CLIPP

Christiani Lehmanni inedita, publicanda, publicata

titulus

Data in linguistics

huius textus situs retis mundialis

http://www.christianlehmann.eu/publ/lehmann_data_in_linguistics.pdf

dies manuscripti postremum modificati

28.06.2004

occasio orationis habitae

Colloquio internazionale 'Di cosa parliamo quando parliamo di linguistica?', Roma 1-2 luglio 2002

volumen publicationem continens

The Linguistic Review 21(3/4)

annus publicationis

2004

paginae

275-310

Data in linguistics

Christian Lehmann

University of Erfurt

Περὶ τῶν ἀφανέων περὶ τῶν θνητῶν
σαφήνειαν μὲν θεοὶ ἔχοντι,
ὡς δὲ ἀνθρώποις τεκμαίρεσθαι.[1]
Alcmaeon of Croton (Diogenes Laertius VIII, 83)

Die Theorien vergehen,
aber das Material bleibt bestehen.[2]
Einar Löfstedt 1942: IX

Le donné linguistique est un résultat;
et il faut chercher de quoi il résulte.[3]
Émile Benveniste 1954(1966):117

Abstract

This article aims to be a contribution to the methodological foundations of linguistics. To answer the question of "what are scientific data?", a semiotic conception of data is proposed according to which they are representations of properties of the object area of a science that serve certain purposes for their users. Kinds of data are distinguished by their ontological status, degree of abstractness, the type of sign representing them and their originality. The methodological status of data in the history of linguistic science is briefly reviewed, and their functions in scientific argument are specified. Various methods of data provision by generation of data or by use of available data are discussed. Since data are representations, they are per se a linguistic issue which, however, is even more complicated for linguistic data proper, because here diverse linguistic levels and diverse levels of abstractness have to be controlled.

Apart from the principal necessity to have clarity on the methodological bases of a science, the issue of the nature and function of data in linguistics acquires increased urgency in a world where the task of documentation of endangered languages is, first and foremost, one of adequate data provision.

---

[1] On invisible and on earthly things, gods have clarity, while to men it is given to infer.

[2] Theories pass away, but the material lasts.

[3] The linguistic datum is a result; and we have to search what it results from.

1. Introduction

Since the etymological meaning of the word *data* is "(things that are) given," it is probably an instance of "nomen est omen" that the notion and role of data in science are generally taken for granted. A representative sample of contemporary publications on methodology and philosophy of science – introductions, manuals and lexica – reveals that the term nowhere constitutes a lemma and the concept is nowhere introduced explicitly, let alone defined. In linguistics itself, endeavors to clarify the role of data are very recent and as yet few.[4]

We will first try to elucidate the notion of data and define it. The conception proposed is essentially a semiotic one: data – not just linguistic data, but any data – will emerge as a certain kind of representation of an object. We will then look at essential properties of data in linguistics and their role in the scientific process, systematize the most important ways of obtaining data and finally come to the genuinely linguistic issue of representations of data. The article is thus a contribution to the methodology of linguistics.

2. The notion of data

2.1. The term

In order to be able to speak about data, we first have to emend the English language a bit. Since the middle of the 20th century, it has become customary in English to use the word *data* as a mass noun. As a consequence, it does not pluralize nor combine with an indefinite article, but instead combines with mensuratives like *a piece* (*of data*). For the sake of the following discussion, we shall undo this linguistic change and reestablish traditional usage: For the concept in question, there is an individual noun *datum* which forms a morphologically irregular, but semantically regular plural *data*. The latter thus does not designate a mass, but a set of individuals.

Secondly, we have to do a linguistic analysis of the word *datum*. This is a relational and, therefore, functional concept in the sense that something is not in and by itself a datum, just as something is in and by itself a sentence. Instead, something functions as a datum for somebody. This is explicable both in terms of scientific methodology – and to this we will come in section 4 – and in terms of the etymology of the word, to which we now turn. The etymological meaning of *datum* is "given," in roughly the same sense which is still productive in modern English, for instance in the complex conjunction *given that*. This conjunction introduces something whose existence and nature is independent from the deictic center and which the deictic center cannot but accept as it is. The extensional meaning of the word *datum* is inde-

---

[4] The largest relevant enterprise is probably the DFG Sonderforschungsbereich 441 on 'Linguistic data structures' at the University of Tübingen. Two independent contributions may be mentioned: Iannàccaro (2000) concentrates on the interference of theory in fieldwork, and Simone (2001) deals critically with several methodological aspects of linguistics, including, in particular (section 2.5.2), the role of data in linguistics. I thank Gabriele Iannàccaro for helpful discussion.

terminate insofar as it designates anything that is "given" (in the relevant sense), independently of its particular nature. Moreover, if *datum* means "given," this, of course, evokes the argument frame of the verb "give": the giver, in this case the producer or source of the datum; the recipient, in this case the discoverer or user of the datum; and the transferred object, in this case the entity which constitutes a datum. At first sight, it might appear that the first two entities of this argument frame are irrelevant for the scientific concept of data. This is, however, not so. Maybe in linguistics more than in many other sciences, it is an absolutely crucial issue who produces the data and who receives them. We will come back to this in section 4.2.

2.2. The concept

In order to clarify the methodological status of data, it is useful to start with the ontology of naïve realism (cf. Lyons 1977, ch. 11.3): First-order entities comprise physical objects; second-order entities comprise states, processes, events and the like. These two kinds of entities are located in space and time and are observable. Third-order entities comprise abstract entities such as propositions, which are not bound to space and time and are unobservable.

A fact is a third-order entity – a proposition – which corresponds to a certain second-order entity. By virtue of this correspondence, the proposition receives the predicate "true." Particular propositions in a science may have the status of a fact. For instance, that Caesar once wrote *veni, vidi, vici* is a fact, or is taken as a fact, in some particular sciences as well as in our civilization in general.

The object area of a discipline is what that discipline wants to learn about. The object area of an empirical discipline has, so we assume whether we are constructivists or not, what will here be called an ultimate substrate, i.e. a basis in the world surrounding the thinking subject which he can perceive. However, the "real world" is complex and multifaceted, and no human cognition is interested in a true copy of it. Every science construes its object area according to its epistemic interest, by delimiting and idealizing what we can perceive and by distinguishing between relevant and irrelevant aspects of these percepts. The object area even of an empirical science (see section 4.1 for the other kinds of science) is, thus, not part of the physical world, but is a mental representation of part of the world. It does not consist of first and second-order entities, but of third-order entities. The object area of linguistics is not a set of states of affairs taking place in the outside world, but a set of situations of linguistic communication construed, delimited, purified and focused upon in ways that vary in the history of science, but which never equal or exhaust some physical reality.

A particular research is devoted to an epistemic object (also called "phenomenon"[5]), which is a construct that is part of the object area. Data concern particular aspects of an empirical epistemic object; and by extension the data of an empirical science represent aspects of its object area.

Consequently, data are neither first nor second-order entities. For instance, a particular pot-shard is not, in itself, an archaeological datum, and a tape which has a speech recorded on it is not, qua physical object, a linguistic datum. Instead, a datum corresponds to a fact, i.e. to a third-order representation of a state of affairs which is considered true. A datum therefore has an inner side which is a mental representation of some state of affairs. For scientist B to

[5] 'Phenomenon' may also be opposed to 'datum'; cf. fn. 19.

accept something which scientist A adduces as a datum means for A and B to share some mental representation which both consider part of the object area of their discipline.

Naturally, mental representations are not the form in which the data are transmitted and analyzed in scientific research. Instead, scientific data are processed in the form of semiotic representations, including linguistic representations, of facts.[6] For illustration, let us briefly look at data of a few different disciplines.

The object area of demography is the structure and dynamism of a population. Particular aspects of it, for instance the sex of a particular person that enters some statistics, are publicly observable. The ultimate substrate of the data of the discipline consists in a set of such particular states of affairs involving members of the population. These are converted into mental representations and into symbolic representations of the latter, for instance in a table, each entry of which refers to an individual and each cell of which represents some property of this individual, for instance his sex. Such a table represents a set of data in this discipline.

The data on which dendrochronology builds its theories are series of numbers, each of which represents the width of an annual ring of some tree and is associated with one in a series of years. The cross-sections of the tree may be stored somewhere for measurements, because they constitute the ultimate basis of reference for certain relevant observations. The data, however, are those series of numbers insofar as they represent facts about these objects.

Finally, history has an ultimate basis in reality which comprises such second-order entities as the event in which prime-minister Begin shook hands with president Sadat on Camp David on Sept 17, 1978. Apart from a couple of exceptions, historians have not witnessed these second-order entities. They have, however, recordings of them available, either iconic records like photos or audio tapes or symbolic representations which historians call sources. They are, then, the form that data take in history.

No science can be run without such representations of facts. It will, therefore, not be necessary, in what follows, to always distinguish between a fact and its representation. Since symbolic representations are a genuinely linguistic problem, we will come back to them below (section 6.2).

Table 1 visualizes the concepts and examples introduced so far and adds a column for linguistics, which we will take up below. There is a further concept, viz. material, which has to be distinguished from data and which may be introduced with respect to the first row of Table 1. In its literal sense, the word *material* designates a collection of physical objects. In some scientific disciplines, the word may be used to designate (parts of) the ultimate substrate of their object area, i.e. a set of first-order entities which the data are based on. For instance, for dendrochronology and archaeology, the physical entities mentioned in the first row of Table 1 constitute the material which research observes and from which it starts. The ultimate substrate of such disciplines is available for repeated direct observation by scientists.

