) point to the existence of a stronger, deductive-like sense. Accordingly, it is only the latter class of collocations that *evidence* shares with nouns such as *proof* (which seem to be confined to a deductive-like sense); adjectives of the former class, in turn, are rarely combined with *proof*, and if this is the case they typically carry certain implicatures (e.g. that there exists some material which is presented as a proof but probably does not deserve this status).[4]

It seems to me reasonable to assume that the tentative polysemy of *evidence* in everyday language could have an impact on its scientific use. The very fact that logicians distinguish between "inductive evidence" and "deductive evidence" (rather than designating these two concepts by entirely distinct expressions) may be seen as an influence from the everyday use of the term. And conversely, if this distinction is obscured in some scientific context we are entitled to ask whether this comes from a neutralization of the sense difference in everyday language.

---

[2] Most European languages also have expressions which are cognate to the English word *evidence* (e.g. German *Evidenz*). Due to the increasing influence of English as dominating language of science, these cognates are, of course, more and more in use in the non-English scientific literature as well, gradually replacing the native expressions and copying the polysemy of the English model.

[3] My frequency analysis in English is based on BNC material. For the search queries, I used PiE ("Phrases in English", http://pie.usna.edu/). For German collocations, I used the tools and corpora of "Wortschatz" (Universität Leipzig, http://wortschatz.uni-leipzig.de/).

[4] Note that *conclusive evidence* is regularly translated as *schlüssiger Beweis* into German, whereas the combination of *unterstützend* (lit.: 'supportive') and *überzeugend* (lit.: 'convincing') with *Beweis* results in expressions which are in a similar way rare and semantically marked like their English counterparts. This could be due to a common-sense reasoning built into the lexical meanings of *proof* and *Beweis* which generate the expectation that the premises are complete and there is also general consensus about their truthfulness. Adjectives indicating incompleteness or expressing perlocutive effects seem to be at odds with these expectations.

Abductive and deductive evidence

So what kind of evidence should be at issue in linguistics? When browsing through linguistic abstracts we soon notice that, with the exception of experimental research in psycholinguistics and neurolinguistics, an overwhelming part of mentions of evidence makes reference to a complex set of pre-analyzed data. The pieces of such selected data are then put together in order to motivate or "support" a particular analysis in a particular theoretical framework (which is not necessarily a single model such as LFG but can also be any hybrid framework). We are dealing here with typical cases of abductive reasoning during which all but the most probable hypotheses are eliminated. This abductive argument about the best hypothesis (analysis) is then often used to confirm or disconfirm the usefulness of certain theoretical constructs and assumptions.

The following points should be emphasized. Usually, publicly presented data (forms, sentences, texts) which enter abductive analyses are to a high degree pre-analyzed, i.e. they are transcribed, translated or paraphrased, glossed with grammatical and lexical types of information, etc. In this respect it does not matter how they are collected, i.e. from one's own intuition, from informants, or from spontaneous speech. The only difference is that in the third case, usually no judgment notation (*, **, ?, ??, etc.) is used, which is based on the implicit assumption that the presented data would and/or should be generally acceptable in the speech community. In this way, they contain a great number of implicit theoretical assumptions (or "unexpressed premises"; cf. Gerritsen 2001). (In Section 3, I will argue that this uncertainty about implicit assumptions, when combined with an uncertainty about the reliability of judgments and their notation or about the reliability of corpora, can have a fatal effect for the scientific community's assessment of what is presented as "evidence".) In abductive reasoning as described above, we often encounter reference to linguistic evidence at different levels of abstraction. In addition to annotated data, sometimes the whole argument (i.e. the argument which is employed to support the new hypothesis) is also called "evidence". Arguments in this sense can be considered a sort of higher-order evidence. When used as counterevidence, they do not refute theories "directly" in the sense that they would show that some predictions made by some theories are not conform with observable outputs of language use (incl. reported metalinguistic judgments). Rather, they put forward reasons why certain (singletheoretical or pantheoretical) assumptions and constructs

constitute a hindrance to expressing important generalizations. This is compatible both with proposals for slight modifications and basic change requests. I will argue that in some contexts we can only have higher-order evidence, e.g. in the context of cross-linguistic investigations. From this, it does not follow that a completely synonymous use of "evidence" and "argument" would be generally justified.[5] In my opinion, it is, for instance, rather misleading in the case of primarily rationalist arguments.[6]

Contrary to what one might assume, it happens rather rarely in actual linguistic practice that what is called "evidence", represents a true example of (positive) "deductive evidence". Of course, "deductive evidence" plays an important role in those subdisciplines of linguistics whose methods are rooted in the experimental paradigm (i.e. psycholinguistics, neurolinguistics, etc.), and also in many fields of

---

[5] The reader should prove for himself/herself how often in an ordinary textbook of linguistics (e.g. Radford 1988, Ouhalla 1999) the terms "evidence" and "argument" are used in a synonymous fashion.

[6] A case in point is the misleading substitution of "argument" by "evidence" in the "poverty of stimulus argument", as sometimes happens in the literature. Some premises of this well-known argument for linguistic nativism are empirically testable (e.g. the premise that children are only provided with positive inputs), others are presumably not. Take for instance the premise that some linguistic patterns cannot be learned from input data alone but only on the basis of an innate predisposition, e.g. an innate language-specialized cognitive mechanism. Here the question arises whether this premise could be refuted by computer modeling alone, provided that it would be possible to demonstrate by means of sophisticated computer algorithms that all relevant structures are, in principle, learnable. Or would this kind of counterevidence be treated as potentially irrelevant, since, as noted by Newmeyer (1998: 88), "that would not [necessarily] mean that the human brain *follows* that algorithm[s]"? If so, the premise in question is not empirically testable (for a detailed discussion of the question of missing evidence in the "poverty of stimulus argument" see Pullum/Scholz 2002). Apart from factual questions, it is the rhetorical presentation which makes this argument so controversial. Its usual presentation manifests more than one argumentation fallacy, where argumentation fallacies are defined as "deficient moves in an argumentative discourse" (van Eemeren/Grootendorst 2004). Asking "how else if not innate" and emphasizing that there is no better theory for language acquisition (Penke/Rosenbach 2004b: 502ff.) manifests the fallacy of charging the burden of proof to the other party and also what is called "argumentum ad ignorantiam" (concluding that a standpoint is true because the opposite has not been successfully defended). Describing non-nativists as reductive empiricists à la Skinner provides an example of the fallacy of the straw man, just like allegations such as "[t]o say that "language is not innate" is to say that there is no difference between my granddaughter, a rock, and a rabbit" (Chomsky 2000: 50). The latter is also a case of the "argumentum ad hominem", by depicting the other party as stupid. Argumentation fallacies corrupt the notion of evidence in the scientific discourse, at least in the Popperian sense of "open science".

computational linguistics. Besides these exceptions, however, it is hardly the case that linguistic material, as typically presented in publications (i.e. in the same form as described above in connection with abductive reasoning), could be understood as empirically tested demonstration of deductively derived conclusions. There are two main reasons why this is so. First, many linguistis work on phenomena and /or languages which are unknown or not very well described. So what they are doing is discovering new facts and providing new hypotheses for them. Second, linguistic theories (of whatever orientation) usually do not make very specific and precise predictions which could be tested against empirically collected independent (!) data-sets, as required by the hypothetic-deductive method.[7] Later below I will argue that the hypothetic-deductive method can only be considered a fruitful research strategy within language-specific investigations – if at all. This is especially so when this method is combined with the idea of "naive" falsificationism, according to which linguistic forms alone (accepted or rejected by some speakers or found in some corpora) might be cited as falsifying counterevidence.

Falsifiability

Unfortunately, recent literature dealing with linguistic evidence is not very clear with respect to the question of falsifiability in language-specific vs. cross-linguistic contexts.

