Language about Language about Languages

Meta-concepts for language comparison

Michael Cysouw

Change dedication in localmetadata.sty

Preface

This book is written in QUATERNARY LANGUAGE, that is, these sentences your are reading are language about language about language about language.

To start from the beginning, PRIMARY LANGUAGE is all language that is considered as object of linguistic study, i.e. the parts that are typically typeset in italics. Up one level, the language as used in the field of linguistics is language that discusses such primary language, so any linguistic theory, including all grammatical terminology, is SECONDARY LANGUAGE ('language about language'). In contrast, the topic of this book is TERTIARY LANGUAGE, namely words (and the underlying conceptual ideas) that are important for formulating a linguistic theory. Thus, because the topic of this book is tertiary language, the sentences that I use in this book itself to elucidate these proposals are 'language about tertiary language', i.e. quaternary language.

From this perspective, this book is a contribution to a meta-theory for linguistics. So, I will not be discussing terms like 'adjective' or 'nasality', i.e. the actual bread-and-butter of linguistic theory. Instead, I will present a collection of abstract meta-concepts that I consider important to be clarified and distinguished, especially when dealing with more than a single language — and most crucially for theories comparing languages. Such meta-concepts might seem like an esoteric subject of high theoretical abstraction, and should thus probably better be left for late hours drowned in ample wine, weren't it for the fact that some widespread confusions in current linguistics are based on insufficiently clarified meta-concepts.

To make it easier to talk about these abstract tertiary concepts, I will propose concrete terminology for the various linguistic meta-concepts, like for example 'languoid' or 'counterpart'. The terms that I will use are sometimes newly coined (like 'doculect'), but will mostly be drawn from the existing English lexicon (like 'construction'). The danger of using existing words is that they will always evoke different denotations and connotations with different readers. Yet, whatever somebody might consider to be the meaning of a word like 'construction', I invite any reader to make an effort not to mingle their own understanding of such

a term with the proposals made here. In effect, an explicitly introduced technical term like 'construction' should be read as a mnemonic device for a concept that might also have been called 'gobbledygook' or which could be referred to as 'Concept 3.2' (because all explicitly defined meta-concepts will be numbered throughout this book). Whether the terms themselves are well-chosen might be criticised, but any such discussion is actually 'just words'. It is the underlying concepts and conceptual distinctions on which I invite scholarly scrutiny and discussion.

Because of the meta-meta-meta-level of each sentence in this book, the reading might at times be terse and demanding. But also the writing itself was tedious. Every word that I tried to use seemed to open up yet another can of worms of scholarly debate, or backfired with resonance onto a different level of meta-ness. I have rephrased almost everything multiple times to remove inconsistencies and to divert around pitfalls as much as possible. If the result feels artificially concocted, it is because I tried to be as concise and precise as possible. If the text does not run smoothly, a soothing prospect is that — at least — it is short.

This book is a very personal take on linguistics, its conceptualisation and its terminology. If this view offers an insightful outlook, it is only because I am standing on the proverbial shoulders of giants while writing this. Extensive discussion with Martin Haspelmath, William Croft, Leon Stassen, Balthasar Bickel, Bernhard Wälchli, Johann-Mattis List, Steve Moran and Jeff Good have had an enormous impact on my thinking, and more often than not they might recognise their ideas in this book. I have tried to give the credit where I remember it, and if I forgot to mention any, I know you will take it as the sincerest form of flattery.

The main goal of this book is not to retell the fine minutiae of the scholarly discussion about the terms and the concepts presented. Instead, I will to try to present a more or less coherent and interconnected ensemble of linguistic meta-concepts. I will thus not go back to Aristoteles and provide a detailed discussion about who all else used comparable terms (though often with different conceptualisations) nor who all used similar concepts (though often using different terms), nor who had exactly which influence on my thinking as presented in this book. The main text will be just me talking, though I will add pointers to personal communication and to previous discussion in the literature by using footnotes.

Contents

1 Form and meaning
Setting the perspective on language

Concept 1.1. The term **LANGUAGE** (without preceding determiner) refers to the phenomenon studied in the field of LINGUISTICS, while **A LANGUAGE** (with preceding determiner) refers to a specific instantiation of language, which is the (primary) object of PHILOLOGY.