---

[6] Unfortunately, the English word *representation* is ambiguous: it can mean either a purely mental entity (German *Vorstellung*) which may or may not correspond to something outside the mind, and it can mean a semiotic entity (German *Darstellung*) in which a perceivable repraesentans can be distinguished from a – mental or semiotic – repraesentatum. The latter is what is meant in saying that data are representations. The former meaning will be expressed, if necessary, by *mental representation*.

For other disciplines including history and linguistics, the real-world entities underlying their object area are not first-order, but second-order entities such as historical events and speech events. These are volatile and therefore cannot constitute any material in the literal sense of the word. In order to be subject to an objective treatment, they first have to be recorded. The recordings are on durable material, but this is material in a different sense, because it is of interest not as a physical object, but only as a representation of the proper object of research. The word *material*, therefore, has a different sense in the natural sciences and in the humanities. In the former, the material is the ultimate substrate of the data; in the latter, the material is a first recording of the data; i.e. it is a particular kind of data. Consequently, the last two rows of Table 1 contain data at different levels of representation.

Table 1. Data and representations in some disciplines

discipline / object	dendrochronology	demography	archaeology	history	linguistics
ultimate substrate	(cross-sections of) tree trunks	population	ruined wall	handshake between two politicians	speech event
epistemic object	chronological position of the tree	gender distribution in the population	the original wall	relationship of the two politicians	utterance in the speech event
original recording	–	–	contemporary mention/drawing of the wall	source mentioning the handshake	video tape of the speech event
derived representation	series of numbers representing measured widths of annual rings	tables and charts representing gender distribution	design of the ruined wall	time-line with major events, including handshake	phonetic transcription of the utterance

Up to now I have proceeded as if data were a special kind of thing. However, what constitutes a datum is not its nature but its function. In the context of empirical scientific research, a datum serves either as the basis for the inductive construction of a hypothesis or as the test for a theorem arrived at deductively. In order to be able to fulfill this function, a datum must have a basis outside of and independent from the researcher.

The issue of the externality and independence of the datum from the researcher will repeatedly occupy us below (cf. especially fn. 14). It essentially means that there must be methods of relating the datum to the ultimate substrate. I will come back to this in section 3.4 and recall here the relationality of the concept "datum," which I introduced by way of the etymology of the term (section 2.1). Nothing is, in and of itself, a datum; instead, it is a datum for somebody (or for a scientific community) in some perspective. Linguist A has a tape which records a story in Yucatec Maya. The recording is A's data. He produces an orthographic representation of the story and publishes it as the result of his research. Linguist B uses A's orthographic representation as data for his grammar of Yucatec Maya, which he publishes as the output of his research. Linguist C is a typologist whose sources of information are grammars.

He uses B's descriptive statements on Yucatec Maya clause structure as data, puts them into a database and arrives at a couple of cross-linguistic generalizations which he publishes as a typology of clause structure.

The example shows that one person's analysis may be another person's data. Something is not a datum by virtue of corresponding to some elementary observation, like the "protocol sentences" of the logical positivists. On the contrary, it may be highly abstract.[7] It may nevertheless function as a datum in some research that assigns it the role of unquestionable evidence in the argumentation. Thus, something is not in and of itself a datum, but it is a datum relative to some particular empirical research.

We must, however, caution against a misunderstanding. The 1970s saw a wave of pragmaticism where scientific concepts were made relative to the scientist who uses them.[8] Now of course, linguists just like other scientists sometimes disagree on whether something counts as a datum or not. The point here, however, is not that the application of a concept to a referent depends on the user. This is a basic semiotic fact which need not be repeated in the definition of each concept. We are not here talking about some pragmatic relativity, but about the intrinsic relationality of the concept. Just as no proposition is, in and of itself, an argument, but may only be used as an argument under certain conditions, so a representation of something is not a datum if taken absolutely, but may function as a datum in a certain research.

Now we can define:

> A datum is a representation$_i$ of an aspect of the epistemic object of some empirical research which$_i$ is taken for granted.

An aspect of x which is taken for granted is a fact about x in the sense defined above. Since a datum is a representation of something, it is a sign. What I propose, thus, is a semiotic conception of the datum; to be sure, not restricted to the linguistic datum, but intended for the datum of empirical research in general. Since a datum is a sign, it may be an icon, an index or a symbol. The examples given in Table 1 include icons, e.g. the videotape of the speech event, and symbols, e.g. the table representing gender distribution. In some sciences there are also indices that count as data, as, for instance, a footprint of a saurian in paleontology.

3. Kinds and properties of linguistic data

Let us now pursue the consequences of this conception in order to see by which particular properties linguistic data differ from data in other disciplines.

3.1. Raw data vs. symbolic representations

Let us first come back to the tape which records a story in Yucatec Maya. The tape is a first-order entity, and the process in which a certain Maya once recorded the story is a second-

---

[7] In Seiffert (1969f), which is a treatise of scientific methodology in general, the term *data* is applied to summary representations of a couple of theories of different scientific disciplines. It is therefore not the case that linguistic data are statements of particular observables, as Chomsky (1964: 28ff.) thinks.

[8] Cf., e.g. the definition of the language universal in Lieb (1974: 494): "A property F is universal in language relative to a person during a time if that person during that time requires that F should be attributed to all languages by any theory of language."

order entity. Now let us set up a research whose epistemic object is this story. Then the second-order entity just mentioned is the ultimate substrate of the epistemic object, whereas the embodiment of the story on the tape is a piece of data which consists in a recording of the ultimate substrate.

As just said before, the tape-recording is an iconic representation of the data in question. In general, photos, audio- and videotapes of speech events, no matter whether recorded in analog or digital technology, are non-symbolic representations of linguistic data, whereas an orthographic or IPA representation of the same event is a symbolic representation.[9] Higher level linguistic analysis of any data commonly presupposes their symbolic representation. In this sense, a recording in the form of a non-symbolic representation constitutes raw data for linguistic research. For most purposes, the first step in the research will consist in their transcription.

The distinction between raw data and symbolic representations must not be confused with the distinction between original recording and derived representation (see the last two rows of Table 1). Some recording is original in the sense that it is not based on another representation. The original recording of a speech event may be a symbolic or non-symbolic representation, and either may be produced by a linguist or a layman. A derived representation of the original recording may conserve its original nature, as when a pure audio record is distilled from a videotape or when a phonetic transcription is converted into an orthographic one; or it may shift it into the other type of representation, as when an audiotape is transcribed or a text is tape-recorded. The production of derived symbolic representations is a typical activity in linguistic data processing, and we will have more to say about it in section 6.2.

3.2. The uniqueness of the linguistic datum

Now consider a particular linguistic datum on that tape, as the occurrence of a certain word on the tape or the fact that it is immediately followed by a certain other word. These data clearly go beyond sheer physical traces on the tape in several respects. The first respect was already mentioned in section 2.2: The fact that word x is immediately followed by word y on that tape is not a first or second-order entity – a physical pattern on the tape –, but instead a mental representation of an aspect of the epistemic object as recorded on the tape.

The other properties of this fact are peculiar to linguistics. First, our identification of a word on the tape presupposes the recognition of a linguistic expression. This is, again, not a set of physical traces on the tape, but instead an abstraction over certain phonetic objects.[10] Second, a semiotic entity has two sides, only one of which is perceivable. The datum in question, however, concerns tokens of certain semiotic entities and consequently the coupling of the significans with its significatum.[11] The identification of even the most elementary linguistic datum therefore presupposes an abstraction and a semiotic operation.

---

[9] Cf. the related notion of discrete vs. analog communication in Watzlawick et al. 1967: 61-68.

[10] This is what Bühler (1933:30-32) calls the 'abstractive relevance of linguistic units.'

[11] This is what Bühler (1933:24) calls the 'semiotic nature of language.'

3.3. Primary and secondary data

Next, let us come back to the example of the typologist who uses the descriptive statements of a Yucatec grammar as the data for his typology. The example was used to show that the notion of data is relative to a purpose. But it also shows, of course, that data may be elementary or abstract to different degrees. For something to be used as a datum in a discipline, it suffices that it be a fact that is empirically relevant to it. It is not necessary that it be an elementary fact. That a certain word on that Yucatec tape is immediately followed by a certain other word may, in a certain perspective and for a certain purpose, be considered an elementary fact which is immediately verifiable by inspection. However, we are used to working with much more complex and abstract kinds of data. In the past fifty years, linguists have gotten used to cascades of example sentences which exhibit a regular structural difference and every other one of which is preceded by an asterisk, accompanied by a commentary in which the linguist adducing the series refers to them as data. Such a linguistic datum is a semiotic object of a higher order. Namely, it is an expression of the object language coupled with a statement of the metalanguage – the latter being highly abbreviated in the form of an asterisk and even its absence – which predicates a certain property over that object. This statement is taken for granted and therefore regarded as a linguistic datum.

Claims made in an empirical science, including both hypotheses arrived at by inductive generalizations over empirical data and theorems derived deductively on the basis of a theory, must be testable on data belonging to its object area. Suppose our typologist launches the (false) hypothesis that if a language has numeral classifiers, it lacks nominal number. We can check this hypothesis for Yucatec Maya on the data contained in the typologist's database, where this language is marked for possession of both numeral classifiers and nominal number. However, we want to go further and verify whether Yucatec does have numeral classifiers. Here we see that although a descriptive statement may be used as a datum at some level, it bears no direct correspondence to anything observable. We therefore have to distinguish between primary and secondary linguistic data. Primary linguistic data are (original or derived) representations of specific speech events with their spatio-temporal coordinates, i.e. of objects with a historical identity. Secondary data are more abstract in some respect. At a first level of abstraction, we get what Lyons (1977:29-31) calls "system-sentences." These are sentences in written representation that lack spatio-temporal coordinates and, therefore, a historical identity. They are being used as types rather than as tokens, but come along with a claim of being usable in some actual speech situation, thus, a claim of being potential primary data. Yet more abstract are facts concerning (primary or secondary) data, including metalinguistic statements on properties of speech events or system sentences and higher order generalizations over such properties. This includes, in particular, the starred example sentences and their non-starred counterparts mentioned before and, more generally, so-called "negative data" (or what Iannàccaro (2000:68 *et pass.*) calls "antiesempio"), i.e. claims on the non-existence of certain phenomena.

