elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 7*. Editor: Peter K. Austin

Language documentation and language typology

OLIVER BOND

Cite this article: Oliver Bond (2010). Language documentation and language typology. In Peter K. Austin (ed.) *Language Documentation and Description, vol 7*. London: SOAS. pp. 238-261

Link to this article: http://www.elpublishing.org/PID/087

This electronic version first published: July 2014

EL Publishing

Language documentation and language typology

Oliver Bond

1. Introduction

Typology is a sub-discipline of linguistics originally conceived around the notion that there is a fundamental basic unity underlying the diversity of the world's languages. Typologists believe that there are certain core properties that languages have in common which can be formulated as generalisations about language in the form of language universals or probabilistic statements about the distribution of language characteristics. One widely cited language universal first proposed by Greenberg (1963/1966) concerns the relative order of subjects (S), verbs (V) and objects (O) in relation to the type of adposition (preposition or postposition) found in languages:[1]

(1) *Languages with dominant VSO order are always prepositional*

The universal in (1) makes the prediction that adpositions will always precede the noun phrase they govern (i.e. will be prepositions) in languages where the favoured order of major constituents in a transitive clause is one where the verb precedes the subject, and the subject precedes the object. Despite the degree of certainty associated with this prediction (suggested by the use of 'always'), most 'universals' are not **absolute** in that they have exceptions, including Greenberg's universal in (1). If you know of a language that has dominant VSO order, yet has postpositions rather than prepositions, or indeed a mixture of both, you are not alone: several exceptions to early universal statements of this kind have since been documented in the typological literature (e.g. see Dryer 2008a, 2008b).[2] In the face of such patterns, contemporary typology is more concerned with the probabilistic (not absolute) statements linguists can make about similarities across languages.

---

[1] This is Universal 3 in Greenberg (1963/1966).

[2] For instance, data used in Dryer (2008a, 2008b) indicates that languages with VSO order and postpositions include Majang (Nilo-Saharan; Ethiopia), Northern Tepehuan and Cora (Uto-Aztecan; Mexico), Koreguaje (Tucanoan; Colombia), Taushiro (Isolate; Peru) and Guajajara (Tupian; Brazil). VSO-dominant languages with both prepositions and postpositions include Murle and Tennet (Nilo-Saharan; Sudan) and Makah (Wakashan, USA).

Oliver Bond 2010. Language documentation and language typology. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 7,* 238-261. London: SOAS.

One important aspect of probabilistic statements is that they cannot be discovered, verified or quantified by studying only one language: by their very nature, probabilistic statements are derived through the examination of large samples of language data (see Song (2001: 17-41) for a sophisticated introduction of sampling issues). Typologists are thus concerned with generalisations that hold cross-linguistically. In the broadest typological work, the conclusions that are reached are based on unbiased samples that include languages spoken throughout the world. However, researchers engaged in language documentation are often most interested in the **typological characteristics** of language groups or language areas. Such linguistic traits (e.g. concerning constituent order, agreement properties, negation strategy, etc.) become of interest when a given language diverges from the normal characteristics of the group to which it belongs, whether a genetic or areal unit.

At its onset, typology was largely concerned with what is possible in language, that is, discovering what the universals of language might be. However, contemporary typology has a more sophisticated agenda: not simply asking 'what is possible?', but examining 'what is where why?', with reference to historical and other factors affecting the distribution of language properties (Bickel 2007). For instance, it is no coincidence that certain morphosyntactic or phonological features are common and distributed across the whole world, while others are restricted to a small domain. The distribution of linguistic features may relate to phenomena such as:

- the geographic isolation of a speech community
- migration and trading patterns
- the types of societies that people live in
- the types of language contact situations that prevail.

Sometimes, the distribution of typological features is due to fundamental and consistent properties of language and its communicative goals.

In this chapter we are going to explore some core concepts in typology and examine how they relate to language documentation and description. At the end of the chapter you will have a sense of what typology is, some of the principles that underlie it, and understand the symbiotic relationship between the two research fields.

2. The typological method

Although there are different approaches to typology, the common denominator in traditional typological studies is the aim to uncover the factors underlying the immense diversity of language structures. This is achieved by using an **empirical approach** to the study of language. The central idea behind the empirical method is that conclusions are dependent on evidence or consequences that are observable by the senses. An empirical methodology involves the use of working hypotheses that are testable using observation or experiment.