The conceptual distinction made here between the fields of LINGUISTICS and PHILOLOGY is introduced here to clearly separate specific goals of scholarship. Both approaches are equally needed to obtain further insight in activities like speaking, listening, signing, writing, reading, etcetera. Scholars of these topics never seem to be exclusively a linguist or a philologist. Everybody seems to be both at the same time, though the majority appears to put an emphasis on the philological side. This emphasis is deservedly so, because without philology (i.e. in-depth study and analysis of actual utterances in specific languages) linguistics (i.e. any higher abstractions on language in general) would be neigh impossible. This meta-methodological essay deals with linguistics, but the explicit assumption is that it is only possible to obtain more general insights into language by studying many different languages and comparing their differences and similarities.

Concept 1.2. Two perspectives on LANGUAGE (→ 1.1) can be distinguished. Language can be investigated both from a **NEUROLOGICAL-COGNITIVE PERSPECTIVE** (or 'inside-human' perspective), which includes the pervasive human qualia of MEANING (→ 1.4), as well as from a **PHYSICAL-LINGUISTIC PERSPECTIVE** (or 'outside-human' perspective), which includes the many facets of LINGUISTIC FORM (→ 1.3).

Both the neurological-cognitive perspective (i.e. 'what happens inside a human being when language is used') and the physical-linguistic perspective (i.e 'what is the structure of language when it is passed between humans') are crucial to obtain a full understanding of language. These two perspectives can be seen as investigating two different phases of language. Both phases of language exist simultaneously and influence each other, both are necessary for language to

function, and none can exist without the other. Yet, the perspective on language in this essay will be strongly biased towards the physical-linguistic perspective, i.e. the physical FORM of language.

What happens during the neurological-cognitive phase is not easy to observe. In the fields of psycholinguistics and neurolinguistics a large arsenal of methods has been developed to tackle questions about what happens 'inside' a human being when language is used. However, any data about this phase of language remains hard to obtain, requiring extensive machinery and highly sophisticated experimental methodologies. In contrast, the physical-linguistic phase of language is extremely easy to observe. Humans do not seem to want to stop exchanging language, be it in the form of sound waves, written language, sign language or in any other physical instantiation.

Concept 1.3. The **FORM** of language is any physical instantiation (outside the human body) of language.

Concept 1.4. The terms MEANING (mostly used for lexical elements), FUNCTION (mostly used for grammatical morphemes and constructions), and SENSE (mostly used for subdivisions of meanings/functions) are not differentiated here. They are all equated here with **EXTENT OF USE**.

The experience of meaning is shared among all humans: the definitive feeling that we understand and know what is the meaning of some sound waves entering our ears, or some scribbled symbols seen through our eyes. This experience seems so natural and easy, and we seem to agree

This works because language is taken here as an 'outside-human' object

Meaning is the big problem: we have just a vague grasp of what meaning really is. Forms and their extent of use are much more empirical concepts to work with

semasiological vs. onomasiological approach: Because we don't know too much about meaning, mostly semasiological

Concept 1.5. The terms POLYSEMY (i.e. different but related meanings are expressed by the same form), HOMONOMY (i.e. one form expresses a range of unrelated different meanings) and SYNCRETISM (i.e. one form expresses multiple meanings which had different forms in an earlier stage of the language) are not differentiated here. They are simply taken to be cases of **ONE FORM WITH MULTIPLE USES**.

assumption: every form is homonymous/polysemous!

homonomy has the intuition of 'unrelated' meanings being combined into one form. However, it is very difficult to decide on 'similarity' between meanings. See semantic maps!

(theory of utterance selection, exemplar approach, etic/emic distinction)

[notetoself] dogma (basic assumption underlying a large body of scholarship)-assumption (defining property of what something means) — hypothesis (note wikipedia: "a proposed explanation for a phenomenon", so I actually mean 'prediction' or 'expectation', as according to wikipedia "Any useful hypothesis will enable predictions by reasoning").

Also: generalizations/parameters — model — theory

2 Doculect and languoid

Disentangling the concept of a language

3 Utterance and construction
Language-specific meta-concepts

The central focus of much research in linguistics is the investigation of the structures of individual languages. At the same time, the larger aim of most linguistic research is to rise above the details of single languages and formulate theories about human language in general (→ 1.1). Yet, the main tenet of this book is that these two goals should conceptually (though not personally) be clearly separated. This chapter introduces the central language-specific meta-concepts, while the subsequent chapters will discuss language comparison.