Most authors provide a rather detailed classification of evidence along different parameters (cf. Penke/Rosenbach 2004b; Kepser/Reis 2005b). Penke and Rosenbach, for instance, distinguish between

a) qualitative and quantitative evidence,

b) "direct" and "indirect" evidence (e.g. according to "how directly the data reflects language knowledge" (p. 487), evidence from corpus data

[7] According to this, the concept of "generative grammar", defined as a "device" which generates and structurally specifies all well-formed sentences of a language, is best considered as a *Gedankenexperiment* (a "thought experiment") whose main purpose is to support a particular philosophical position. It is, however, not a concept which has ever been taken literally (i.e. as a testable device) in the mainstream of descriptive generative work, which is empirical in the sense of "data-driven", but has a puzzle-solving orientation, rather than aiming to provide exhaustive description of languages. From this, a general incompleteness of premises follows, so that it is never the case that the conclusions necessarily follow from the premises (as they should in a deductive argument) but are, at best, supported by them to a certain degree. In this sense Chomsky's above-cited characterisation (p. 6) is certainly to the point.

and grammaticality judgment being considered by them as "direct", evidence from experimental effects as "indirect"), and

c) evidence from elicited and evidence from spontaneous data.

Further types may be added depending on the area of investigation (synchronic vs. diachronic studies, language acquisition, etc.), and, in the case of experimental data, depending on the relevant subdiscipline (psycholinguistics, neurolinguistics, etc.) or the specific experiment design. What is striking in this kind of classification is that usually no reference is made to meaning. So for Penke and Rosenbach (p. 486) e.g., "qualitative direct" evidence only involves the existence of certain forms or constructions:

> Using data qualitatively simply means that we use data to show that a certain form/construction is possible in a specific context or that a certain experimental effect occurs in an experimental setting.

This is a very important point, since ignoring meaning, rather than treating it on a par with form, when evidence is concerned, could be one of those features of the general "methodological climate" which have not changed very much since the seventies. Indeed, the good old debate between advocates of introspection/elicitation and those of observation/corpus was a debate between two sections of form-oriented linguistics. The argument of the non-mentalist fraction that observable evidence would be superior to evidence relying on intuition (which is only accessible to the outsider by metalinguistic reports) makes sense only when we are primarily interested in forms rather than in "unobservable" interpretations. Although corpora provide an extremely valuable source of hypotheses about meanings, without any doubts, every linguist working on subtle semantic questions knows that they will not suffice, especially not when the linguist is not a native speaker of the language investigated. Native speakers' intuition is indispensable in such cases, if not for any other reason than to obtain the whole range of possible interpretations of corpus attestations.[8]

---

[8] Presumably, it is not by chance that among Labov's arguments for the superiority of observational data (when there is a discrepancy between these and introspective judgments), the most convincing ones come from phonology (cf. Labov 1975). On the other hand, it is probably also significant that at that time when Chomsky's claim that intuition about grammaticality is separable from intuition of meaning became popular wisdom in generative textbooks, many of his followers worked on their own languages, i.e. under conditions in which prior knowledge of meaning can be taken as self-evident and therefore ignored.

Considering forms and constructions as basic entities in qualitative evidence prompts another question. Should this hold true only within language-specific boundaries or in cross-linguistic investigations as well? Is it, for instance, possible to provide "evidence" in the form of sentences (or other linguistic forms) from Language X in order to support an analysis in Language Y? And conversely, can sentences (or other linguistic forms) from Language X be presented as "counterevidence" against an analysis primarily based on Language Y (and perhaps some other languages)? If the answer is "yes", what does "evidence" mean in such cases?

As far as the issue of falsification is considered, linguistic literature is dominated by a somewhat ambivalent attitude to the Popperian view on falsification in empirical sciences (cf. Popper 1972). On the one hand, it is generally acknowledged that empirical hypotheses and theories cannot be "verified" in a strict sense, by simply adducing further data, but can only be falsified by demonstrating that the predictions derived from the hypotheses do not fit data collected by observation or experiments. On the other hand, it is a weaker version of the principle of falsifiability that became accepted in linguistics, following some arguments of the extensive criticism against Popper, particularly those put forward with respect to research in social sciences (cf. Diekmann 2005: 153ff.). Now, it is generally agreed that a single counterexample will not do the job of refuting a hypothesis or even a complex set of hypotheses (a "theory"); ideally, a systematically collected set of counterexamples is needed (cf. Penke/Rosenbach 2004: 482ff.; cf. also Dahl 1980). Moreover, it often happens that something that appears to be a counterexample is as such not relevant, because it does not belong to the domain of the hypothesis (cf. again Dahl ibid.).

However, the problem in linguistics lies not so much in the weakening of theories by reformulating universal statements into statistical tendencies. Rather, there is a potential for a serious clash when linguists commit themselves to one side of the Popperian idea (i.e. to the superiority of the deductive-hypothetic model over inductive methods) without fully accepting its other side (clear criteria for falsifiability). It seems to be sinking into oblivion that Popper strongly required not to shift the burden of falsification to the contrahents. He insisted, rather, that it is the obligation of the scientist who puts forward a theory to specify in advance under which conditions she or he would accept that her or his theory is falsified (cf. Popper 2004: 53). In other words, making it clear in advance that what the successful conditions of a falsification would be is the price for being allowed to make audacious hypotheses;

this controlling device should prevent us from "immunizing" scientific claims, e.g. by unsystematically pointing to the irrelevance of counterexamples suggested by other members of the scientific community. To summarize, the postulate "No evidence without precise specification of counterevidence" should help us to prohibit the misuse of science; as such, this postulate is, in my opinion, still attractive. At what level of abstraction "evidence" and "counterevidence" should be formulated, is, of course, a specific business of linguistics.

Now let us come back to the question asked above: is it possible or reasonable to consider (sets of) sentences from a particular language as "evidence" or "counterevidence" for or against an analysis which is based on data from another language (e.g. for or against the "subjacency principle" which was originally motivated by English data[9])? In my

---

[9] It is very revealing what Chomsky says about evidence in an interview he gave in 1983 (cf. Rieber 1983). Using "the subjacency principle", he explains in this interview the difference between evidence "for explanatory theories" (based on patterns of acceptable and unacceptable utterances) and evidence "for psychological reality" (based on psycholinguistic experiments). As for the former type of evidence, Chomsky describes a pretty inductive process of generalization, where the starting point is a single speaker, then one looks for comparable arrays of acceptable and unacceptable utterances with other speakers in the same language, and finally with comparable utterance configurations in other languages (!). The lack of idiosyncrasy is explained by growing probability when one finds more and more instances intra- and cross-linguistically. Note also the expressions "the speech of some speaker" and "accounts for what that person is doing," which clearly point to an understanding of "evidence for explanatory theories" as "evidence from E-language:"

> Well, let's keep within the range of what is called, in what seems to me a rather misleading locution, "linguistic evidence." So, let's suppose that I'm investigating the speech of some speaker – let's say, myself – and I find that there is a strange array of acceptable and unacceptable utterances. Suppose I'm considering interrogative expressions. I find that some are well-formed (for example, "who do you think won the game") while others are not (for example, "who did you ask what game won," meaning: "who is the person x such that you asked what game x won"). Suppose now I find that I can explain the array of possible and impossible questions by assuming some abstract principles that constrains [sic!] the grammar. Then somebody comes along and says, how do you know that's not idiosyncratic. We know how to find out: I look at the next person and see whether he has a comparable array of possible and unacceptable interrogative expressions and a comparable system. Suppose I find that I can explain that person's array of acceptable and unacceptable utterances by the same principle, and so on. … Suppose the subjacency principle to account for a certain informant's judgments about what is and what is not a properly-formed question, as in the examples I just mentioned. … Suppose I then proceed to show that for the next person I study the same principle of subjacency accounts for what that person is doing, and

view, this is only possible in a very weak sense of the term "evidence", i.e. as a hint that it would be worth taking a closer look into the language in question. What I mean here is not the impossibility of comparing languages, which is certainly possible and reasonable when, for each language in question, investigations are carried out at the same (sufficient) level of depth (see below). Instead, I am referring here to the common practice of using isolated collections of translated or otherwise "copied" sentences in order to show that the target language has a pattern similar to or different from the source language (i.e. the language motivating the original hypothesis). The problem with such translation-like sentences is that they wrongly suggest that, beyond the relevant sentences, everything else would be equal. This is certainly almost never the case. With the exception of dialects and closely related languages, all those factors which potentially have an intervening effect tend to be different. This is an enormous source of what Popper had called "immunization" of theories: if it is both possible that a) similar effects arise due to the interplay of different language-specific principles and b) different effects arise due to different conditions restricting similar principles, then we will not only mistrust sentences as sound confirmatory evidence but also mistrust them as feasible counterevidence, pointing to the possibility that the unexpected state of affairs (in judgments or occurrences) might be caused by some idiosyncratic properties of the language in question; therefore, the apparent counterexample would not be a counterexample at all. Indeed, this kind of reasoning is a common linguistic practice in protecting hypotheses (cf. also fn. 10 below).