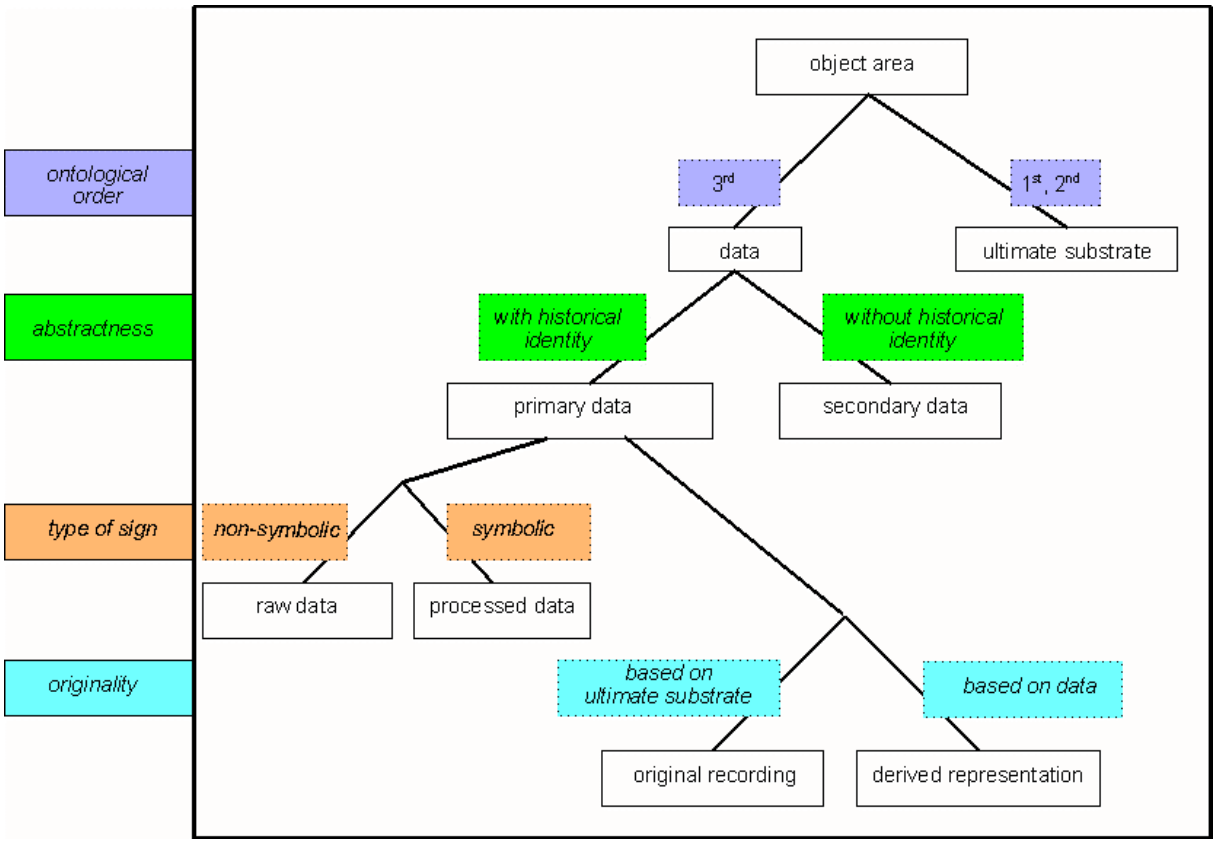
A primary linguistic datum necessarily represents a certain spatio-temporal variety of a language. Certain branches of linguistics, including much of descriptive linguistics and a fortiori typology and universals research, are usually not interested in intralinguistic variation and want to make statements concerning a language as a whole. Such statements may be based on primary data, but abstract from their spatio-temporal setting. More often than not, they are based on system-sentences and other kinds of secondary data.

I have now introduced two related distinctions: one between primary and secondary data, and one between original and derived representations. The latter distinction obtains inside the category of primary data. In section 6.2, we shall see that the production of derived representations generally involves abstraction. Therefore, the transition from primary to secondary data in a sense continues an increase in abstraction that also holds for the progression from original to derived representations.

3.4. Data and operational procedures

Table 2 summarizes the kinds of entities, in particular of data, introduced so far. It is to be read as a decision tree from top to bottom. The criteria of classification and their values are arranged in colored boxes.

Table 2. Types of data



At the bottom level, the two distinctions according to type of sign and according to originality are independent of each other. Both of them obtain within the class of primary data. Secondary data are, of necessity, symbolic and derived.

In general, for a datum to be accepted as such in the discipline, there must be operational procedures of relating secondary to primary data, and primary data to the ultimate substrate. Such procedures are part of the methodology of that discipline, viz. of the methods that allow scientists to control the relationship between the theory and the data. The methods establish the relationship in both directions. On the one hand, they are standardized procedures[12] that may be applied routinely to a set of raw data and allow the scientist to convert these into sym-

[12] the "survey protocols" and "analytical techniques" of Simone (2001, section 2.3f)

bolic representations and analyze the latter, i.e. to develop hypotheses on them. On the other hand, such methods constitute the operationalization of theoretical constructs, i.e. they specify the conditions under which a concept may be applied to a phenomenon and under which a theorem is considered falsified by a datum. If there are no such operational procedures, then firstly there is no basis on which the datum can be taken for granted, which means it is not a datum in the sense of our definition; and secondly, there is no way of relating a theory to a perceptible epistemic object, which means it is not an empirical theory. This is a field in which linguistics has not excelled during its existence. In section 6.2, some elementary standards of working with derived data are proposed.

4. Role of data

4.1. Methodological status of data in science

Only in an empirical science can data have any import at all. If there is sometimes talk of data in a logical discipline such as philosophy and mathematics (cf. fn. 7), it makes sense only to the extent that some specific research occasionally does make use of empirical methods; and the same can be said about hermeneutic disciplines such as literary studies. On the other hand, in an empirical discipline such as biology and chemistry, data are the ultimate basis, the point of reference and touchstone for any scientific statement.

The case of linguistics is, again, more complicated, since it shares properties with all three kinds of sciences mentioned.[13] To the extent that the object of linguistics is a construct of our mind, linguistics is a logical science. To the extent that it is observed in the world around us, it is an empirical science. And to the extent that the object requires understanding, linguistics is a hermeneutic discipline. These facets of our field have influenced diverse conceptions of data in it. If linguistics is a logical science, then it needs no data at all. Witness of this attitude are a couple of linguistic theories, e.g. the one of Coseriu (1958), which were launched without consideration of a single linguistic datum or in which linguistic data only play the role of illustrations which render the theoretical statements more readily intelligible for the consumer. If linguistics is an empirical science, then it depends on data, both in inductive work that analyzes and generalizes over the data and in deductive work which tests hypotheses on the data. A good example of this approach is the kind of sociolinguistic work represented in Labov (1982). Finally, if linguistics is a hermeneutic discipline, then its object is not data in the sense of facts ultimately reducible to observable entities, but instead mental representations conveyed and modified between understanding subjects, including the participating scientist. Stolze (1992) is an example of this approach.

In the history of the discipline, various brands of linguistics have gravitated towards one or another self-image. Accordingly, the role played by empirical data has varied enormously. With some simplification, we may distinguish three phases. The first is the logical strand of linguistic research, which stems from antiquity. For Plato, Aristotle and the Stoics, grammar was a precondition for rational argumentation. This conception prevailed in medieval modism

[13] Simone (2001) is rather negative on the status of linguistics as an empirical science. The term he applies to linguistics, 'meso-science' (intended to mean 'half-science'), however, also has a positive interpretation, viz. 'center-science.'

and in rationalist general grammar. Since the advent of modern linguistics at the beginning of the 19th century, the logical approach to language has been strong in typology, e.g. in the work of Friedrich Schlegel and Vladimir Skalička, and of course in those approaches that aim at elaborating linguistic theories, including the work of Louis Hjelmslev, Eugenio Coseriu and Noam Chomsky. In this tradition, linguistic data play no role in research, since the linguist only externalizes what is already in his mind. Representations of speech events are neither needed as a basis of inductive generalizations nor as touchstones of empirical verification or falsification of hypotheses deduced from the theory, since it is not an empirical theory.[14] If tokens of linguistic units like words or sentences appear in scientific treatises, it is as illustrative examples in order to facilitate understanding. Consequently, the logical trend in linguistics has not produced a culture of linguistic data. On the contrary, to the extent that it prevailed in modern linguistics over a certain period, it suppressed any methodology that would care for a responsible treatment of data, with the consequence that knowledge and skills in this area are relatively underdeveloped in today's descriptive linguistics, if compared with neighboring disciplines such as anthropology or sociology.