Concept 3.1. An UTTERANCE is an concrete well-formed instantiation of language usage actually attested in a DOCULECT (→ **??**), normally being a complex construal of various interconnected CONSTRUCTIONS (→ 3.2).

Utterances are the basic data for any linguistic investigation. An utterance can be very short and simple (e.g. just saying *yes* in English), but it normally is a complex hierarchical combination of various interconnected constructions.

By definition, an utterance is a one-of-a-kind affair, as each utterance is actually uttered in a concrete time and place, and its interpretation can only be assessed in relation to the specific situation of utterance. In practice, the field of linguistics often abstracts away from this detail and assumes that it is possible to investigate an utterance in isolation or with only minimal context.[1] Such an abstract utterance might be called a SENTENCE. In many cases this abstraction does not lead to unwarranted interpretations, but it amounts to a (possibly misleading) simplification. This is most clearly exemplified by the widespread occurrence of sentences with many different interpretations, depending on the context of utterance. Utterances are almost never ambiguous, in contrast to sentences.

Utterances are in principle always 'well-formed', i.e. every attested utterance is a valid object of (primary) language to be studied in linguistics (i.e. in secondary

[1] There is a growing awareness that linguistics can also profitably be directly based on individual utterances, or EXEMPLARS. Such a token-based approach is obviously taken in research using corpora. In the context of language change, see the THEORY OF UTTERANCE SELECTION (Croft 2000)

language). This assumption might be referred to as the DESCRIPTIONAL DOGMA of modern linguistics. Note that there is a bewildering number of reasons a speaker might bring up to disqualify an utterance as being 'not well-formed.' In principle, each such reason is a respectable (secondary) reflexion on the speaker's own language, which is a piece of information that should always be taken serious by linguistic theory (even if eventually dismissed by explicit argumentation). Also note that in much of modern linguistics the speaker and the linguist are the same person. This combination makes utterance and reflection much closer and quicker, though it also includes the danger of influence from theoretical expectation on the reflection.

Utterances belong to a specific doculect ('language'), i.e. utterances are language-specific. However, note that a doculect might of course be a mixed language, or an instance of code-switching, which can lead to different 'languages' being combined into a single utterance. This will still be called 'language-specific', actually only to prevent the more cumbersome-sounding 'doculect-specific' (→ **??**).

Concept 3.2. A **CONSTRUCTION** is an abstract language-specific combination of CODING DEVICES (→ 3.3) and other embedded constructions, which is used recurrently with a reasonably coherent set of meanings/functions. Constructions typically also include SLOTS (→ 3.4) to be filled by embedded constructions.

The term 'construction' is used in a bewildering number of meanings in current linguistics, so there is a great potential for confusion when using this word. Still, I have decided to retain this term here, although the current conceptualisation will add yet another variation to the meaning of this word.[2]

A construction will here be understood to be a combination of language-specific elements that are used recurrently in a particular language. So, in its most simple instantiation, a construction can be a word, a morpheme, or a stock phrase. More interestingly, constructions are often combinations of such elements including 'slots' to be filled (possibly recursively) by other elements. For example, the English Progressive is a construction consisting of a form of the verb *to be* together with a slot for a verb, which has a suffix *-ing*.

Actually, all language-specific grammar can be seen as such nested constructions, and I think that it is possible to profitably discuss divergent proposals as made in different approached to grammar in current linguistics, and reach consensus in the majority of cases. Probably the most difficult issue preventing consensus in current linguistics is the question how much unification of construction

[2] The current concept of constructions is close the the one in Croft (2001). It also seems to be basically same as Haspelmath's (2010:664) DESCRIPTIVE CATEGORIES.

is deemed necessary. Constructions are always abstractions over actual variation, but some (more 'descriptional') linguists might prefer separate listing of less abstract constructions, even when they are only slightly different, while others (more 'theoretical') linguists would rather unify many constructions into one underlying more abstract construct, often invoking some kind of movement to derive the actual constructions.