The notorious problem that counterevidence will not be taken seriously because it is always possible to give reasons post festum why they should not be considered relevant exists at higher levels of abstraction as well, e.g. when languages are cited *en bloc* as evidence against certain universal statements or constructs, or against certain typological generalizations.[10] My favorite example is Tagalog, which

---

for the next person. Suppose the result extends to other phenomena and other languages.

[10] While discussing some problems of the "generativist approach to typology", Newmeyer refers to a work by Gilligan (1987) who tested the predictions Rizzi and Safir made in connection with the "null subject parameter". Gilligan attested several languages which were excluded by the hypotheses of both linguists in question. This is commented by Newmeyer (1998: 359; 2005: 93) in the following way:

> It is possible that the cases of non-predicted subject inversion, for example, are the result of something other than the null subject parameter. As Gilligan

became well-known in the seventies for posing serious problems for the notion of the subject (Schachter 1976). Tagalog has a series of particles (*ang/ng/sa*) that simultaneously fulfill the function of marking reference and grammatical relations. Accordingly, two types of approaches have been established in the meantime: one type treats these particles as determiners, the other as case markers (or subject/topic markers). Without going into detail, we can say that both approaches have some merits and some drawbacks. Kornai and Pullum (1990: 34), who proceed from the determiner analysis, mention Tagalog in a discussion of early X-Bar theory as a counterexample, violating assumptions about the optionality of determiners. It would be easy to reject this counterevidence by pointing out that the particles in question are actually case markers (or: that they show this unexpected behavior because they also have a case-marking function). In the same way, Tagalog, presented as potential counterevidence against some assumptions of case marking theories, might be subject to the opposite objection (i.e. pointing to the referential function of the particles in question). We can find thousands upon thousands of similar cases in the literature and there is no easy escape from this dilemma, neither by ignoring such cases (there are too many exceptions) nor by formulating extremely abstract universal principles which cannot be linked to language-specific categories and constructs in a principled way. In my opinion, Tagalog provides good (abductive) evidence for the assumption that languages may conflate referential and case marking functions. It is exactly this possibility which should be considered in the design of descriptive frameworks.

---

himself points out: 'Perhaps the Rizzi hypothesis is correct but its effects are obscured in [Brazilian Portuguese and Mandarin – two languages with a cluster of properties predicted not to exist] because of some yet unanalyzed aspect of languages' (1987: 90).

So what! In my opinion, the mentioned hypotheses are clearly refuted in their original form, provided that the 100 languages are adequately analyzed. Refutation implies corrections and modifications but not necessarily that the hypotheses would have to be given up entirely. It could be the case that the "null subject parameter" provides an adequate generalization for a subset of languages; in that case the theory has to account for this division of languages. Or, the hypotheses are too coarse-grained and have to be refined with respect to those "unanalyzed aspects of languages". Saying, however, that a hypothesis is "perhaps correct" but makes wrong predictions due to "unanalyzed aspects of languages" is nonsense.

Where has all the evidence gone?

It is common place in linguistics that categories from different languages cannot be assigned to each other in a one-to-one fashion (cf. Croft 2001). Since the seventies, there is an ever-growing body of empirical arguments against the universality of virtually every category or construct which has been assigned a universal status in some theory. Asking what kind of evidence could possibly refute the UG hypotheses in the Chomskyan linguistics, Tomasello (2004), e.g., points out that there is seemingly no such evidence since all those empirical arguments which come to mind as potentially having a falsifying capacity are still ignored, at least in certain versions of the model.[11] Why is, for instance, X-bar syntax still in use in many varieties of Minimalist Program (as a universally applicable representation) even though it is more or less acknowledged in the scientific community that this representational device is not particularly adequate for non-configurational languages?

The phenomenon that large amounts of valuable information that were empirically collected and as such provide new theoretical hypotheses do not become integrated into general theory is by no means a unique feature of Chomskyan linguistics. In the so-called functionalist literature, for instance, it has been repeatedly demonstrated that the concept of "basic word order" is not a workable construct for many languages, at least not when it is understood in terms of grammatical relations such as subject and object. As early as 1981, Comrie (1981: 88) noted that "word order typology" should be considered for this subset of languages as "not relevant". Nonetheless, many functionalists still work with the good old Greenbergian typology as if it had a universal applicability; and we are still waiting for a subordinate typology which takes different kinds of "non-applicabilities" in this domain into account.
Should we assume that surviving linguistic ideas have a priori withstood falsification, i.e. simply by surviving in the scientific community? Surely not. And should we assume that more or less ignored ideas have been refuted, i.e. simply by being not taken seriously in the scientific community ("refutation by neglect")? My answer is "no" in this case as well. Empiricalness is not only a question of the possibility of falsification, it is also a question of whether members of the scientific community are able to see a systematic causal connection between empirical results and theoretical modifications.

---

[11] Tomasello (2004: 642) gives a nice summary of different assumptions about what exactly is considered as a necessary part of UG.

Browsing through older literature from the seventies, we encounter plenty of psycholinguistic "evidence" for (and, of course, also "counterevidence" against) the "psychological reality" of theoretical constructs which are out of use in the current versions of the relevant models, e.g. for the psychological reality of specific transformation rules, then for deep structure generally, and so on. One is wondering whether such pieces of "evidence" (or the quantitative/qualitative relation of "supporting evidence" against corresponding counterevidence) have ever played a key role in the subsequent modifications of GB and Minimalist Program? Is it, for instance, due to the fact that the amount of counterevidence exceeded that of supporting evidence that the (relevant types of) transformation rules have been abandoned and later the difference between d-structure and s-structure as well,[12] or did such modifications happen for some entirely different considerations (e.g. for considerations about descriptive usefulness or due to some other types of evidence)? If the latter is the case, the question arises how different types of evidence are to be ranked, when some of them are allowed to cancel others.[13]

The current methodological challenge in linguistics is how to effectively combine different sources of evidence to make accurate complex predictions. There are two popular misconceptions, each of which forms an obstacle to the successful integration of knowledge, one

---

[12] Of course, one could take the view that such decisions are never made on a purely scientific basis (i.e. by weighing up supportive evidence against counterevidence) but mainly for sociological reasons, as suggested by Carlson (2003: 78) with respect to the abandonment of movement rules: there was, in the long run, an "insufficient preponderance" of linguists willing to accept the supposed evidence in favor of movement rules as satisfying "normal standards of evidence" (cf. also Wanner 1977).

[13] With respect to this question, there is a difference between a) strong mentalistic theories which equate theoretical descriptions with mental representations (i.e. assume that theoretical descriptions of linguistic phenomena are descriptions of mental phenomena and that "a description of mental phenomena is also a representation of what is actually going on in the mind or brain" (cf. Hoffmann 1989: 219)) and b) weak mentalistic theories which only maintain that linguistic descriptions should have "psychological reality" without strongly committing themselves to reified representations. For the latter, conflicts between results of linguistic analyses and psycholinguistic results are less of a problem as linguistic analyses may be generally given higher priority. In contrast, the former need independent evidence a) that the reification hypothesis is valid and b) that the hypothetical description is also a valid description (cf. Hoffmann ibid.). In addition, strong mentalistic theories have to account for cases in which descriptions valid according to psycholinguistic/neurolinguistic experiments and descriptions elegant according to traditional criteria of linguistics and philosophy of science do not coincide.

is being prevalent in some (but not all) generativist circles, the other in some (but not all) functionalist circles.