Only slightly younger is the hermeneutic phase of linguistics, starting with the school of Alexandria, with Dionysios Thrax (ca. 160 – 95 BC) and Apollonios Dyskolos (first half of 2nd cent. AD), and continuing in the philologies to this day, but in linguistics itself essentially discontinued since its inception proper, at the beginning of the 19th century. In this tradition, linguistic data mostly take the form of manuscripts. They are preserved essentially on account of their content, not because of the linguistic data they contain. The latter, again, are not viewed as tokens of a type, as instantiating patterns of the linguistic system, but instead as expressing unique messages, sent by a member of a certain society and culture to the scientist as the recipient, but a member of a different society and culture. Although this tradition has developed an admirable skill and diligence in editing, transmitting, archiving and interpreting texts, no notion of data as representing the object area of linguistics has developed.

The view of linguistics as an empirical science is a latecomer in the history of the discipline. There were timid beginnings in historical-comparative linguistics since Franz Bopp, a century later ousted to a large extent by European structuralism as launched by Ferdinand de Saussure. Awareness of the linguistic datum, its nature, role and dignity, evolved first in those branches of linguistics that actually executed fieldwork. These were essentially European dialectology at the end of the 19th century and, in the first half of the 20th century, American structuralism in its contact with anthropology. The achievements of the latter were essentially annihilated, in the way just alluded to, by generative grammar. They were, however, taken up and refined in the seventies by the modern disciplines of sociolinguistics and psycholinguistics. These imported the methodology of such thoroughly empirical sciences as sociology and experimental psychology into linguistics, including their methods of obtainment and manipulation of data. Here we meet, for the first time, an elaborate conception of the linguistic datum. It is, however, outside the institutional core of the linguistic discipline, whose object is

[14] This assertion itself, although not its reason, is freely conceded by the advocates of the investigation of the 'I-language' instead of the 'E-language' (cf. Simone 2001:58 against Chomsky 1986:22). Note that the construct of I-language (internal language) is meant to avoid the requirement stated in section 2.2 that a datum must have a basis outside of and independent from the researcher.

linguistic systems and which produces descriptions, comparisons and theories of this object. It has only been for a relatively short time that descriptive linguistics, taking up insights of European dialectology and American structuralism, has been struggling to raise its methodological standards concerning the treatment of data to the level established in socio- and psycholinguistics. In the past few decades, more and more linguists have dedicated themselves to fieldwork and to the documentation of endangered languages, involving the recording, representation, elaboration and archiving of primary data for their own sake. There are so few relevant traditions and established methodological standards in our field that the urgent necessity of documenting endangered languages has given a decisive thrust to the elaboration of techniques concerning the processing of linguistic data.

4.2. Functions of data

The functions of data in empirical research are various. They derive from the relationality of the concept of the datum as introduced in section 2.1: The datum is related to its producer or source and to its user or recipient. We will start with the former.

4.2.1. The relation of data to the producer

In the past fifty years, scientific standards regarding sources of data and their identification have raised considerably. Up to the middle of the twentieth century, it was customary in descriptive linguistic work – although not in dialectology and the philologies – to use examples without any indication of their source. If the language being described was the author's native language, one could bet that he had produced the examples himself, because it was not customary to base descriptive work on corpora or on fieldwork. In the past decades, the use of corpora has become both easier and more wide-spread, and it is now standard to identify the source of the data. Diligence in this respect still varies a lot, but there is a trend towards more accuracy concerning the spatio-temporal coordinates of the production of the data so that the consumer gets a chance of controlling the particular variety of the language represented by them.

Here also belongs the issue of the representativity of the data (on which more appears in section 5.3). Most linguistic descriptions purport to deal with a language. Then, of course, the question arises how the data must be sampled in order to justify such a claim. A good example of sound methodology in this respect may be found in frequency dictionaries (e.g. Gougenheim et al. 1967). Methods of delimiting the linguistic variety to be documented and to define a balanced sample that represents this variety are firmly established in the tradition of their production.

The producer of the data is normally a person in his capacity as a speaker of the language in question. Respect for his role has increased, too. While his renaming from "informant" to "consultant" may safely be neglected as an outflow of political correctness, the career that he has made from the background to the foreground of linguistics is much more noteworthy. While he was not even mentioned in the earliest linguistic publications based on fieldwork, it

is now standard to render him due attention, and not seldom has he advanced to the position of co-author of the linguistic description (e.g. Hofling and Tesucún 2000).[15]

What is perhaps even more significant is the fact that speech communities are no longer content to serve as mere sources of data. On the one hand, they want to have a say in the research project, determining what is to be recorded and published and what not; and this development has, alas, not always served to support scientific work. On the other hand, they have developed a genuine interest in the data that are produced and processed in linguistic projects, as that may be the form in which the language survives after it is no longer used by its speech community. The data here fulfill an important function in documenting the way of life of a society for posterity, so that future generations of the community may at least learn how their forefathers lived, and maybe even revive the traditional language (cf. Lehmann 2001). As a consequence of this new function of linguistic data, an awareness of the necessity to develop standards of quality has arisen, and we currently witness an unprecedented upgrowth of research projects that develop standards and technological facilities designed to represent, process and archive linguistic data.[16] The progress in all of these respects is, of course, a direct consequence of the fact that linguistics has, during the same period, increasingly become an empirical science.

The quality of data concerns their user, and to this we will come in the next section, but it also concerns their producer. Ceteris paribus, data whose content and form satisfies esthetic and spiritual standards are more highly valued than junk data. What appears to be a truism for philologists and for members of a community whose traditional language is threatened by extinction is an unwonted thought for most linguists. People who are professionally more interested in structure than in content tend to neglect such qualitative criteria. However, where resources are limited – and they certainly are in the documentation of an endangered language – quality of the data is the essential selection criterion besides representativity (see Lehmann 2001, section 5), and it is the professional task of linguists to respond to such demands.

4.2.2. The relation of data to the user

4.2.2.1. The relation of the data to the researcher

As to the user of the data, we may distinguish between the researcher who takes them for granted and the addressee of the research. Starting with the researcher, we can distinguish between the research itself and the report on the research delivered to the consumer. In the research itself, the data are used either as the basis for induction or as the test of theorems that were deduced. In the report, the data play the argumentative role of evidence for the theory (cf. Simone 2001, section 2.6).

Let us first come back to the literal meaning of the term *datum*, "given." In real life, the researcher is, of course, not merely a passive recipient of the data. It was said in section 2.2 that data are not a specific kind of thing waiting to be discovered by the scientist. It would

---

[15] Iannàccaro (2000, section 6f) insists that informants just like linguists have their linguistic theories which may shape the data they furnish.

[16] The DoBeS (Dokumentation bedrohter Sprachen) program of Volkswagen-Stiftung, which has been running since 2000, is a representative example.

therefore be naïve to assume as a normal course of things that the ever-attentive scientist hits upon a set of data and then feels impelled to develop a theory that accounts for them. This may happen from time to time, but even then he has the choice of ignoring the data. In general it is the scientist's epistemic interest that triggers the research, including the supply of data.

The essential difference between an empirical and a logical science is that the object area of the former has an ultimate substrate outside the scientist, and insofar its properties do not depend on him. Of course, the scientist defines his object area, delimits it, and as we shall see in section 5.2, he can bring the existence of the data about. But even if he does so and in his research controls a number of variables, he does not determine the dependent variable, i.e. that property of the data which he is investigating. This is, at the same time, an essential difference between empirical and hermeneutic methodology. Cases in which the empirical researcher nevertheless influences his data fall into a variety of categories. The category of fraud is methodologically least interesting. The other problems with objectivity of data are essentially bound up with the way they are obtained and will therefore be discussed in section 5.

The greatest practical problem for the researcher is usually that the kind of data required for the particular research project is not available, so that part of the project is precisely data provision. While this is usually not a problem for languages spoken in the research team, it may cost considerable time and money in other cases. The expenses tend to get disproportionate in typological projects, both because these need data from many diverse and often remote languages and because a typological project should exploit finished descriptions instead of having to procure and analyze primary data in the first place. This gives rise to the question of why it is not possible, in descriptive and, a fortiori, in typological linguistics, to make use of available data instead of having to provide them in the course of the research in question.

One must say at the outset that linguistics is not the only science that has this problem. Probably 95% of all the projects in the empirical sciences, from chemistry over neurobiology and sociology to psychology, produce their own data on which they base their research. Very seldom indeed does a project take up the data of another project to either examine them critically or to use them with a different epistemic interest. On this background, the linguist could just reject the demand of using other people's data.[17] The motivation for requiring the availability of independent data in linguistics probably stems from the fact that the processing of linguistic data to the point that they can serve as the basis of higher-level analyses is extremely laborious. While the natural sciences are accustomed to automatizing their data processing to a high extent so that it costs more material and machine-power than manpower, automatization of data processing in linguistics is rather underdeveloped. First of all, trips to remote places and months of fieldwork may be necessary to obtain a sizable corpus of raw data. Secondly, processing linguistic raw data to the point that they appear as texts with several tiers of analyses plus translation, as described in section 6.2, requires a linguist who knows the language. And finally, even such a person takes years to fully analyze a reasonable corpus of texts. It thus becomes intelligible that researchers, especially those with more theoretical interests, do not always want to start from scratch, and funding agencies tend to feel the same.

---

[17] And, true enough, some researchers do not make their data available to others.

Such considerations lead to the elaboration of representative corpora of languages, such as Svartvik and Quirk (1980) for English, the *Mannheimer Corpus* for German, ADMYTE for Spanish, Archivio dell'Italiano Parlato and so on, and in the past two decades they even gave rise to a new branch of linguistics, viz. corpus linguistics. As a result, both very large corpora for some European languages and standards of elaborating such corpora have been developed. And there can be no doubt that such corpora are increasingly being made use of in empirical research. The problem is, however, that progress in science leads to new kinds of questions, questions which the producers of the corpus could not foresee and therefore did not think about tagging their texts for. Currently we have arrived at the point where users require the corpus, with all its sophisticated derived representations, to come along with the raw data that it is based on, i.e. they want to check the data on the original videotapes (cf. section 6.1). At this moment, it is hard to prognosticate whether satisfying this demand will solve the problem of the insufficiency of "free" corpora which the investigator has not tailored to his specific problem. It must be admitted that apart from some corpora that are frequently used by different researchers, there are also large data cemeteries, collections of data that were gathered and processed with enormous expenditures for some specific research project and which are completely useless after the end of that project. And there is no doubt either that the use of information technology, i.e. digital representation of the data, has tended to aggravate rather than alleviate this problem.