The typological method involves cross-linguistic comparison. Often this is through a representative sample of the world's languages. It typically involves classification of either (i) components of a language or (ii) languages themselves. A component of a language is a particular **construction** (e.g. relative clauses) or **feature** (e.g. oral plosives) that can be compared across languages. Devising typologies at the constructional level is fruitful because a language may have move than one strategy to achieve a particular goal (i.e. belong to two 'types' at the same time). Languages have also been classified into types based on shared properties. This kind of typology is known as 'holistic' typology. For example, we might say that a certain language is a type X language while another language is a type Y language; holistic typology would propose that type X languages have certain properties, while type Y languages have certain other properties. One popular holistic typology classifies languages into types according to their morphological characteristics. These types are referred to as 'isolating', 'agglutinating', 'polysynthetic' or 'fusional'. Holistic typologies are generally less revealing than those that those that involve implicational relationships between different language parameters.

Typology is primarily concerned with classification based on formal features. Typology does not group languages together into families. Likewise, typology does not classify languages into types based on geographical location, or based on the number of speakers a language has. Typology classifies constructions in languages based on the forms out of which they are composed; these can be at any level, including sounds, morphemes, syntactic constituents, and discourse structure. Since these elements are employed to convey meaning, typologists are naturally concerned with semantic categories such as 'event' and 'agent', which are manifested by formal units of language (see also Sells, this volume).

Typologists form generalisations that are based on observations, so typological research is concerned with the study of patterns that occur systematically across languages. The aim to uncover diversity makes typology unlike a Universal Grammar type model of language, which seeks to abstract away from linguistic diversity to uncover innate restrictions on language. In

addition to creating taxonomies, contemporary typologists also seek plausible explanations for typological patterns in 'extragrammatical' domains such as discourse, pragmatics, physiology, cognition, speech processing, language contact, social influences on language use, etc. The kinds of explanatory models that typologists tend to use include concepts such as competing motivations, economy, iconicity, and semantic maps underlying some sort of conceptual space that speakers have (see Croft (2003) for discussion of these models).

Another significant dimension of typological work is that many grammatical phenomena are fundamentally diachronic, and not just synchronic. Within typology, structures that exist are usually considered from a historical perspective as well because we know that languages change through time, that certain patterns are more time-stable than others, and that unusual typological patterns often arise when a particular construction is in flux between two more common or major types.

So with all of this in mind, we can ask: "is typology a theory?" The kinds of theories that are prevalent in linguistics like Minimalism, Government and Binding Theory, Functional Grammar, Cognitive Grammar, Role and Reference Grammar, and Lexical Functional Grammar are designed to model how language works (see Sells' chapter in this volume for references and discussion). They provide a framework for linguistic analysis. Typology is not really a theory of grammar in that it does not use an abstract architecture to account for or formalise explanations. Rather, typology is concerned with identifying cross-linguistic patterns and correlations between these patterns. For this reason, the methodology and results of typological investigations are (in principle) compatible with any grammatical theory (and with language documentation). Unlike Universal Grammar however, typology is not concerned only with purely innate aspects of language, but also with the communicative and diachronic processes that result in the geographical and genealogical distributions of features across languages. These include:

- population movements and language contact
- socio-anthropological influences on linguistic structure
- cognitive and communicative pressure on processing and acquisition
- grammaticalisation and other historical processes.

All these areas relate to aspects of language documentation in one way or another.

The standard strategy in typological research, as presented by Croft (2003: 14), involves three key steps:

(i) determine the particular semantic(-pragmatic) structure or situation type to be explored;

(ii) examine the morphosyntactic construction(s) or strategies used to encode that situation type;

(iii) search for dependencies between the structures used for that situation and other linguistic factors, including other structural features and external functions expressed by the construction in question, or both.

The first step involves defining a domain of research. The second step requires identifying variation across languages to examine constructions or strategies used to encode that situation type in a variety of languages. The ultimate goal is to search for dependencies between different features of language. Typology is not just the classification of strategies into types, but also concerns which functions can be shared by structures, and what predictions can be made about a language based on its structural characteristics.

To exemplify the application of the standard typological method, we now look briefly at Comrie and Kuteva (2008) who examined the distribution of relativisation strategies in a sample of 166 languages.