According to the first misconception (which is more widespread among generative linguists), "linguistic theories" whose hypotheses are generalizations based on a few languages could be "tested" on all other languages of the world. This idea corresponds to the first part of the Popperian claim that in science, we are always "guessing", so what really counts is to find the right problems and then put forward interesting "guesses" (hypotheses); it does not matter where they come from. The idea is, however, in serious conflict with the second part of the Popperian claim, simply because we linguists are not able to make precise statements about what could be accepted as falsifying evidence with respect to approx. 7000 languages of the world. So what typically happens is that the lack of knowledge about languages is itself used as immunizing strategy against falsification, as shown above (p. 14). At best, we could say that the idea of "theory testing", when understood as testing theoretical accounts of generalizations based on very few languages, is an extremely uneconomic research strategy at the current state of linguistic knowledge.[14]

According to the second misconception (which is more widespread among functional linguists), it would be possible to get cross-linguistic insight by simply collecting "data from reference grammars" in a large number of languages and then "generalize over them to formulate

---

[14] It is well-known that such initial theories frequently differ in their basic assumptions only because the initial set of languages accidentally contains one language instead of another, e.g. Italian rather than French, French rather than Spanish, Mohawk rather than Cayuga, etc. Ideally, it would not matter with which language we start (with English and Dutch or with Chinese and Vietnamese), because we get converging results by successive falsifications and modifications. Unfortunately, this is not what happens in linguistic practice where the initial ("theory-providing") languages heavily determine the heuristic path of knowledge integration. Some linguists commit themselves explicitly to the goal of describing some languages in terms of others (e.g. to the "goal of explaining nonconfigurational languages through configurational [ones]" (cf. Legate 2002: 38, fn. 12)). Even if this is not the case, post festum parameterisation, as it was propagated in GB, has its own drawbacks and, as a method of successive extension triggered by falsification, it is rather problematic. Relying on fine-grained differences, as observed in the initial set of languages, necessarily leads to an exponential growth of parametric differences ("microparameters") when new languages are successively "tested", i.e. to a messy collection of thousands or millions of parameters (cf. Newmeyer's (2004) arguments against ascribing parameters to an innate UG). Organizing "microparameters" into "macroparameters", in turn, necessitates some initial ideas about frequent clusterings of microparameters in the languages of the world, otherwise the whole enterprise is doomed to failure by unfalsifiability due to lack of knowledge (cf. fn. 10).

universals" which have functional explanations (cf. Haspelmath 2004: 568). Haspelmath, for example, who emphatically advances this view, claims that for typological work in this sense, "phenomenological descriptions" at a shallow level are sufficient, while descriptive frameworks which go beyond "basic linguistic theory" would simply be "irrelevant".[15] Interestingly, this kind of breadth-first approach to linguistic diversity runs into the same trouble as the depth-first approach mentioned before: by heavily underestimating the fact that there is no such thing as "analysis-free" data (and reference grammars are especially not the place where we would find them), it generates unfalsifiable statements because, under such conditions of generalization, every statement may be encountered by counterarguments that the language-specific inputs of the generalization are incomplete, misinterpreted, or otherwise inappropriate from a language-specific perspective. Moreover, disclaiming the necessity of descriptive frameworks in cross-linguistic research means failing to recognize that much of what is presented as language-specific evidence in a cross-linguistic context constitutes arguments ("higher-order evidence") against established ways of analysis, established constructs and categories of existing descriptive frameworks. Thus, insisting on "phenomenological descriptions" amounts to deliberately putting up with the loss of such evidence.

It is almost 100 years since Franz Boas (1911) formulated his famous postulate that languages (like cultures) should be analyzed in "their own terms". This has sometimes been taken as a statement against the

---

[15] Haspelmath's claim involves three independent issues. While propagating "phenomenological descriptions" as sufficient and descriptive frameworks as irrelevant, he ties up to Chomsky's earlier distinction between "observational adequacy" and "descriptive adequacy" (cf. 2004: 569), somehow assuming that "phenomenological descriptions" are less distorted by theoretical assumptions than those provided by formalized frameworks and that they are also neutral with respect to "cognitive reality" (cf. the critique by Aissen/Bresnan 2004). The term "basic linguistic theory" (coined by Dixon 1997: 128) refers to a theoretical hybrid, incorporating influences from traditional (prestructuralist) grammar, structuralist methodology, early generative grammar, and different philologies. What is highly problematic in this concept is not the eclecticism per se or the lack of formalization (in a computational sense) but its lack of a critical stance regarding hidden (and not always compatible) theoretical influences from different traditions. Finally: reference grammars in our times are by no means restricted to "basic linguistic theory" but are frequently written in specific (generative and nongenerative) frameworks. Apart from this, they do not provide, as a rule, comparable and complete information necessary for typological generalizations which go beyond some simple universals of Greenbergian style (for a critique of using only secondary sources in typological research cf. Newmeyer 1998: 326ff.).

possibility of comparing languages. However, we can also read it as a methodological prerequisite for language comparison. Only if it is guaranteed that a particular analysis for a particular language fits the optimal analysis we would arrive at proceeding from a language-internal perspective for that language, might we expect that evidence drawn from this analysis will not be rejected very soon as poor evidence. Requiring that languages should be analyzed in their own terms means first of all to apply the same methodological standards to all of them. When we accept, for instance, Occam's razor as a methodological imperative in some languages, then we should also accept that it determines what counts as an "optimal analysis" in all other languages in an analogous way. Consequently, a cross-linguistic hypothesis which implies an "elegant" analysis for some languages but a "messy" one for others should be ruled out. Moreover, optimality of analysis should be required for every language globally rather than locally, e.g. "elegant" syntactic analyses at the cost of terribly redundant morphological analyses (and vice versa) should not be allowed. In this sense, many of the well-known debates about the universality of traditional categories (e.g. word classes; cf. Sasse 1993) are debates, first of all, about the universality of methodological standards rather than about ontological assumptions. They are debates about methodological double standards and the question of how far we should go in pressing languages into a procrustean bed. The question of shared ontologies needs other types of evidence.

All things considered, the availability of new methodologies in our time still requires a further process of rethinking, as justly emphasized by Ackerman (2007):

> *Concerning methodologies*, we will not gain much by utilizing *new quantitative and experimental methodologies* if we simply apply them to familiar phenomena from the usual languages in order only to bolster cherished beliefs, rather than permitting them to help us identify *new objects of analysis* that may require *new theoretical toolkits*, especially when the latter connects linguistics with cutting edge research in cognitive (neuro)science.

3 A Case Study: Evidence in Aspectual Studies

This case study about aspectual classification will point out certain characteristic difficulties linguists face when trying to weigh up linguistic evidence, i.e. to critically assess its validity and usefulness. For several reasons, aspect is particularly instructive in demonstrating problems with linguistic evidence. First, determining the aspectual types

of "verbs" (both in the sense of lexical elements and syntactic tokens) heavily depends on the availability of subtle semantic interpretations. This is especially true of languages which lack a homogeneous morphological category of aspect so that aspectual categorization appears to be a matter of interpretation rather than a grammatical property as noted for German by Bäuerle (1994: 15). But even in "true aspect languages" such as Romance languages or Modern Greek, ambiguities and tricky semantic alternations are the rule rather than the exception. Second, there is a strand of tradition in aspectology which transcends linguistic boundaries such as "formal vs. functional", "generative vs. non-generative", etc. in a remarkable way. This is the family of time-schema approaches which are first and foremost associated with the work by Vendler (1957/1967) and Dowty (1979): the basic notions of the Vendler/Dowty classification, first mostly used in formal semantics, have been adopted in a wide range of grammatical theories, including generative, functional and cognitive theories (cf. Levin 2000; Foley/Van Valin 1984; Kochańska 2000), and extended from the description of English to a likewise wide range of languages. Despite conceptual differences and differences in formal elaboration, the respective approaches thus have a common core of semantic distinctions, or at least a common starting point (the four Vendlerian categories), and a common methodological tool (the Dowtyan tests, which are frequently adopted, when possible, or used as the basic model for further tests). This leads to a striking similarity among linguistic data presented as evidence in different frameworks, which, in turn, allows us to focus on the methodological quality of such pieces of evidence from a more general perspective, by abstracting from the specific theoretical and representational differences in the literature for the current purpose.[16]

Aspectual classification à la Vendler and Dowty is partly based on the compatibility of verbs with certain adverbs, with so-called "aspectual verbs" (e.g. *finish*) or with grammatical constructions (tense/aspect morphology) and partly on the resulting sentence meaning if such cooccurences are allowed. Typical statements concerning the compatibility of verbs are:

---

[16] Sasse (2002) provides a very good overview of the history of aspectual theories and the most important controversies within current models (cf. also Kabakčiev 2000). Behrens (1998) contains a detailed study of theoretical and methodological problems usually involved in the application of aspectual tests. For an exhaustive assessment of different inventories of situation types see Tatevosov (2002).