The deficiency of available data on most languages of the world is possibly even direr when it comes to secondary data. This is both a problem of quality and of quantity. Linguistics is still a relatively young science; in the beginnings scientific standards were not particularly high, and there has not been sufficient manpower to describe the thousands of existing languages in breadth and depth. As a result, many questions of contemporary typology and universals research do not find an answer in available grammatical descriptions.

Many of the problems in this area are methodological problems, in particular problems of standardization. The form in which linguistic data should be presented and in which corpora should be arranged, and the structure that a grammatical description should observe and the kinds of questions that it should answer are routine issues that are amenable to standardization to a much higher extent than many linguists appear to believe. At the moment, we are still at the stage where 30% of the linguists think they can get away with a phonetic representation which ignores IPA and where everybody thinks up his own conventions of interlinear morphological glossing. It would contribute to the emancipation of our science if basic methodological operations that are really routine could automatically obey an established standard. This would free linguistic work for those tasks that really require mental energy.

4.2.2.2. The relation of the data to the addressee

Finally, we have to look at the function of the data for the addressee of some scientific report. If the addressee is more interested in the theoretical results of the research than in the data it is based on, then he may wish to join the author of the report in taking the data for granted. At the level of interpersonal communication, his consideration of the data will then depend on the confidence that he has for this author and his sources of data. In anticipation of this attitude of the consumer, many linguistic publications do not include their data or at most relegate them to an appendix.

On the other hand, it is possible that the addressee of the research report wants to check the data, be it that he mistrusts the author, be it that he develops an interest in the data that

goes beyond the function that the author had destined them for. Such a reader requires full explicitness in the presentation of data, including both verifiable information on their provenience and such a representation of their linguistic structure that is sufficient for their controlled understanding. Here we again hit upon the issue of the "datum for its own sake." Data are representations of such aspects of the research object that correspond to the epistemic interest of the researcher. He cannot foresee the diverse epistemic interests of his addressees, and he cannot exhaustively represent in his data every aspect of the epistemic object. This is a practical problem that can only be solved by steering a middle course. If the researcher makes use of a published corpus, then he has the right to simplify the representation of the data for the sake of his epistemic interest, provided that he refers the addressee to the original. If the research is based on original data, then it is the duty of the researcher to make them available in such a form that the addressee of his report can fully control them (Himmelmann 1993). This involves standardized linguistic representations of the data, which we will come back to in section 6.2.

For the addressee of a research report, data may also fulfill a function of visualization and illustration. However, this is only a secondary function of linguistic data. Their primary function as stated in section 2.2 is to serve as the basis of induction and as the test for deduction. The expository or illustrative function is actually the primary function of examples, not of data. In other words, the mere fact of being used as an example in the exposition does not confer the methodological status of a datum to some linguistic expression. Some examples come along with a claim of being data, others not.

5. Obtainment of data

In the last section (4.2.2), the methodological relation between the data and the researcher was introduced. We now have to deal more in detail with the ways of obtaining linguistic data (cf. Iannàccaro 2001). We will concentrate on primary data in the sense of section 3.3 and on the ways in which the method of data provision affects their quality.

5.1. Introspection

The linguist who is a speaker of the language he describes can himself produce and interpret data of this language. The method of using one's own language competence in both these linguistic operations and in the associated metalinguistic operations like grammaticality judgements, paraphrases and the like, is called introspection. It has a venerable tradition in linguistics, as the forefathers of logical linguistics, Plato and Aristotle, already relied on it (see section 4.1).

Taking up the distinction between production and understanding, we can observe that the role of introspection differs. While some linguists have requested, in more or less unsystematic ways, grammaticality judgements of fellow speakers for example sentences the linguist had thought up, the interpretation of primary data originating in his speech community has mostly been regarded as the professional task of the linguist that he must respond to and for whose completion he does not depend on others. Speech recognition experiments might appear to be an exception to this generalization. These, however, do not aim at learning about the meaning that members of the speech community assign to some utterance produced by

another member. Instead, the meaning of the test sentences is presupposed among the controlled variables, and the research interest is in the mechanisms that subjects apply in interpreting them.

Even if it is granted that a linguist (who uses the hermeneutic method) is a professional interpreter, it is worth noting that this capacity of the linguist has its limits even for his native language. I am referring to linguistic varieties not covered by his competence. Most conspicuous here is the problem of interpreting child language data, where the hermeneutic intuition of the researcher may fail and possibilities of metalinguistic interaction with the native speaker are limited.[18]

As for the production of example sentences upon introspection, this has been an established custom in descriptive linguistics to our day. To the extent that linguistics is a logical discipline, this is unobjectionable (cf. section 4.1). However, introspection has been treated as a safe empirical method, under the pretense that a linguist's speech behavior and grammaticality judgements should be (at least) as good as those of any other native speaker, and moreover the data he produced could be counterchecked by the other linguists whom he addresses with his research report, so that objectivity was guaranteed. It has become clear for some time now that these suppositions are false and that this use of introspection is a misuse of the concept and associated ethos of empirical science. First of all, linguists are members of a closed circle in their speech communities who share a sociolect that is narrowly delimited, subject to traditional normative standards and nothing less than representative of their language. Second, the procedure in which a linguist produces data on which he constructs a theory which he then tests by these data is, of course, circular. The data do not actually fulfill any control function, and the procedure has nothing to do with scientific method. And last but not least, few linguists have escaped the temptation to dress the data they produce according to the theory they cherish.

The net result of all of this is that introspection is necessary and useful as a heuristic tool in linguistic work, but it is not part of empirical methodology, and the data thus produced have no status in empirical research besides illustrating what the linguist theorizes.

5.2. Production vs. discovery of data

Coming to the serious ways of obtaining linguistic data, there are essentially two of them: data may be found or may be produced.[19] Of course, data that are found have been produced at some point. However, what is important here is that the researcher may participate causally in the production of the data or may just come across pre-existent data. Disciplines differ crucially in this respect. In traditional archaeology, data accepted by the methodology must be pre-existent and discovered.[20] Data produced under the influence of the scientist have the

---

[18] While interpretation of child language data is certainly possible to various degrees, their production upon the researcher's introspection is outright impossible.

[19] A related distinction is made in Iannàccaro (2000, section 3) and Iannàccaro (2001:25): 'phenomena,' which are found (without being searched for), are distinguished from 'data,' which are searched (or produced).

[20] Modern archaeology employs methods of the natural sciences, and here things are again different.

status of fakes. In neurolinguistics and experimental psychology, on the other hand, all the data that are of relevance are produced under the control of the scientist. And in a natural science such as neurobiology, a certain rat brain is not an interesting datum in itself, but only if a certain area of it exhibits a certain change of color produced by the experimenter. The immediate conclusion from this is that both kinds of data are well-established in scientific methodology. It may therefore cause no astonishment if both of them are used in linguistics, too.

In the purely empirical sciences, which do not have the hermeneutic component which linguistics has, reliability of research methods is a must, and it entails as a corollary that results must be reproducible. Consequently large amounts of similar data are produced so that one can apply statistical methods to them. This kind of approach does exist in linguistics, too, chiefly in neurolinguistics, experimental psycholinguistics and in statistical linguistics. There are, however, limitations and drawbacks to such methods which will be mentioned in section 5.4.2. Much linguistic research is devoted to data that are not reproducible, be it for contingent reasons, because the factual preconditions for their production can no longer be met, be it for theoretical reasons, because they do not have the status of tokens, but of a type (cf. section 3.3). Although these two situations are methodologically totally different, they do share the uniqueness of their data. This is typical for a science that has a share of hermeneutics: The epistemic object is a manifestation of the mind of an individual human being, and, insofar, it is not measured and no induction is applied to it, but instead it is understood.

In other disciplines, the distinction between produced and discovered data is handled with utmost diligence. In neurobiology, no ambiguity ever arises over the issue of whether the color of the brain area was there before or after the experiment. In this respect, there is fault with much linguistic work. Especially in work dealing with grammar, there is a tradition going back to antiquity for the researcher to use, side by side, example sentences that he found in a corpus and examples that he has produced himself. And if examples are taken from spontaneous recordings, they are normally edited, because speech errors and the like are of no interest to the grammarian. Normativism lurks behind every corner, and objective data become indistinguishable from illustrations of what the researcher thinks should be the case.

Primary linguistic data are tokens of linguistic signs. Human beings have two converse relations to them: either as a speaker or as a hearer. The linguist takes the same two perspectives, as shown in Table 3. In the perspective of the hearer, he is confronted with utterances produced by somebody else. He analyzes their form and structure, interprets this and thus arrives at the meanings and functions carried by the data. This is the semasiological perspective, which is typical of structural linguistics. In the perspective of the speaker, the linguist starts from some cognitive or communicative function which is to be fulfilled by linguistic signs. The utterances produced in this way are functional variants of each other, and so the linguist sees which structural means the language uses to fulfill such a function. This is the onomasiological perspective, typically taken in functional linguistics.