There are a few important definitional characteristics of the current conceptualisation of constructions. First, all constructions are language-specific. By current definition there are no constructions that occur in more than one language — but of course note the difficulty of what actually counts as a single language (→ ??). Not all is lost, though. There are various possible similarities between constructions from different languages: they might be homologous (→ ??), or they might belong to a similar strategy (→ 5.5) or pattern (→ 5.4). But they are never the same construction. Second, constructions are linguistic abstractions, i.e. they are secondary language. Instances of constructions are used in actual utterances, or, formulated reversely, a construction is a grouping of all occurrences in actual utterances. Third, the identification of a construction is not self-evident. I would even coin the proper recognition of a constructions to be a 'discovery.' Scholars will and should discuss about the preferred description of any construction; reconsider, reformulate, and refute them until ideally a consensus arises.

Finally, as stated in the definition above, constructions are normally expected to express a 'reasonably coherent' set of meanings. The reason for this rather vague formulation is that it is extremely difficult (if not impossible) to propose a general analysis of meaning on the basis of just a single language (→ 1.4).

Concept 3.3. A CODING DEVICE is an abstract minimal functional construction of a specific language.

The term 'coding device' is proposed here for a slightly more general conceptualisation than the traditional notion of a morpheme, i.e. the minimal meaning-bearing (or function-bearing) unit of language. Typically, coding devices are individual morphemes or lexemes, but they can also be non-linear morphology (tonemes, ablaut, umlaut, metathesis, etc.) or even specification of relative ordering of parts, and amount of morphological fusion between parts.

In empirical practice, coding devices are the minimal building blocks necessary to formulate all proposed constructions. Formulated differently, coding devices are the endpoints ('leaves') of the hierarchical structure of constructions in sentences ('trees'). As such, all coding devices are also constructions, and the introduction of this new term is mainly for convenience.

There is a widespread assumption in linguistics that each utterance can be broken down into a disjunct ('non-overlapping') set of coding devices. This assumption that there is a single underlying separation of each utterance into a set of smallest building blocks can be called the SINGLE-LEVEL CODING HYPOTHESIS.

Concept 3.4. A **SLOT** is a part of a construction that *defines* a class of constructions (typically lexemes), namely, all constructions that can fill the slot belong to the class.

The idea of a slot in a construction is often defined reversely from the above formulation. Namely, by assuming that a language has classes of coding devices (e.g. 'word classes' or 'inflectional classes'), then a slot can be characterised by a specific such class of elements that are allowed to occur in the slot. The problem with that approach is that there is no independent definition of the classes themselves. By reversing the definition (as proposed here), classes are *defined* by a specific slot in a construction. That implies that each slot in each construction *a priori* defines a separate class, and it becomes an empirical question how tightly all these classes fit together into a single overarching word class division for the whole language. The assumption that each language has a small set of overarching classes can be called the WORD CLASS HYPOTHESIS.

Concept 3.5. The **BEHAVIORAL POTENTIAL** of a construction (typically a lexeme) is defined as the collection of slots that the construction can fill.

Behavioral potential is

4 Comparative concept and counterpart
Comparing languages

To be able to compare data from different languages, it is necessary to make sure that the data is comparable across languages. It is crucial to realise that grammatical terminology like 'subject', 'genitive' or 'diminutive' are not sufficiently defined by tradition to be automatically useful across languages.[1] The central prerequisite for comparability is an explicit definition of the topic to be compared.

Concept 4.1. A DOMAIN is a cross-linguistic topic of investigation. A domain consists of three different parts, namely (i) a name for the domain, (ii) an explicit definition, and (iii) the actual data selected for comparison. The name is called DOMAIN LABEL, the explicit definition is called COMPARATIVE CONCEPT (→ 4.3), and the data selected for comparison is called COUNTERPART (→ 4.4).

Names for domains are just labels that have to be explicitly defined anew for every comparative investigation. Unfortunately, the same labels are often used both for language-specific constructions (→ 3.2) as well as for cross-linguistic domains (→ 4.1). To avoid confusion between the two, it is good practice to clearly distinguish DOMAIN LABELS from CONSTRUCTION LABELS. When writing in English, it is possible to use capitalisation to differentiate the two. Cross-linguistic domain labels are commonly written in lowercase (e.g. 'diminutive'), while language-specific construction labels start with a capital letter (e.g. 'Diminutive'). To prevent any possible confusion, it is strongly preferred to additionally add a language name (→ **??**) to a construction label (e.g. 'Dutch Diminutive').[2]

[1] The most forceful recent plea against the assumption that such terms are automatically well-defined is Haspelmath (2010). However, the necessity to explicitly define the topic of a cross-linguistic investigation has been recognised throughout recent decades of scholarship. REFERENCES.