(1) a) Punctual (or "instantaneous", "momentary") verbs cannot be combined with adverbials which imply durations.

b) Stative verbs cannot be used in the progressive.

(1a) is then usually illustrated by sentences such as (2a) (cf. Smith 1997: 42; cf. also Van Valin/LaPolla (1997: 95), who adduce the very same sentence, or Verkuyl (1993: 42) and Van Geenhoven (2005: 111), the latter using a slightly modified variant of (2a) to illustrate (1a)). (2b) shows a popular example for illustrating (1b) (cf. Smith 1997: 40), frequently adduced not only in aspectual studies but also in descriptive grammars and introductory books on semantics (cf. Saeed 1997: 119).

(2) a) ?#*The bomb exploded slowly* (cf. Smith 1997: 42)[17]

b) **Kim is knowing the answer* (cf. Smith 1997: 40)

Searching through corpora, however, one finds surprisingly many attestations of *explode* in combination with *slowly*, and also a good number of uses of *know* in progressive. In (3) and (4), we see some of the attestations I found by Google search:

(3) a) The bombs kept moving, they struck the Hecate [a ship; LB]...The Hecate **exploded slowly**, fire ripped through every deck, systems crashed, and the hull ripped itself apart.

[17] Smith makes a distinction between "semantically ill-formed" sentences, marked with a hash (#) as in (2a), and "grammatically ill-formed" sentences, marked with an asterisk (*) as in (2b). For many other linguists, (2a) is simply ill-formed just like (2b) and, as such, it is marked with an asterisk. Unfortunately, Smith does not explain how she understands the difference between these two categories of ill-formedness. Some of her examples give the impression that the difference in marking is simply a matter of badness, sentences marked with an asterisk being considered as somehow worse. In other cases she seems to make a distinction between those diagnostic elements of the aspectual tests which are a) traditionally treated as part of English grammar (i.e. progressive) and b) lexical elements (i.e. adverbs). The use of progressive with stative verbs violates a well-known prescriptive rule of traditional English grammar but less so the use of certain adverbs with certain verb classes. The term "semantic ill-formedness" seems to me particularly misleading in the current case, since the relevant adverb systematically triggers a process reading, always leading to a reasonable utterance meaning. What is not meaningful or "semantically ill-formed" here cannot thus be the resulting sentence but only an artifact of a simple compositional model, according to which *explode* retains its punctual reading when combined with *slowly*. For a possible difference between the two cases suggested by corpus data compare footnote 19.

b) The Romulan ship **exploded slowly** from its warp core outward, ripped apart at the seams…

c) My UrQ spoiler got some moisture inside the foam and then **exploded slowly**.

d) The monitor crackled as it **exploded slowly**.

e) He said it looked like a slow moving firework that even **exploded slowly**, spreading this odd light everywhere.

(4) a) We bristled at the quip that research "**was knowing more and more** about less and less."

b) Professor Diamond: Let us remember, RCUK [Research Councils UK] is a relatively smooth coordinating organization and what RCUK is doing is providing the essential glue that enables that interaction to take place, and over time will, I believe, give a focal point for RDAs [Regional Development Agencies] to come in, and I think they **are knowing** that **increasingly** over time. (from the Homepage of the UK Parliament)

c) We **were knowing** of nothing **when** we hired him, we are all very surprised", she said.

Now the basic question is: how should we, linguists, treat such corpus attestations? Do they represent counterevidence against the validity of (1a) and (1b)? Or do they invalidate only the underlying assumptions that *explode* would be a "true" punctual verb and *know* a "true" state verb (with the consequence that (1a) and (1b) might be correct but (2a) and (2b) are infelicitous examples for illustrating them)? Or do such corpus examples even point to a deeper inconsistency in traditional linguistic practice concerning the relation between a) concepts such as "ill-formedness", b) the meaning of the notation of judgments (i.e. the meaning of *, #, ?, etc.), and c) the way of how linguistic tests are applied?

In order to answer these questions let us first think about how it does come that Carlota Smith (like many other linguists) rejects (2a) and (2b), by marking them by a hash or an asterisk, although similar sentences may be easily found in corpora?

(5) a) Did she undeliberately overlook some relevant contexts when judging (2a), e.g. by imagining a situation in which smaller bombs are involved (whose explosion might be perceived as lasting a very short time) rather than big objects such as atomic bombs, planets, big ships, etc.? Interestingly, it is frequently with such big objects that an explosion is described in corpora as happening "slowly" (cf. (3a, b); but cf. also (3c, d), which describe the explosion of a spoiler and a monitor).

b) Or is Smith's rejection of (2a) motivated by certain extralinguistic assumptions (e.g. by the belief that there is a language-independent concept of "explosion" which would refer to the ignition point) rather than by true introspection about how the English verb *explode* might actually be used. Note that in a large part of the attested examples in which the English verb *explode* is combined with *slowly*, reference is made to a situation which not only includes the ignition point but also the following effects of destruction.

c) There is a third possibility of how to interpret the rejection of (2a) and also that of (2b): in this case, the notation signs (# and *) do not simply mark the ill-formedness of these sentences in principle. Rather, they refer to judgments as if they were made under certain fixed readings of the verbs (the punctual reading of *explode* and the state reading of *know*) – the fact that the same verbs may occur with process readings in certain contexts, as shown in (3) and (4), will be, in this case, quite recognized but these readings will be considered as "derived". So the well-formedness of the corresponding sentences under these "derived" readings is treated as if it would be a matter of separate judgments. This strategy of presenting judgments involves a prior analysis of lexical ambiguity, meaning shift, markedness, etc., on the basis of which it is decided that, from the possible readings a verb may have in context, one is to be considered as a "primary" reading and all other ones as "derived" readings. However reasonable such analyses may be, they deserve their own independent evidence. More importantly, such pseudo-judgments of linguists do not reflect their actual intuition about the acceptability of linguistic expressions. As such, they have a status basically different from the ordinary judgments we can directly receive from naïve native informants.

It is not possible for me to exactly decide which of these three possible factors (lack of imagination, ontological beliefs replacing results of true introspection, pseudo-judgments reflecting additional linguistic analysis) play the most decisive role in the particular cases we are discussing. In general, all of them may come into play simultaneously, contributing to a situation in which the cited judgment data are difficult to interpret, i.e. difficult to confirm or discard, or classify as irrelevant.

Lack of imagination (cf. (5a)), i.e. the inability of native speakers to simultaneously imagine a huge range of contexts while judging the acceptability of linguistic expressions is, in my view, a problem still underestimated in linguistic methodology. According to current research, introspective data are complex performance data (cf. Schütze 1996; Kepser/Reis 2005b; Behrens 1998). Using the conventional terminology (which associates competence with grammaticality and performance with acceptability), this means that there is no such thing as direct "grammaticality judgments". What people are doing when judging sentences is evaluating their acceptability with respect to certain contexts they just have in mind. Generalizing over acceptability judgments might certainly be considered as a method of reconstructing competence; it should, however, be clear that it is much easier to do so in strongly grammaticalized and not (or only weakly) context-sensitive areas of investigation. I think that every linguist working on a highly context-sensitive area such as aspect is familiar with the phenomenon that native informants tend to repeatedly change their opinions about the same linguistic expressions because new situational contexts cross their minds over and over again. In such a case, the relevant data are not single judgments but a set of judgments, each judgment being connected to a particular contextual condition.