Table 3 Two perspectives on linguistic data

viewpoint	**basis**	**semiotic operation**	**perspective**
hearer	forms and structures	interpretation	semasiological (“structural”)

speaker	cognitive and communicative functions	production	onomasiological ("functional")

A comprehensive description of a language system is arranged either by structural or by functional criteria, and partial descriptions are devoted either to some structural device or to some functional domain. The descriptive linguist is therefore in need of data that share either their structure or their function. Natural language users, however, are only attentive to structures and functions in their respective contexts. They do not by themselves aim at producing discourse that only fulfills a given function, or at only understanding discourse that exhibits a certain structure. Such one-sided interest is exclusive of a professional approach. The linguist does have ways of selecting data of the kind that correspond to his approach. However, just as natural language users are both speakers and hearers, the linguist must be aware that the semasiological and onomasiological approaches complement each other; to take only one of them gives a biased picture of the language. With this in mind, we will examine, in the following two sections, the use of linguistic data in the two approaches.

5.3. Spontaneous data and the semasiological approach

The semasiological approach to linguistic description presupposes a large corpus. Two entirely different methodological situations must be distinguished here. The first is defined by a corpus language. For languages such as Hittite and Accadian, we dispose of large, but finite corpora. We cannot change this empirical situation, so if we want to describe such a language, we have to take the semasiological approach. The other situation can be characterized as the exploration of a corpus whose production was triggered by the researcher in the first place. In this, he may pursue diverse purposes. The linguist may want to document a language in danger of extinction. Then the task is to produce a corpus of the language that should survive it and be available to future scientists and laymen alike. Or he may need data that represent a particular variety of a language, as for instance when child language data are recorded in order to investigate a particular problem of language acquisition.

The two situations have a problem in common, which is the representativity of the data in the corpus. In the case of a corpus language, obviously only the written language is represented. The lack of oral communication means that not only the basic mode of communication, including all of the phonetics and most of the phonology, remains unknown but also – especially for a language from an ancient society, where only a small percentage of the population was literate – what is quantitatively the bulk of communicative events remains unrepresented. Written communication is often restricted to very specific genres, not only in antiquity. For instance, for many earlier linguistic stages or extinct languages of the Americas, we only have catechisms and similar religious literature. On the basis of such data, one can hope to reconstruct only a very approximate image of the language.

If it is the linguist who triggers the spontaneous production of a corpus, the problem of representativity should, in principle, be solvable. If research is limited to a well-defined linguistic variety, as in the example mentioned before, the main problem is usually a practical quantitative problem, in the sense that the sample of different idiolects must be large enough to warrant generalizations concerning the linguistic variety as such. In the case of the documentation of a language, the problem of representativity has not been solved to this day. Lin-

guists who described languages in fieldwork have concentrated, again and again, on a very few genres of texts, viz. myths, tales, autobiographic stories, in short: narratives. To be sure, this happened for good reasons: These speech events are easy to record, and they tend to be well structured and to contain a relatively high portion of complete and grammatical sentences. However, they are not at all representative of communication in the community, because they are essentially monological, while most communication, especially in scriptless communities, is dialogical or polylogical. The problem of what constitutes a representative corpus of a language is a new one in linguistics and is a challenge both for linguistic theory and methodology and for practical linguistic work. (See Lehmann 2001.)

There are essentially two ways of linguistically analyzing a corpus. If one is interested only in a particular structural feature, for instance the genitive and its functions, then one scans the corpus for all genitive forms, produces a concordance of them, classifies the examples found and comes up with a semasiological analysis of this structural device. If, on the contrary, one aims at a comprehensive description of the language system, the choice method is to analyze each successive sentence of the corpus as regards its internal structure at all levels and its linguistic and extralinguistic context, to assemble the structural categories, relations and processes of the language in this way, to classify them by structural criteria and then to assign each structural device its functions. The approach is well-established in the philology and linguistics of corpus languages. Linguistic descriptions of languages such as Accadian and Hittite have been elaborated essentially in this way.

The approach was formalized in structural linguistics during the first half of the twentieth century. The two principal operations of segmentation and classification are applied at the lower levels of linguistic structure, and the result is a structural description that covers at least the phonology and morphology. At the higher levels of linguistic structure, i.e. the syntactic and textual levels, the recognition of patterns and the correct assignment of a given token to a pattern involve procedures of interpretation that are not easily formalized. Here again the hermeneutic approach comes in, which offers both the advantage of perceiving contextual relations, disambiguating polysemous or homonymous structures and figuring out the deeper sense of an utterance, and the corresponding drawback of subjectivity.

Given that in language, structure serves function, no comprehensive analysis of linguistic structure without reference to its function is possible. Some schools of American structural linguistics made it a point to apply "cryptanalysis," i.e. to come up with a structural analysis of primary data without knowing their meaning. These attempts must be considered failed. The story of script decipherment is rich in illustrative test cases. All cases of successful script decipherment involved some kind of historical or archaeological information or even a bilingue in addition to the texts themselves. Wherever such information is not available, as in the case of the Indus Valley script, any linguistic analysis fails.

It is quite a different issue whether one needs to be a native speaker of a language in order to analyze it. The answer given to this question by the success story of linguistics is a clear "no." Apart from some peripheral items such as nursery rhymes, whose knowledge may serve as the ultimate touchstone of the native speaker, many persons (including linguists) have acquired a full, native-like command of a second language. And this is not even necessary in order to do a linguistic description of a language, since understanding and controlled interaction with native speakers may serve the same purpose. Witness are dozens of excellent grammars of languages not mastered by their authors.

Human beings, including linguists, are not genetically equipped to focus on linguistic data that have a certain structure, and are fallible in this task. Even researchers of good will are not immune to uncontrolled distortion of their data. It has been shown (Cutler 1981), for instance, that collections of speech errors that were observed under non-controlled circumstances are often not reliable because of the subconscious hermeneutic interaction of the person who notes them. That is, people, including linguists, subconsciously change their perceptual input. It is also well-known both to anthropologists and to linguists that the kind of participating observation that is typical of much fieldwork inevitably distorts the data; and even if the fieldworker is aware of it, he cannot eliminate the bias altogether.

The procedure of scanning the corpus for data that have a given structure is, in principle, automatizable. Tools that do this service are available at diverse levels of sophistication. The basic level affords just a search for certain allomorphs or word-forms and makes a concordance of them, while the most advanced level involves algorithms of speech recognition and grammatical analysis. Needless to say, tools of the latter kind are available only for a handful of languages. For all the other languages of the world, the corpus will first have to be tagged for the kind of structural information that one may want to retrieve. This, however, presupposes just the kind of structural analysis that we are talking about.

The advantage of working with a corpus is, of course, the enhanced objectivity of the data and of all the research that is based on it. In comparison with the other approaches, the possibilities for the researcher to manipulate the data are minimized. Another great advantage is that a corpus the researcher has not produced himself may be varied, heterogeneous, full of surprises and a constant source of inspiration. Exposing oneself to spontaneous data is, in fact, the safest way of discovering those categories of a language that are peculiar to it and that the researcher did not expect. The heterogeneity of spontaneous data has, it is true, two sides. Multiplicity and richness is the positive side. The negative side is wild variation. It is the task of the linguist to systematize and interpret variation. But a good deal of the variation present in a corpus is not due to corresponding differences in function or in the context, but is just dysfunctional: idiolectal idiosyncrasies, dialect differences that one would factor out if one could control them, false starts and other kinds of speech errors with their repairs etc. And the undeniable drawback of a corpus is its incompleteness. Certain lexical items, morphological forms and syntactic constructions will be lacking even from a very large corpus. However, this just confirms what was said above about the complementariness of the semasiological and onomasiological approaches.

5.4. Generation of data and the onomasiological approach

In the onomasiological perspective, the researcher wants to know how a certain cognitive or communicative function is fulfilled in the language, and the task is to obtain primary linguistic data of utterances that fulfill it. This presupposes a theory of the cognitive and communicative basis of language which is subdivided according to functional domains such as Concept Formation, Reference, Determination, Possession, Spatial Orientation, Temporal Orientation, Participation, Interpropositional Relations etc. (cf. Lehmann in press, section 2.2). Each domain is spelled out down to the level of typological grammatical categories. Depending on the specific research interest, for instance collecting new data in the domain or classifying available data by some functional parameter, the concepts are then operationalized in the form of questionnaires, example sentences, test frames and the like. Since our interest here is the ob-

tainment of data, the methods that are more appropriate for the classification of data, especially test frames, will be foregone, but a few of the others will be singled out.

5.4.1. Elicitation and translation

In working with informants, an established elementary method of obtaining data in a predefined functional domain is to elicit them with the help of metalinguistic procedures. If a morphological paradigm is wanted, then the grammatical parameter in question is assigned each of its values in turn, and the informant is asked to provide the corresponding forms. An analogous procedure can be applied at the level of syntax, by transforming a sentence into a minimally different one according to some relevant functional parameter.

The translation method consists in preparing example sentences of the background language (i.e., the regional lingua franca that the linguist and the informant use for communication) and to ask the informant to translate them into his language. The example sentences have systematic paradigmatic relations to each other so that they cover the expected variation in the functional domain in question.[21]

However, in its simple form, the translation method is intrinsically invalid. If the task is to find out those grammatical categories of the target language which render certain functional categories, it is methodologically inappropriate to present the latter in the disguise of the grammatical categories of another language, as this obviously leads to interference from the latter. On the other hand, the method has a couple of advantages, and it is therefore worth refining. One way of doing this consists in translation questionnaires (see, for example, Dahl 2000, appendices). Here, characteristic little stories or situations are constructed, in which the sentence to be translated is embedded. The context is configured in such a way as to force the association of that sentence with the cognitive category which is at stake and whose expression in the target language is to be tested. In the original version of the questionnaire, the category would appear in its English grammatical manifestation, but that is suppressed by presenting its host word as a mere lexeme, without any grammatical categories and, in particular, without any hint to the grammatical category being tested. This is, of course, done in order to minimize interference from the background language used. The following is a typical example from such a questionnaire:

> *Perfect questionnaire (Dahl (ed.) 2000:803, #37)*
> It is cold in the room. The window is closed. A asks B:
> You OPEN the window [and closed it again]?