[2] This tradition originated with Comrie (1976). This capitalisation trick cannot be used in all orthographic traditions. For example, German nouns always have to be capitalised, so it is not possible to use a lowercase nouns in regular German orthography. The addition of a language name becomes crucial in such situations.

Concept 4.2. Two domains are **SUBDOMAINS** of an overarching domain when languages recurrently use the same counterpart (→ 4.4) to express both subdomains.

Cross-linguistic research often investigates the relation between various domains. Or, formulated differently, it investigates the structure of subdomains within a single domain (→ 5.4). For example, the domains 'instrumental' (how does a language express that something is used as an instrument) and 'accompaniment' (how does a language express that a participant is accompanying another participant) are strongly linked into an overarching domain, because many languages use the same construction for both (e.g. English *with*).[3]

In current linguistics, the OPTIMAL DOMAIN HYPOTHESIS seems to be widespread. This hypothesis says that it is possible to define a finite number of domains that are necessary and sufficient for the analysis of all human languages.[4] Whether this hypothesis holds is an open question. There are at least two different ways in which it might not be true. First, there might turn out to be just one domain, i.e. all domains proposed might turn out to be subdomains of one overarching domain. Or, second, there might turn out to be an infinite number of sensible domains.[5]

Concept 4.3. A **COMPARATIVE CONCEPT** is a definition of a domain (→ 4.1), which is primarily meaning/function-based definition, but which frequently includes additional form-based constraints. This definition is used to select CONSTRUCTIONS (→ 3.2) from the different languages to be compared. The set of constructions selected from a single language is called a COUNTERPART (→ 4.4).[6]

A comparative concept is necessarily universally applicable across all human languages. If it turns out that a specific comparative concept is not applicable in certain languages, then the definition has to be revised. This approach can be called the PRINCIPLE OF UNIVERSALLY APPLICABLE DOMAIN. To make them universally applicable, comparative concepts are mostly based on meaning/function.

[3] Subdomains are also known in the literature as SENSES (Haspelmath 2003), ETIC GRID (Levinson & Meira 2003) or ANALYTICAL PRIMITIVES (Cysouw 2010).

[4] This hypothesis is approximately the same as CATEGORICAL UNIVERSALISM from Haspelmath (2010:663).

[5] Note that the observation that there is more than one sensible definition of the term 'tense' (an example discussed in Haspelmath 2010:679) does not contradict this hypothesis. Maybe linguists can agree at some point that we need, say, 17 different definitions of 'tense' for a full fledged analysis of human language (though hopefully using different labels), then the hypothesis is still true.

[6] The term COMPARATIVE CONCEPT was proposed by Haspelmath (2010). In much of the typological literature, the term TERTIUM COMPARATIONIS is used with the same intention.

Such definitions are most easily applicable across the widely different constructions as attested in the world's languages. Additional form-based constraints can be used, but care should be taken to formulate such constraints in a universally applicable manner. There appear to be only very few universally applicable form-based criteria, possibly not much more than (i) size of construction, (ii) number of counterparts, (iii) relative position construction parts, and (iv) extent of morphological fusion of parts (→ 5.3). In the practice of worldwide language comparison, the process of formulating and adjusting the definition of a domain is a central part of proper research. The details of the definition itself are often more important than their actual empirical application.

Further, there is a PRINCIPLE OF UNRESTRICTED DOMAIN CHOICE, which simply means that there are no *a-priori* reasons to restrict the kind of definitions that are allowed. It is important to realise that this principle does not imply that every method is equally suitable. However, whether a definition is useful can only be judged *ex post*, i.e. by judging the usefulness of any insights that arise from the definition.[7] There are various *ex post* reasons that might be brought up as an argument for a specific comparative concept. For example, someone might argue that comparative concepts should be chosen in such a way that the counterparts form a coherent category in all individual languages. The claim that this is possible might be called HOMOGENEOUS COUNTERPART ASSUMPTION. A different approach might argue that definitions of domains should be based on (non-linguistic) insights into the functioning of the neurological-cognitive phase of language (→ 1.3). This might be called the COGNITIVE GROUNDING ASSUMPTION.

Concept 4.4. A **COUNTERPART** is a set of language-specific constructions (→ 3.2) — though in practice often just a single construction — that is selected from a language on the basis of a comparative concept (→ 4.3).