Ringen (1980: 117) and others have pointed out that the notion of "clear cases" applied to reported judgments allows two different interpretations: a) the judgments are shared by all (or a significant sample of) informants (no interpersonal variation) and b) the judgments of a single informant remain constant over time (no intrapersonal variation). Whereas interpersonal variation has always been attracted much interest, potentially corresponding to sociolinguistic variation in production data, intrapersonal variation has never been taken seriously. Rather, it is mostly used as a general argument against the introspective method. However, when judgments would be controlled systematically for biasing contexts and when such conditions would be reported regularly with all judgments used as linguistic data, it could turn out that many of the unclear cases due to intrapersonal variation are rather

clear and systematic. It could even be the case that some of the interpersonal divergencies in judgments are only apparent divergencies, going back to the fact that in comparing judgments we compare apples with oranges, i.e. what we compare is judgments made under different contextual assumptions. Such considerations point to the possibility that there is nothing wrong with introspection per se but only with the way it is usually applied as a method of measurement in linguistics. To make a comparison with other empirical sciences: imagine a chemist would not take care of certain variables such as temperature (or s/he simply would not report them), although s/he knew that these variables have a crucial impact of the measurement results.[18]

An old argument against corpus research was that corpora are basically incomplete, however large they are. Rather than capturing the principled distinction between what is possible and what is not possible in a language, they would "only" reflect frequencies. Ironically, a similar objection might be made against introspection, considered as a heuristic tool for acquiring linguistic knowledge. It is basically afflicted with incompleteness. When employed at a certain time, introspection reveals only selected aspects of the complex knowledge of speakers, corresponding to their linguistic and extralinguistic preferences. Corpus data and introspective data complement each other also in the sense that the former may provide crucial information about relevant patterns and relevant contexts, which are in their entirety not accessible to single speakers. By generating new and more specific hypotheses corpora can thus improve the application on the introspective method.[19]

---

[18] Of course, there are many linguists who work with very sophisticated eliciting techniques, regularly providing informants with several background contexts. The problem is that, outside of psycholinguistics and perhaps also outside a new discipline within linguistics called "language documentation", linguists are not used to systematically record and/or publish the details of their elicitation or those of their own introspection.

[19] My corpus data suggest a difference between the combinability of *explode* with *slowly* on the one hand and that of *know* with progressive morphology on the other hand. We find such a broad range of uses of *explode* in combination with *slowly* that it cannot be considered exceptional in any sense. In contrast, *know* occurs in progressive only in a few linguistically restricted types of contexts. The most prominent context types I found are: a) reference is made to a gradual change of state (knowing increasingly more and more; cf. (4a, b)) and b) reference is made to a temporary state which is bounded, as a background state, by another event (being in the state of knowing, when…; cf. (4c)). This second context represents a classical case of taxis-relatedness, i.e. the "incidence case", in which "true aspect languages" (e.g. Modern Greek, Russian) normally use an imperfective form (cf. Sasse 2002), and which is also reported to be a favorable condition for the use of progressive as an

Expressing ontological assumptions in form judgments (cf. (5b)), rather than relying on the actual intuition about the tested sentences, deprives us of the fundamental sense of linguistic tests. The reason why aspectual tests are necessary is exactly the point that there is no universally neutral or basic association between situations in the world and linguistically realized situation types. Just because languages might show varying lexicalization patterns and varying patterns of aspectual alternations we do need controlling devices such as tests in order to deliminate language-specifically entailed meanings (i.e. the meanings of particular expressions in particular languages) from language-independent ontological assumptions.

There is a widespread linguistic practice to mark sentences as ill-formed only with respect to one possible interpretation and then to make a sidenote (e.g. in a footnote) that the same sentence can be well-formed under some other interpretations (cf. (5c)). Carlota Smith, for instance, remarks for (2b) that "some speakers allow sentences like [this] on the ingressive reading, in which the sentence presents preliminary stages of an explosion" (Smith 1997: 42; cf. also Smith 1996: 261).

As early as 1973, Householder mocked the "oddity" of asterisked sentences which are said to be ungrammatical (or unacceptable) only in a particular sense, emphasizing that here, "we are judging a structure-meaning pair, instead of just a structure" (Householder 1973: 369). However, this is probably what naïve (non-linguist) speakers are always doing when asked to judge linguistic forms, namely judging form-meaning pairs rather than judging forms alone, so that rejecting a linguistic form entirely could simply represent an extreme case, i.e. the case where informants are not able to associate the tested form with any intelligible meaning. At least aspectual tests are generally form-meaning tests, which is somewhat obscured by their frequent labeling as "syntactic and semantic tests" (cf. Foley/Van Valin 1984: 37), as if there were two types of tests which could be used independently of each other. However, when looking at the tests cited in Dowty (1979: 60), we

---

emergent category in some languages, e.g. in Estonian (cf. Metslang 1995). In this respect, the use of progressive in these contexts fits very well cross-linguistically generated expectations. What the corpus does not reveal and what should be tested with the introspective method is the question whether the progressive alone can license a process reading in English (like *slowly* with *explode*) or a temporary state reading (like progressive with predicates such as *be silly*) or whether an additional explicit verbal context triggering such readings (e.g. *more and more*) is necessary for the progressive to be accepted.

will soon realize that several of them constitute ordered pairs of tests[20], jointly contributing to a final classification.[21] The first test (the "syntactic" or cooccurence test) captures the prior distinction between the extreme case (no meaning available) and meaningful results, this way excluding situation types manifesting the extreme case. The second test ("semantic" or interpretative test) proceeds from the positive results of the first test and makes further divisions on the basis of the available meanings.

Languages may vary on a continuum between a more permissive and a more restrictive type: in the first case, most verbs will occur in all (or a variety of) diagnostic environments with alternating interpretations; here only interpretative tests will be of interest. In the second type of languages, we will find both true cooccurrence gaps and semantic alternations (cf. also Dowty 1979: 62), so that both types of tests have to

[20] Such ordered pairs of tests in Dowty (1979: 60) are test 1 and test 6 (non-stative tests), test 3 and test 5 (*for*-test), and test 4 and test 10 (*in*-test).

[21] It is almost superfluous to mention that we cannot use the same tests cross-linguistically that are presented in Dowty (1979) but, for every language, we have to develop its own battery of relevant tests.

be applied successively[22]. In any case, lexical classification[23] should capture the entire potential of semantic alternations[24].

How then should we interpret side remarks such as that made by Smith, i.e. the remark that "some speakers allow" sentences like *the bomb exploded slowly* on the interpretation in which it refers to the preliminary stage of the event rather than to the event itself (cf. 1996: 261)? Do the well-known scalar implicatures apply here: some but not all speakers? This would mean that marking the sentence as ill-formed reflects the intuition only of part of the speakers involved. Or do all speakers involved accept this sentence under the "preliminary" reading, but some of them find this interpretation perhaps marginal? One is also wondering whether the same speakers would reject meanings in which the sentence refers the subsequent stages following the ignition point (as attested by most of the corpus examples), rather than to preliminary

---

[22] Dowty himself is not always consistent on this question of successive application. According to the test logic, as described above, a negative result with the first type of test should correspond with a "does not apply (d.n.a.)" for the second type of test. In spite of this, Dowty sets a question mark for "achievements" in the case of the first test (non-stative tests) and characterizes the subsequent test (test 6, progressive) as "d.n.a." for achievements. He justifies the question mark (instead of "not") with pointing out that some verbs classified as "achievements" by Vendler (e.g. *die*, *arrive*) do "not really sound so bad" (1979: 130) when used in progressive. From the perspective of lexical classification, the question is, of course, whether all potential achievement verbs in English allow the use of progressive in the relevant reading (preparatory phase), similar to *die*. If the answer is "yes", we have to use further interpretative tests for distinguishing the class of achievement from other classes such as activities. If no, we have to distinguish between two subclasses within the Vendlerian domain of achievements.

[23] The following point should be emphasized: the Vendlerian/Dowtyan tests use the logic of lexical tests: they test the lexical disposition of verbs and the lexically established disposition of abstract phrases (e.g. *build a house* as a phrase abstracted from inflection). As such, these tests proceed from the utterance meaning like many other lexical tests which try to reconstruct lexical meaning on the basis of utterance meaning (e.g. many ambiguity tests). Applying lexical tests has, however, no implication for the main controversies in aspectology: are situation types lexically established or only at the level of syntax or on both levels, and if the latter is correct, should we prefer a "recategorizing" model, a "selecting" model or another model in order to describe the mapping between the lexicon and the syntax? In any case we need lexical tests and the adequacy of a model may also depend on the language type.