The example presupposes a functional concept which may be described as "temporal localization of an event in the immediate past prior to the speech act such that, not the state logically resulting from the event itself, but a physical consequence of it persists at speech act time"; and it is asked which structural category the target language uses to express it. Some

---

[21] In its simplest form, the method goes back at least to dialectology (see, for example, Weijnen et al. 1975-9). In a more modern form, it underlies the series *Archivo de Lenguas Indígenas de México* launched by Jorge A. Suárez and now edited by Yolanda Lastra (1974ff). Here, the documentation of a language consists of the translation of a set of several hundred standardized sentences into the target language. The sentences are chosen in such a way as to maximize chances that their translations will exhibit the central grammatical categories and vocabulary of the language.

languages (like German) would use the perfect tense here; others (like Spanish) would use the simple past; yet others (like Yucatec Maya) would prefer the perfective aspect.

With both the elicitation and the translation methods, the responses of the informant are recorded, analyzed with his help and counterchecked with other native speakers. Both methods are frequently applied in fieldwork on underdescribed languages. They are popular because they are inexpensive in every respect. To a certain extent, they are necessary to systematically complete data obtained by other methods. However, as has long been known, they are, to some extent, both unreliable and invalid. They are unreliable because the linguist, the informant and their relationship are sources of error which render the data faulty. Elicitation and translation are more a hermeneutic than an empirical method, as the two persons in their interaction jointly construe some meaning. The two methods are invalid to the extent that they are meant to reveal the grammatical categories that the language possesses. In fact, they only reveal such categories that the analyst expects and therefore codes in his questionnaires, example sentences and paradigmatic operations. It is therefore crucial that the onomasiological method does not rely on the grammatical categories of the analyst's language, or on any grammatical categories at all, for that matter. Instead, it must rely on a universal (i.e., language-independent) system of cognitive and communicative functions. To the extent that linguistics does not (yet) dispose of such a system, it cannot be guaranteed that these two methods will discover the grammatical categories of the language. Consequently, what is true of any scientific method at all is a fortiori true of the methods of elicitation and translation: They must never be applied in isolation, but must always be complemented by other methods.

5.4.2. Induced speech

Sometimes data that are relevant to the research topic are too rare in the corpus or otherwise hard to come by. Another set of methods within an overall onomasiological approach involves induced speech, that is, the elicitation of linguistic behavior by non-linguistic stimuli. The Max Planck Institute for Psycholinguistics at Nijmegen has been developing, over the years, a sizable set of tools, kits and experiments to be employed for this purpose in diverse cognitive and communicative fields. One type of method involves the representation of little scenes with puppets or by silent movies, which are then to be described or retold by the native subjects. There may also be communicative problems to be solved, such as the task of orienting a fellow in space or instructing him to mount a device. All of these methods presuppose a certain functional domain and a set of cognitive or communicative operations in it. The setup of the experiment is designed and the task is defined in such a way as to maximize chances that the linguistic solution to the task will make use of the grammatical devices that the object language possesses in that area.

Similarly, speech errors are valuable data for reconstructing the mechanisms underlying speech production. There are large corpora of speech errors, but the conditions under which a datum was entered into the corpus are often opaque, so that no statistical methods can be applied. Then experiments may be conducted in which subjects are prompted to produce speech errors of a certain kind, for instance metatheses, which are then sufficient to develop systematic hypotheses on their origins (Baars 1980).

The advantage of methods of induced speech against those methods which involve metalinguistic elicitation and translation is that they exclude interference from other languages. However, the experimental setting itself is not always entirely natural, leaving aside that the mere fact of being in an experimental situation is bound to trigger uncommon linguis-

tic behavior. Moreover, methods of induced speech have the disadvantage that they are relatively costly in terms of time and money.

6. Representations of data

The ultimate substrate of linguistics is speech events. These are directly observable and may be recorded. However, just like in other sciences, the data of linguistic research are almost never tokens of the ultimate substrate itself, but only representations of it. This is essentially due to the volatility of speech events, but also to the fact that linguistics is only interested in the linguistic aspects of speech events – those aspects that are semiotic in nature. We saw in section 2.2 that data are representations of the epistemic object, and consequently they are signs. Linguistics differs from other disciplines in that its epistemic object itself is semiotic in nature, so that the object and representations of it may become indistinguishable. The two central modes of representation are the auditory and the visual mode, and to some extent they are interconvertible with preservation of those features that most linguists are interested in. As a consequence of this situation of the data, linguists have worried very little about whether a particular datum was an original or a derived representation. In this section, we will sort out the relationships of the various representations to each other.

6.1. Raw data

Language is an activity, not a (static) object. The ultimate substrate of linguistics consists of second order entities, not of first order entities. This is true regardless of whether the speech event in question is one of speaking or of writing. Therefore, the closest, most faithful rendering of the ultimate substrate is a sound movie. A sound movie represents the process in which the original utterance was produced, with its hesitations and editing operations. It represents the complete phonetics of a spoken utterance, including pauses and prosody. It shows the paralinguistic communicative behavior of the speaker, with his mimics and gestures. The movie represents the whole speech situation, with the addressee and his reactions and the extralinguistic context which is presupposed and referred to by the deixis and which is sometimes changed by speech, for instance in commands. In short, the movie renders most of what speakers naturally make use of in producing and interpreting speech.

Needless to say, a movie is only a representation, not the original. At any given point in time, the spectator only sees the scene in one perspective. Most of the time, this is the perspective of the addressee, not of the speaker. Since only the auditory and visual senses are involved, the spectator does not feel or smell what the speaker and hearer feel or smell. And there are various other reductions and distortions in a movie. Nevertheless, it currently presents the most faithful way of rendering a speech event. For many purposes inside and outside the linguistic discipline, especially for the documentation of endangered languages, but also for various didactic purposes, the best data are raw data in this sense.

Although the auditory and visual modes are interconvertible to a certain extent, the process of converting a sound movie or an audio tape into a symbolic representation and deriving various other symbolic representations from the latter is unidirectional. That is, as long as the raw data are available, one can always fall back on them and distill better secondary representations from them. The converse does not hold: Once the original recording is lost and only

symbolic representations are left, certain questions about the original speech event will always remain unanswered.

6.2. Symbolic representations

The epistemic object of linguistics has many facets and is capable and in need of many levels of abstraction in order to be fully understood. Processing linguistic data therefore essentially involves their representation at diverse symbolic levels. Depending on their particular epistemic interest, linguists represent an utterance at least[22] at the levels enumerated in Table 4.

Table 4. Levels of representation of linguistic data

n°	**level of representation**	**code and symbols of representation**
1	segmental phonetic	IPA
2	prosodic phonetic	intonation curves, stress levels, etc.
3	lexical-phonological	morphophonemes, morpheme boundaries
4	orthographic	standard orthography
5	morphological	interlinear gloss with vocables of background language
6	grammatical	grammatical categories and relations
7	semantic	translations in various languages

All of these are written representations, which means they necessitate a change of mode. This is the first step towards the reification (or hypostatization) of the epistemic object of linguistics. There is no way of avoiding it in scientific work, but one must be aware of it, or else one will fall victim to what Harris (1980:6-18) calls "scriptism." The notion of grammar (τέχνη γραμματική) as the "art of writing" is deeply rooted in linguistics.

Practically all of these representations, except the intonation curves of n° 2, are symbolic. They bear complex relations to each other which need not be analyzed here in full. Some of them, especially n° 1 – 3, render properties of the significans of the language sign, while others, especially n° 7, render properties of its significatum, and yet others, especially n° 5 and 6, represent aspects of its structure. Some representations, especially n° 1 and 2, render properties of the raw data as closely as a symbolic representation possibly can, while others, especially n° 3 and 6, are abstractions from more concrete representations. Most representations render individual linguistic items, while others, especially n° 6, show classes instead of individuals.

Correspondingly, the conversion of one representation into another one first and foremost changes the data. Most of these changes are reductions; a few are refinements. For instance, the conversion of n° 1 into n° 3 involves loss of phonetic information, while the conversion of n° 3 into n° 5 is accompanied by an upgrading because it involves resolution of homonymy. Each representation may be used as linguistic data for some purpose; each renders different questions and answers possible. It is only necessary to keep in mind that while a datum by definition represents only an aspect of the epistemic object, derived representations reduce and distort the original even more. Methods which relate derived representations ultimately to

[22] It is not important that the list of Table 4 be complete; it suffices for it to be representative.

original representations of primary data are an important subset of those procedures called for in section 3.4, which guarantee linguistics the status of an empirical science.

Thus, the two main operations which produce derived linguistic representations are abstraction and the semiotic operation of coupling a significans with its significatum. As we saw in section 3.2, the necessity of applying just these two operations constitutes the uniqueness of the linguistic datum. It is for this reason that the analogies between linguistics and other sciences sought in section 2.2 come to their end here, and Table 4 has no counterpart in any other science.