A comparative concept is a definition that specifies how to select constructions (possibly one, possibly more than one) from a language. The resulting set of constructions for each language will be called a COUNTERPART.[8] The central procedure of language comparison is to compare counterparts. The technique to compare counterparts across languages is not automatically obvious, because counterparts are language-specific constructions (for a full discussion, see Chapter 5).

In the field of linguistic typology, the counterparts are often extracted from reference grammars or by using questionnaires to be filled out by specialists in

[7] Haspelmath's (2010) CATEGORICAL PARTICULARISM seems to be a similar principle.

[8] The term COUNTERPART is proposed by Good.

the languages investigated. The problem to properly define comparative concepts and select counterparts can also be solved by using experimental stimuli with visual context (e.g. pictures or movie clips), or by using linguistically expressed context in the form of parallel texts.

Comparative concepts and counterparts are also used in other subfields of linguistics. In quantitative historical linguistics, the 'Swadesh-style' WORDLIST APPROACH is widespread. Using the terminology as developed in this chapter, the term 'wordlist' can now be clarified. The starting point is a set of lexical comparative concepts, or CONCEPTLIST. In most practical instances, these comparative concepts are rather ill-defined, often just relying on individual English (or Spanish or Russian) words with minimal explicit definition. This is the same fallacy as discussed at the start of this chapter: assuming that domain labels are sufficiently defined automatically — which they mostly are not. A WORDLIST is then the list of counterparts in a specific language.

5 Shape, strategy and pattern
The structure of a typology

The central problem of LINGUISTICS (→ 1.1) is to compare constructions from different languages to each other. This comparison poses a problem in a very trivial sense, namely in that linguistic elements from one language are different objects from elements in another language. A conceptually simple kind of comparison across languages (though empirically quite tricky to establish) is to investigate whether two CONSTRUCTIONS (→ 3.2) are historically related. For example, the English suffix *-ly* is historically 'the same' element as the German suffix *-lich*. These constructions are HOMOLOGS (→ **??**). However, such a comparison of homologs is only possible for a rather limited number of constructions, because the languages in question have to be quite closely genealogically related, or the construction has been borrowed from one language to another. These kinds of comparison will be discussed in more detail in Chapter 7.

This chapter deals with a more abstract kind of comparison that has become known under the term 'linguistic typology'. Starting from a universally applicable DOMAIN (→ 4.1), this approach to language comparison defines universally applicable TYPOLOGICAL CHARACTERISTICS (→ 5.2) to compare constructions across languages, leading to a TYPOLOGY (→ 5.1).

Concept 5.1. A **TYPOLOGY** is a METRIC (→ 9.1) on COUNTERPARTS (→ 4.4). Such a metric is defined on the basis of TYPOLOGICAL CHARACTERISTICS (→ 5.2). In the most basic instantiation, a typology is just a partitioning of all counterparts into a few discrete groups, so-called TYPES.

Counterparts are the linguistic expressions from different languages to be compared to each other (for more details, see chapter 4). The actual comparison of these counterparts can take many different manifestations. In the most abstract sense, any such comparison is a kind of metric, i.e. the mathematical concept of distance (or its inverse, similarity). A more extensive discussion of different kinds of metrics can be found below (→ 9.1).

Counterparts can be more or less similar to each other, and the decisions about their similarity constitutes the typology. In the simplest possible metric, all

comparisons between counterparts lead to only two outcomes: either 'the same' (i.e. the counterparts belong to the same type) or 'different' (i.e. they belong to different types). Under such a metric, all counterparts end up in different groups. Within each group all counterparts are 'the same'. Such groups are traditionally called TYPES, hence the widespread use of the term TYPOLOGY for this endeavour.

Such measurement of the similarity between counterparts has to be based on characteristics that are universally applicable across all languages (→ 5.2). To establish such characteristics is not trivial because many conceptual ideas about language often turn out to be strongly influenced by the few languages that are intimately familiar to a scholar. The central, and often most arduous, step in establishing a typology is actually the proper definition of the typological characteristics. The importance of such definitions should not be underestimated: only slightly different definitions will already often lead to rather divergent typologies.

A typology is thus a metric on counterparts, and each counterpart (→ 4.4) is a (possibly diverse) set of constructions (→ 3.2) from a languoid (→ **??**). A seemingly small detail of the practical establishment of such typologies is that a counterpart sometimes consists of multiple constructions. How to deal with such more complex counterparts is often non-trivial, but those complications will be ignored here. In what follows I will silently assume that each counterpart from each languoid will consist of a single construction.