[24] Some authors recognize that plural subjects may trigger an iterative interpretation of *explode slowly* (one bomb explodes after the other) (cf. Van Geenhoven 2005: 11). This type of quantification alternation differs however from that one responsible for the attestations in (3). Actually, we find corpus attestations of *explode slowly* with plural subjects in which things distributively explode slowly, e.g.: *The 2 bombers both exploded slowly as the atmosphere on the inside combusted,...*

stages? Or does the side remark in question only serve the didactic purpose of informing the reader of the own awareness of possible interpretations, which, however, are considered altogether irrelevant in the context of the argumentation, because the negative judgment only targets that sense of *explode* which is reconstructed as the primary one on the basis of the overall analysis of English?

Of course, lexical classification should go beyond the plain description of the syntagmatic distributional patterns, in that it should also account for paradigmatic properties such as recurrence and frequency in the whole vocabulary, capturing the difference between systematic and idiosyncratic, directional and non-directional, etc. alternations. Corresponding hypotheses about the best macroanalysis of the lexicon may be supported by different types of higher-order evidence, e.g. by being based on series of experiments on interpretational preferences and markedness across the vocabulary, and, of course, also by being based on series of aspectual tests of distributional style. However, in order to avoid circularity, it should not be allowed that such complex hypotheses about the entire lexicon influence what is intended to provide evidence at a lower epistemological level (i.e. the results of judgments).[25]

Doing so would disqualify judgment data as "evidence" in the sense of publicly trustworthy pieces of knowledge: it would render them even more unintelligible for other researchers, beyond the notorious problem of context-dependency as discussed above, and as such it would also make them incomparable both to other kind of data sources (e.g. corpus data) and to truly elicited judgment data from non-linguist informants. As a consequence, this insecurity in assessing judgment data would be passed on to the assessment of complex argumentations which use such data, and ultimately to the assessment of competing theoretical models

[25] Such problems of circularity are perhaps also rooted in the more general problem that modern linguistics in many cases inappropriately rely on structuralist methods. It has been repeatedly pointed out by many linguists such as Lyons (1991: 17) that we leave the domain of distributional analysis when we start to operate with notions such as "normal", "marked", etc. Even though the Chomskyan revolution brought a renunciation from the structuralist paradigm, it did this only in philosophical but not in methodological respect. This may be connected, paradoxically, to the theoretically-motivated antipathy against the "discovery procedures" of the old paradigm. So the distributionalist techniques of structuralists continued to be used as basic methods in generative linguistics, as well as outside of it, without being refined and adjusted to the new theoretical requirements and also without being purged of inherent inconsistencies.

(cf. footnote 23) which, in turn, are supported by corresponding arguments from particular languages.

4 Conclusion

The problems discussed in the previous chapter do by no means invalidate introspection as a useful tool of data acquisition. In many areas such as verbal aspect, we are compelled to rely on native speakers' intuition. Apart from this, every data collection method counts, and every kind of hypothesis that is based on a certain type of data should be checked against other data sources, so hypotheses based on natural discourse data against the intuition of native speakers just as much as vice versa. It is rather the manner of how introspective data are traditionally treated in terms of evidence in linguistic argumentation that appears to me unsound. But this problem is part of a more general issue, i.e. the way evidence is dealt with in the actual working context of linguistics. As emphasized in this paper, the current challenge in linguistics is more than ever the question of how to reach a sufficient level of commonly accepted standards of evidence (beyond different interests in and different ontological claims about language) which apply both crosslinguistically and crossdisciplinarily and thus allow comparing and integrating incomplete pieces of knowledge from different fields. It is certainly not theory-dependency that hampers successful knowledge integration but the still low degree of interest in many branches of linguistics (cf. above pp. 17ff.) to make transparent and publicly controllable the implicit theoretical and methodological assumptions and restrictions which are involved when something is presented as evidence. Rhetorically appealing to "evidence" as if it were a "hard proof" when it is no more than a "weak hint" may work in persuading those who are already "on board" but hardly in convincing skeptics. In the long run, this strategy of narrowing the group to which evidence is presented has the consequence, as argued by Carlson (2003: 78f.), that "what can be satisfactorily shown" would become "less and less connected to the grounding of common standards of evidence in society at large, and also less and less connected to the standards of evidence invoked across academics in general". My impression is that agreeing with this statement and urging the vital necessity of new methodological devices is not a matter of belonging to an established subculture such as "generative" and "functional" linguistics but something that cross-cuts

these old divisions. This too would be a really interesting aspect in the new epistemological debate.

5 References

Ackerman, Farrell (2007). *Construction Theoretic Approaches to Grammar: Lessons for Syntax from Wordbased (Word and Paradigm) Morphology. Szuperkurzus – June 4-8, 2007.* MTA Nyelvtudományi Intézet, Budapest. (http://www.nytud.hu/program/szuperkurzus2007jun/budahandoutfinal.pdf, 28.7.2007)

Aissen, Judith / Bresnan, Joan (2004). Remarks on Description and Explanation in Grammar: Commentary on Haspelmath. In: Penke, M. / Rosenbach, A. (eds.) (2004a). 580-583.

Arppe, Antti / Järvikivi, Juhani (forthcoming). Every Method Counts - Combining Corpus-based and Experimental Evidence in the Study of Synonymy. Corpus Linguistics and Linguistic Theory. (http://www.ling.helsinki.fi/~aarppe/Publications/CLLT_Arppe_Jarvikivi-07.pdf)

Ballmer, Thomas T. (1976). Inwiefern ist Linguistik empirisch? In: Wunderlich, Dieter (ed.). *Wissenschaftstheorie der Linguistik*. Kronberg: Athenäum Verlag. 6-53.

Bäuerle, Rainer (1994). Zustand - Prozeß - Ereignis. Zur Kategorisierung von Verb(alphras)en. *Wuppertaler Arbeitspapiere zur Sprachwissenschaft 10.* Wuppertal: Bergische Universität-Gesamthochschule Wuppertal. 1-32.

Behrens, L. (1998). *Ambiguität und Alternation. Methodologie und Theoriebildung in der Lexikonforschung.* Habilitationsschrift, München.

Boas, Franz (1911). Introduction. In: Boas, Franz (ed.). *Handbook of American Indian Languages* (Vol. 1). Smithsonian Institution, Bureau of American Ethnology, Bulletin 40. Washington: Government Printing Office. 1-83.

Camazine, Scott / Deneubourg, Jean-Louis / Franks, Nigel R. / Sneyd, James / Theraulaz, Guy / Bonabeau, Eric (2001). *Self-Organization in Biological Systems*. Princeton University Press.

Carlson, Greg (2003). On the Notion 'Showing Something'. In: Moore, John / Polinsky, Maria (eds.). *The Nature of Explanation in Linguistic Theory*. Stanford, CA: CSLI. 69-82.

Chomsky, Noam (1993). A Minimalist Program for Linguistic Theory. In: Hale, Kenneth / Keyser, Samuel Jay (eds.). *The View from Building 20:*

Essays in Honor of Sylvain Bromberger. Cambridge, MA: The MIT Press. 1-52.

Chomsky, Noam (2000). *The Architecture of Language*. New Delhi: Oxford University Press.

Comrie, Bernard (1981). *Language Universals and Linguistic Typology. Syntax and Morphology.* Oxford: Blackwell.

Croft, William (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Dahl, Östen (1980). Is Linguistics Empirical? A Critique of Esa Itkonen's Linguistics and Metascience. In: Perry, Thomas A. (ed.) (1980). 133-145.

Devitt, Michael (2006). *Ignorance of Language*. Oxford: Clarendon Press.

Diekmann, Andreas (2005). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Dixon, R.M.W. (1997). *The rise and fall of languages.* Cambridge: CUP.

Dowty, David R. (1979). *Word Meaning and Montague Grammar*. Dordrecht: D. Reidel.