Abstraction involves reduction of variation. In converting representation n° 1 into n° 3, phonetic variation is neutralized, and in converting n° 3 into n° 5, allomorphy is eliminated. Some of this variation is part of the linguistic system; some of it is lectal or just irregular. Coughs, hesitations, slips of the tongue, false starts, etc. tend to be suppressed in the production of more abstract representations, at the latest at levels n° 6 and 7. The production of derived linguistic representations therefore also involves editing the original representation (cf. Simone 2001:57). This is principally done with a view to the norm. The clearest cases may be observed when texts that were recorded in the field are prepared for publication.[23] However, the norm is not something arrived at inductively in an empirical science, but something set by groups of speakers including linguists (see section 5.2 on normativism). The whole process is directed towards the distillation of system sentences; but if these are what is wanted, there are shorter ways of getting them, namely by introspection (see section 5.1). In processing linguistic data, two rules must therefore be observed: First, the editing must be transparent; second, derived versions must not replace, but accompany the original version.

Raw data are the most theory-free form of data that one can get in linguistics. The production of all the representations of Table 4 involves some analysis and consequently presupposes some theory.[24] Especially at levels n° 5 and 6, representations are conceivable from which the reader can abduce the entire grammatical theory of the author. This only serves to once more underline the point that data must not be confused with primary data. Whether anybody regards any of the representations of Table 4 as linguistic data depends on his purposes and on his conviction that the representation can be related back to primary data by standard methodological procedures. There is, alas, no clear-cut distinction between data and constructs; a representation is, by definition, a construct. The most one can say here is that the progression from raw data to derived representations and finally to secondary data replaces the primary data by increasingly abstract constructs.

On the other hand, some of the relations between the various levels of representation are highly regular, so that one representation can be derived from the other by the application of rules. This means that, despite the restrictions of section 4.2.2, some of the processing of linguistic data is automatizable. Here I am referring, in the first place, to the achievements of

---

[23] For instance, Manuel J. Andrade recorded many Yucatec Maya texts on discs in the 1930s. Refugio Vermont-Salas provided a close phonetic transcription of them in 1971, which was microfilmed, but never published. Hilaria Máas Collí produced an orthographic transcription in 1984, with heavy editing of the original text, and this one was published (*Cuentos mayas yucatecos*. Mérida, Yuc.: Universidad Autónoma de Yucatán, 1990).

[24] The data are "theory-laden"; see, for instance, Iannàccaro 2000:53f and Simone 2001:60.

speech technology and corpus linguistics which pertain to tagging and markup,[25] morphological analysis, interlinear glossing etc.[26] Most of these procedures will probably require interaction or control of a linguist for the rest of the lifetime of linguistics, but their automatization is nevertheless useful for the reasons mentioned in section 4.2.2: computers perform them with more consistency and efficiency and free linguists for more exacting work.

7. Conclusion

I noted above the fact that neglect of data has been the rule in linguistics, at least since the beginning of structural linguistics. However, an indulgent interpretation of this fact is possible: It appears that the theories and methodological concepts of structural linguistics first had to be developed and tried out on object languages that linguists controlled by introspection, for which data provision was no problem because any wanted amount of data could be generated at any moment, and which have an age-old descriptive tradition so that descriptive tools did not need to be developed from scratch. At the start of the twenty-first century, linguistics has become mature and now enters a new phase of its development. Thanks mostly to fieldwork on diverse languages, to descriptions that are both functional and structural, and to typological comparison, the discipline is now in a position to approach in a responsible way the rest of the world's languages whose methodological situation is less comfortable.

This moment in the history of the discipline happily coincides with new and urgent demands being made on it from outside, viz. from the speech communities of languages threatened by extinction. As if awakening from sleep in a scientific greenhouse, the discipline has suddenly become aware of the fact that its capacity is urgently needed for the documentation and description of most of the languages of the world, both for the sake of their speech communities and their interest in their cultural tradition and for the sake of the very database of the discipline itself. Language documentation has become a slogan in today's linguistics. As is usual in such cases, some members of the scientific community who are more flexible in publicizing the work they had always been doing than in adapting to urgent demands from outside have adopted the new term as a more effective label under which to sell traditional linguistic description. Most of us, however, have understood that the new situation demands a rethinking of our methodological bases. In endangered languages, data constitute a value for their own sake because they are irreplaceable. Consequently, we need to develop methodological standards for their scientific and practical treatment so that future generations can make the best possible use of them.

References

ADMYTE (1993 ff). *Archivo digital de manuscritos y textos españoles.* Madrid: Micronet.

---

[25] SGML, XML and the Corpus Encoding Standard developed by EAGLES represent important progresses in the representation of analyzed linguistic data.

[26] From among the producers of relevant software, the Summer Institute of Linguistics (www.sil.org) must be singled out.

Baars, Bernard J. (1980). On eliciting predictable speech errors in the laboratory. In: *Errors in linguistic performance. Slips of the tongue, ear, pen, and hand*, Fromkin, Victoria A. (ed.), 307-318. New York & London: Academic Press.

Benveniste, Emile (1954). La classification des langues. *Conférences de l'Institut de Linguistique de l'Université de Paris* 11:33-50. Reprint in: Benveniste, Emile (1966). *Problèmes de linguistique générale*, 99-118. Paris: Éd. Gallimard (Bibliothèque des Sciences Humaines).

Bühler, Karl (1933). Die Axiomatik der Sprachwissenschaften. *Kant-Studien* 38:19-90.

Chomsky, Noam (1964). *Current issues in linguistic theory*. London: Mouton (Janua linguarum series minor, 38).

Chomsky, Noam (1986). *Knowledge of language. Its nature, origin, and use*. New York: Praeger (Convergence).

Coseriu, Eugenio (1958). *Sincronía, diacronía e historia. El problema del cambio lingüístico*. Montevideo: Facultad de Humanidades y Ciencias.

Cutler, Anne (1981). The reliability of speech error data. *Linguistics* 19:561-582.

Dahl, Östen (ed.) (2000). *Tense and aspect in the languages of Europe.* Berlin & New York: Mouton de Gruyter (Empirical Approaches to Language Typology, EUROTYP 20-6).

Gougenheim, G., P. Rivenc, R. Michéa, and A. Sauvageot (1967). *L'élaboration du français fondamental (1er degré). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier (Nouvelle éd. refond. et augm.).

Harris, Roy (1980). *The language makers.* Ithaca: Cornell University Press.

Himmelmann, Nikolaus P. (1993). Material ambitions. *Languages of the World* 7(2):66-68.

Hofling, Charles A. and Tesucún, Félix Fernando (2000). *Itzaj Maya grammar*. Salt Lake City: University of Utah Press.

Iannàccaro, Gabriele (2000). Per una semantica più puntuale del concetto di "dato linguistico": un tentativo di sistematizzazione epistemologica. *Quaderni di Semantica* 21(1):51-79.

Iannàccaro, Gabriele (2001). Alla ricerca del dato. In: *Dati empirici e teorie linguistiche. Atti del XXXIII Congresso Internazionale di Studi della Società Linguistica Italiana,* Federico Albano Leoni, Rosanna Sornicola, Eleonora Stenta Krosbakken and Carolina Stromboli (eds.), 23-35. Roma: Bulzoni (SLI, 43).

Labov, William (1982). *The social stratification of English in New York City,* 3rd printing. Washington: Center for Applied Linguistics.

Lastra, Yolanda (ed.) (1974ff). *Archivo de lenguas indígenas de México.* México, DF: Colegio de México.

Lehmann, Christian (2001). Language documentation: a program. In: *Aspects of typology and universals,* Walter Bisang (ed.), 83-97. Berlin: Akademie Verlag (Studia Typologica, 1).

Lehmann, Christian (in press). Documentation of grammar. In: *Lectures on endangered languages: 3. From Kyoto Conference 2001,* Miyaoka, Osahito (ed.),.xxx-xxx. Osaka: Osaka Gakuin University (ELPR Publication Series C004).

Lieb, Hans-Heinrich (1974). Universals of language: quandaries and prospects. *Foundations of Language* 12:471-511.

Löfstedt, Einar (1942). *Syntactica. Studien und Beiträge zur historischen Syntax des Lateins. Erster Teil: Über einige Grundfragen der lateinischen Nominalsyntax*. Lund: Gleerup (Acta Reg. Societatis Humaniorum Litterarum Lundensis, X:1) (2., erw. Auflage).

Lyons, John (1977). *Semantics*. 2 vols. Cambridge: Cambridge University Press (Rep. 1990-1991).

Seiffert, Helmut (1969-70). *Einführung in die Wissenschaftstheorie. Bd. 1: Sprachanalyse - Deduktion - Induktion in Natur- und Sozialwissenschaften* (1969). Bd. 2: *Geisteswissenschaftliche Methoden: Phänomenologie - Hermeneutik und historische Methode - Dialektik* (1970). München: Beck (Beck'sche Schwarze Reihe, 60-61).

Simone, Raffaele (2001). Sull'utilità e il danno della storia della linguistica. In: *Storia del pensiero linguistico: linearità, fratture e circolarità. Atti del Convegno della Società Italiana di Glottologia, Verona, 11-13 novembre 1999,* Giovanna Massariello Merzagora (ed.), 45-67. Roma: il Calamo.

Stolze, Radegundis (1992). *Hermeneutisches Übersetzen. Linguistische Kategorien des Verstehens und Formulierens beim Übersetzen*. Tübingen: G. Narr (TBL, 368).

Svartvik, Jan and Randolph Quirk (1980). *A corpus of English conversation*. Lund: CWK Gleerup.

Watzlawick, Paul, Janet Beavin Bavelas, and Don D. Jackson (1967). *Pragmatics of human communication. A study of interactional patterns, pathologies, and paradoxes*. New York: Norton & Co.

Weijnen, Anton et al. (eds.) (1975-9). *Atlas linguarum Europae (ALE). 3 vols.* Assen: van Gorcum.