Concept 5.2. A TYPOLOGICAL CHARACTERISTIC is a universally applicable property of a counterpart. There are at least three different kinds of such properties: SHAPE (→ 5.3), PATTERN (→ 5.4) and STRATEGY (→ 5.5).

To establish a typology in practice, all counterparts are analysed according to some typological characteristics. Many typologies are based on just a single typological characteristic. For example, a single typological characteristic that distinguishes three different strategies (→ 5.5) trivially leads to a typology with three types. However, many different typological characteristics about the same counterparts can also be combined into a MULTIVARIATE TYPOLOGY.[1] Looking forward, this seems to be a much more fruitful approach, namely to start off distinguishing many different, often very simple, typological characteristics, and to derive more complex typologies on that basis.

In its most extreme form, a typology might include very many different, highly detailed typological characteristics, which might lead to a situation in which all counterparts are different from all other counterparts. In such a typology each

[1] Bickel

language is its own type, leading in effect to a TYPOLOGY WITHOUT TYPES. Such a typology might seem to defy the rationale of language comparison, because it only seems to tell us that all languages investigated are different. However, not all counterparts are equally different. In a highly detailed multivariate typology some counterparts are more similar to each other than they are to others. This metric on counterparts represents an empirical restriction of the theoretically possible linguistic diversity.

Concept 5.3. A typological **SHAPE** is a TYPOLOGICAL CHARACTERISTIC (→ 5.2) classifying the FORM (→ 1.3) of a COUNTERPART (→ 4.4).

There is only a very limited number of possibilities to compare the SHAPE of counterparts across all the world's language. There are at least four different kinds of typological characteristics of form — and probably not more: (i) size of constructions, (ii) quantity of counterparts, (iii) relative position of construction parts, and (iv) extent of fusion of construction parts.

First, the SIZE OF A CONSTRUCTION can, for example, be measured pronunciation time, in number of phonemes or in number of coding devices (→ 3.3). Low values on such a measurement indicate that the domain under investigation is deeply ingrained in the structure of those languages ('strongly grammaticalised'). In reverse, a large size of a construction indicates that some form of circumlocution is apparently necessary.

Second, the QUANTITY OF COUNTERPARTS is established by counting the number of counterparts that are selected on the basis of a comparative concepts (→ 4.3). This is a measure for the explicitness with which a language deals with the domain investigated (also known as paradigm size or paradigmatic diversity). Some randomly chosen examples are the number of phonemic vowels, the number of different grammatical cases, or the number of different past tense constructions.

Third, the RELATIVE POSITION OF CONSTRUCTION PARTS is of course best exemplified with the infamous word-order typologies. For such a typological characteristic, the comparative concept (→ 4.3) needs to specify different parts within each counterpart whose order can then be investigated. For example, within the domain of possession, one can investigate the order of the possessor and the possessee.

Finally, the EXTENT OF FUSION OF CONSTRUCTION PARTS is a characteristic typically used to distinguish morphologically bound marking from marking by separate words (also known as synthetic vs. analytic marking). For example, in the domain of tense marking, a common distinction made is between marking tense as a morphological category on the verb vs. marking tense using separate particles or adverbs.

Concept 5.4. A typological **PATTERN** is a TYPOLOGICAL CHARACTERISTIC (→ 5.2) classifying the similarity between counterparts (→ 4.4) within the same language. To establish such a pattern it is necessary that the DOMAIN (→ 4.1) is subdivided into SUBDOMAINS (→ 4.2). The similarity between the counterparts of these subdomains represents the pattern.