Everett, Daniel L. (2005). Biology and language: a consideration of alternatives. *Journal of Linguistics* 41: 157-175.

Everett, Daniel L. (2007). *Cultural Constraints On Grammar In Pirahã: A Reply to Nevins, Pesetsky, and Rodrigues.* MS. (http://ling.auf.net/lingBuzz/000427)

Featherston, Sam (2005). The Decathlon Model of Empirical Syntax. In: Kepser, S. / Reis, M. (eds.) (2005a). 187-208.

Foley, William A. / Van Valin, Robert D. (1984). *Functional Syntax and Universal Grammar*. Cambridge et al.: CUP.

Gerritsen, Susanne (2001). Unexpressed Premises. In: van Eemeren, Frans H. (ed.). *Critical Concepts in Argumentation Theory*. Amsterdam: Amsterdam university Press. 51-79.

Gilligan, G. (1987). *A Cross-Linguistic Approach to the Pro-Drop Parameter*. Doctoral dissertation, USC, Los Angeles.

Haspelmath, Martin (2004). Does Linguistic Explanation Presuppose Linguistic Description? In: Penke, M. / Rosenbach, A. (eds.) (2004a). 554-579.

Hoffmann, Robert R. (1989). Some Ambiguities in the Study of Ambiguity. In: Gorfein, D.S. (ed.). *Resolving Semantic Ambiguity*. New York et al.: Springer. 204-222.

Justeson, John S. / Katz, Slava M. (1991). Redefining Antonymy: The Textual Structure of a Semantic Relation. In: *Using Corpora. Proceedings of the Seventh Annual Conference of the UW Centre for the*

New Oxford English Dictionary, September 29 - October 1, 1991, Oxford, England. 138-153.

Kabakčiev, Krasimir (2000). *Aspect in English*. Dordrecht et al.: Kluwer.

Kepser, Stephan / Reis, Marga (eds.) (2005a), *Evidence in Linguistics. Empirical, Theoretical and Computational Perspectives*. Berlin / New York: Mouton de Gruyter.

Kepser, Stephan / Reis, Marga (2005b). Evidence in Linguistics. In: Kepser, S. / Reis, M. (eds.) (2005a). 1-6.

Kochańska, Agata (2000). Verbal Aspect and Construal. In: Foolen, A. / van der Leek, F. (eds.). *Constructions in Cognitive Linguistics*. Amsterdam/Philadelphia: Benjamins.

Kornai, András / Pullum, Geoffrey K. (1990). The X-bar Theory of Phrase Structure. *Language* 66 (1): 24-50).

Kuhn, Thomas S. (1976). *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt/M.: Suhrkamp Taschenbuch Verlag.

Legate, Julie Anne (2002). *Warlpiri: Theoretical Implications*. Doctoral dissertation, MIT, Cambridge, MA.

Levin, Beth (2000). Aspect, Lexical Semantics, and Arguments Expression. In: Conathan, Lisa J. et al. (eds.). *General Session and Parasession on Aspect (BLS 26)*. Berkeley, CA: BLS. 413-429.

Mayr, Ernst (2004). *What Makes Biology Unique?: Considerations on the Autonomy of a Scientific Discipline*. Cambridge: CUP.

Metslang, Helle (1995). The Progressive in Estonian. In: Bertinetto, Pier Marco / Bianchi, Valentina / Dahl, Östen / Squartini, Mario (eds.). *Termporal Reference, Aspect and Actionality. Vol. 2*. Torini: Rosenberg & Selier. 169-183.

Newmeyer, Frederick (1998). *Language Form and Language Function*. Cambridge, MA / London: The MIT Press.

Newmeyer, Frederick (2004). Typological Evidence and Universal Grammar. In: Penke, M. / Rosenbach, A. (eds.) (2004a). 527-548.

Newmeyer, Frederick (2005). *Possible and Probable Languages. A Generative Perspective on Linguistic Typology*. Oxford: Oxford University Press.

Ouhalla, Jamal (1999). *Transformational Grammar: From Principles and Parameters to Minimalism. (2nd Edition)*. London: Arnold.

Penke, Martina / Rosenbach, Anette (eds.) (2004a). *What Counts as Evindence in Linguistics? The Case of Innateness.* Special Issue. Studies in Language 28 (3).

Penke, Martina / Rosenbach, Anette (2004b). What Counts as Evindence in Linguistics? An Introduction. In: Penke, M. /Rosenbach, A. (eds.) (2004a). 480-526.

Perry, Thomas A. (ed.) (1980). *Evidence and Argumentation in Linguistics*. Berlin / New York: Walter de Gruyter.

Popper, Karl R. (1972). *Objective Knowledge. An Evolutionary Approach*. Oxford: Clarendon Press.

Popper, Karl R. (2004). *Ausgangspunkte. Meine intellektuelle Entwicklung*. München/Zürich: Piper (Piper Serie).

Pullum, Geoffrey K. / Scholz, B.C. (2002). Empirical Assessment of Stimulus Poverty Arguments. *The Linguistic Review* 19: 9-50.

Radford, Andrew (1988). *Transformational Grammar. A First Course*. Cambridge: CUP.

Rieber, Robert W. (1983). The Psychology of Language and Thought. Noam Chomsky interviewed by Robert W. Rieber. In: Rieber, R.W. (ed.). *Dialogues on the Psychology of Language and Thought*. New York: Plenum.

Ringen, Jon D. (1980). Linguistic Facts. A Study of the Empirical Status of Transformational Generative Grammars. In: Perry, Thomas A. (ed.) (1980). 97-132.

Saeed, John I. (1997). *Semantics*. Oxford: Blackwell.

Sasse, Hans-Jürgen (1993). Syntactic Categories and Subcategories, In: Jacobs, Joachim / von Stechow, Arnim / Sternefeld, Wolfgang / Vennemann, Theo (eds.). *Syntax. Ein internationales Handbuch zeitgenössisches Forschung*. (HSK 9) Berlin / New York: de Gruyter. 646-686.

Sasse, Hans-Jürgen (2002). Recent Activity in the Theory of Aspect: Accomplishments, Achievements, or just Non-Progressive State? *Linguistic Typology* 6: 199-271.

Schachter, Paul (1976). The Subject in Philippine Languages: Topic, Actor, Actor-Topic or None of the Above? In: Li, Charles N. (ed.). *Subject and Topic*. New York: Academic Press. 491-518.

Schütze, Carson T. (1996). *The Empirical Base of Linguistics. Grammaticality Judgments and Linguistic Methodology*. Chicago / London: The University of Chicago Press.

Smith, Carlota S. (1997). *The Parameter of Aspect (Second Edition)*. Dordrecht et al.: Kluwer.

Smith, Carlota S. (1996). Aspectual Categories in Navajo. *IJAL* 62(3): 227-263.

Tatevosov, Sergej (2002). The Parameter of Actionality. *Linguistic Typology* 6(3): 317–401.

Tomasello, Michael (2004). What Kind of Evidence Could Refute the UG Hypothesis? Commentary on Wunderlich. In: Penke, M. /Rosenbach, A. (eds.) (2004a). 642-645.

Van Geenhoven, Verle (2005). Atelicity, Pluractionality, and Adverbial Quantification. In: Verkuyl, Henk / De Swart, Henriette / Van Hout, Angeliek (eds.). *Perspectives on Aspect*. Dordrecht: Springer. 107-124.

Van Valin, Robert D. (1990). Functionalism, Anaphora, and Syntax. Review of Functional Syntax, by S. Kuno. *Studies in Language* 14: 169-219.

van Eemeren, Frans / Grootendorst, Rob (2004). *A Systematic Theory of Argumentation. The Pragma-Dialectal Approach*. Cambridge: CUP.

Van Valin, Robert D. / LaPolla, Randy J. (1997). *Syntax. Structure, Meaning, and Function*. Cambridge: CUP.

Vendler, Zeno (1957/1967). *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.

Verkuyl, Henk (1993). *A Theory of Aspectuality. The Interaction between Temporal and Atemporal Structure*. Cambridge: CUP.

Wanner, Eric (1977). Review of "The psychology of language, by J. A. Fodor, T. G. Bever, & M. F. Garrett." *Journal of Psycholinguistic Research* 6: 261-80