Patterns are very powerful typological characteristics, which also play a central role in semantic maps (see Chapter 6). The basic idea can be exemplified by the hand-arm polysemy as attested in various languages.[2] This characteristic can be analysed as being based on two subdomains, roughly identified as the body parts called *hand* and *arm* in English. The expression of these two subdomains can show two patterns: either the two subdomains are expressed identically, or they are expressed differently. A second example is notion of ergativity, based on three subdomains, leading to five different possible patterns. Ergativity is based on the subdomains commonly called S ('subject'), A ('agent') and P ('patient'). Depending on the polysemy between the constructions used for these three subdomains, there are theoretically five different possible polysemy patterns: nominative/accusative (A=S≠P), ergative/absolutive (A≠S=P), tripartite (A≠S≠P), neutral (A=S=P), and the apparently unattested transitive/intransitive role marking (A=P≠S).[3]

The number of patterns grows very fast with the number of subdomains. Two subdomains allow for two patterns, three subdomains for five, four subdomains lead to 15 different patterns, and five subdomains generate 52 different patterns. With 10 subdomains there are already more than 100.000 different theoretically possible patterns. The number of patterns for n subdomains is the so-called nth Bell Number B_n. The size of B_n grows very fast, even faster than exponential.[4]

Another way to look at patterns is to interpret a single pattern as a highly detailed multivariate typology. A pattern can be described as a large collection of characteristics, namely one characteristic for each pair of subdomains. For example, with 3 subdomains, there are $\frac{3*2}{2} = 3$ different *pairs* of subdomains, or more general: with n subdomains, there are $\frac{1}{2}n(n-1)$ different pairs. Specifying the similarity for each of these pairs completely describes the pattern.

In the examples above, patterns were exemplified as polysemy-patterns between subdomains, i.e. each pair of subdomains either use the same construction,

[2] Ref to WALS

[3] Sapir 1917

[4] There is no closed-form formula for the calculation of Bell Numbers. For all details, see the On-Line Encyclopedia of Integer Sequences at http://oeis.org/A000110. As for the rate of growth: technically speaking, for large enough n (specifically $n > 4$) it holds that $2^n < B_n < 2^{n^2}$.

or a different constructions. This is of course only the most basic approach: it is equally possible to use fractional values of similarity between two constructions (i.e. constructions within a single language can be more or less similar to each other). In the words of the definition of a pattern as formulated above (→ 5.4), a pattern is a metric between the expression of the subdomains. Or, slightly reformulated, a pattern is a metric *within* a domain.

The empirical appeal of typological patterns lies in the fact that to establish a pattern it is only necessary to compare constructions *within* a single language to each other. Patterns thus provide a ingenious solution to the problem of comparability, because no actual comparison between constructions from different languages is necessary.

Concept 5.5. A typological **STRATEGY** is a TYPOLOGICAL CHARACTERISTIC (→ 5.2) that classifies COUNTERPARTS (→ 4.4) on the basis of other occurrences of the counterparts outside of the current DOMAIN (→ 4.1).

A typological strategy is a widespread 'shortcut' to classify counterparts. After counterparts from many languages are collected for a specific domain, these counterparts are classified in groups (i.e. partitioned) on the basis of in what kind of other uses the counterparts are attested. For example, a specific possessive construction can be called 'locative' because the possessor is marked with a case or adposition that also has locative functions.[5] I consider such typological strategies to be 'shortcuts' because the other uses are often much more vaguely defined than the domain itself, and the relation between those other uses is mostly not included in the discussion. From this perspective, typological strategies could also be called 'centric patterns' (→ 5.4), because with strategies only polysemy relative to a single central domain is investigated, ignoring the polysemies between the surrounding uses.

[5] stassen in WALS

6 Meaning and semantic space
Mapping language diversity

meaning as extend of use
language-specific: contextual vectors
cross-linguistic: typological patterns
semantic space as the ‘average’ extend of use (often using dimensional reduction technique — note that semantic space is very likely to be curved!)
semantic map as the language-specific intstantation

7 Homology and stratum
Historical language comparison

historical process leading to similarity

homology: historically related form calque: historically related meaning (important for creoles, and large-areal consistencies‼!) stratum: point of transfer

note: 'all words that start with [p]' (ringe 1992) is strictly speaking a typological characteristic! It is not a historical comparison, because it does not take into account sound change.

homology, recent stratum: loan homology, old stratum: cognate calque: always loan?

8 Code point, glyph and grapheme

Writing systems across languages

Using letters

- orthography vs. transcription (language-specific vs. comparative)
- using simplified transcription in historical comparison
- glyph/character vs grapheme (etic vs. emic)
- tailored grapheme cluster

9 Measurement and metric
Towards quantitative language comparison

Concept 9.1. Metric: definition of distance (similarity) between elements. Specification traditionally as a formula, but in practice often a large table with all distances, often semi-manually specified.

distance matrix / similarity matrix (Gramian, kernel)