2.1 Relativisation on subjects

The first step necessary to conduct typological work on relativisation strategies is identification of the research domain: in this case, what is meant by the term **relativisation**? For Comrie and Kuteva (2008), a relative clause is:

> a clause narrowing the potential reference of a referring expression by restricting the reference to those referents of which a particular proposition is true

Thus, a relativisation strategy is a type of grammatical structure used to restrict the reference of a referring expression of the type identified above. An example of relativisation from English can be seen in (2). This sentence contains a relative clause *who just greeted us* (indicated by square brackets), which narrows the potential reference of the referring expression *the girl* (i.e. the head noun) to referents of which the proposition *the girl just greeted us* is true.

(2) I teach the girl [who just greeted us]

The definition used here is composed to capture a largely semantic-pragmatic function expressed in various ways cross-linguistically, so it is worded to avoid inherent reference to structure, or rather to make as little reference to structure as possible (notice that the notion of 'clause' is relevant to the typology so it must be included in delimiting the research domain).

Comrie and Kuteva (2008) identify four main types of strategy for relative clause formation across the languages in their survey based on empirical observation of how the referent is indicated within the relative clause: the **relative pronoun** strategy, the **non-reduction** strategy, the **pronoun retention** strategy, and the **gap** strategy. The first type of strategy is:

> **Relative pronoun strategy**: the position relativized is indicated inside the relative clause by means of a clause-initial pronominal element, and this pronominal element is case-marked (by case or by an adposition) to indicate the role of the head noun within the relative clause.

The relative pronoun strategy can be exemplified with data from German (Germanic, Indo-European). In (3) the referent whose reference is narrowed by the relative clause is *der Mann* 'the man':[3]

(3) German (Comrie and Kuteva 2008)

Der	*Mann,*	*[der*	*mich*	*begrüßt*	*hat],*	*war*
the	man.NOM	REL.NOM	me	greet.PTCP	has	be.3SG.PST

ein	*Deutscher.*
one	German

'The man [who greeted me] was a German.'
(cf. The man greeted me.)

[3] The abbreviations used in this paper are: ACC = accusative, DAT = dative, DEM = demonstrative , DIR = directional, DIST = distal, EXCL = exclusive, NOM = nominative, OBL= oblique, PFV = perfective, PL = plural, POT = potential, PRS = present, PST = past, PST2= past (second most recent), PTCP = participle, RDP = reduplication, REAL = realis, REL = relativiser, SG = singular, SUBJ = subject.

The relative pronoun *der* is case-marked to indicate the role of the head noun within the relative clause. What makes this strategy distinctive is the presence of a case-marked relative pronoun form at the beginning of the relative clause.

The second strategy is:

> **Non-reduction strategy**: the head noun appears as a full-fledged noun phrase within the relative clause.

In Maricopa (Yuman, Hokan; USA), rather than using a relative pronoun in the relative clause to indicate the referent with restricted reference, the referent is indicated by a noun, as in:

(4) Maricopa (Gordon 1986: 255)

[aany=lyvii=m	*'iipaa*	*ny-kw-tshqam-sh]*	*shmaa-m*
yesterday	man	1-REL-slap.DIST-SUBJ	sleep-REAL

'The man [who beat me yesterday] is asleep.'
(cf. The man beat me yesterday.)

Here, the noun phrase *'iipaa* 'man' is in the middle of the relative clause; it is not expressed outside the relative clause.[4]

The third strategy is:

> **Pronoun-retention strategy**: the position relativised is explicitly indicated by means of a resumptive personal pronoun.

This strategy is found in Babungo (Bantoid, Niger-Congo; Cameroon); note that the resumptive *ŋwɔ́* 'he' in (5) would not appear in a regular main clause. A literal translation of this example into English, would be *I have seen the man who he has beaten you*.

[4] We know *'iipaa* 'man' is inside the relative clause because the surrounding material is incompatible with an analysis in which the head noun is external to the restrictive clause.

(5) Babungo (Schaub 1985: 34)

mə	*yè*	*wə́*	*ntɨ́ə*	*[fáŋ*	*ŋwə́*	*sɨ́*	*sàŋ*	*ghô]*
I	see.PFV	person	that	[who	he	PST2	beat.PFV	you]

'I have seen the man [who has beaten you].'
(cf. The man has beaten you)

Finally, the gap strategy is overwhelmingly the most popular strategy across the world's languages:

> **Gap strategy**: there is no overt case-marked reference to the head noun within the relative clause.

The gap strategy is found in Turkish (Turkic, Altaic). In (6) the head-noun *ögr̆enci* 'student' is outside the relative clause and there is no reference to 'the student' within the relative clause (cf. the relative pronoun and pronoun-retention strategies), i.e. there is a 'gap' the relative clause where the subject should be.

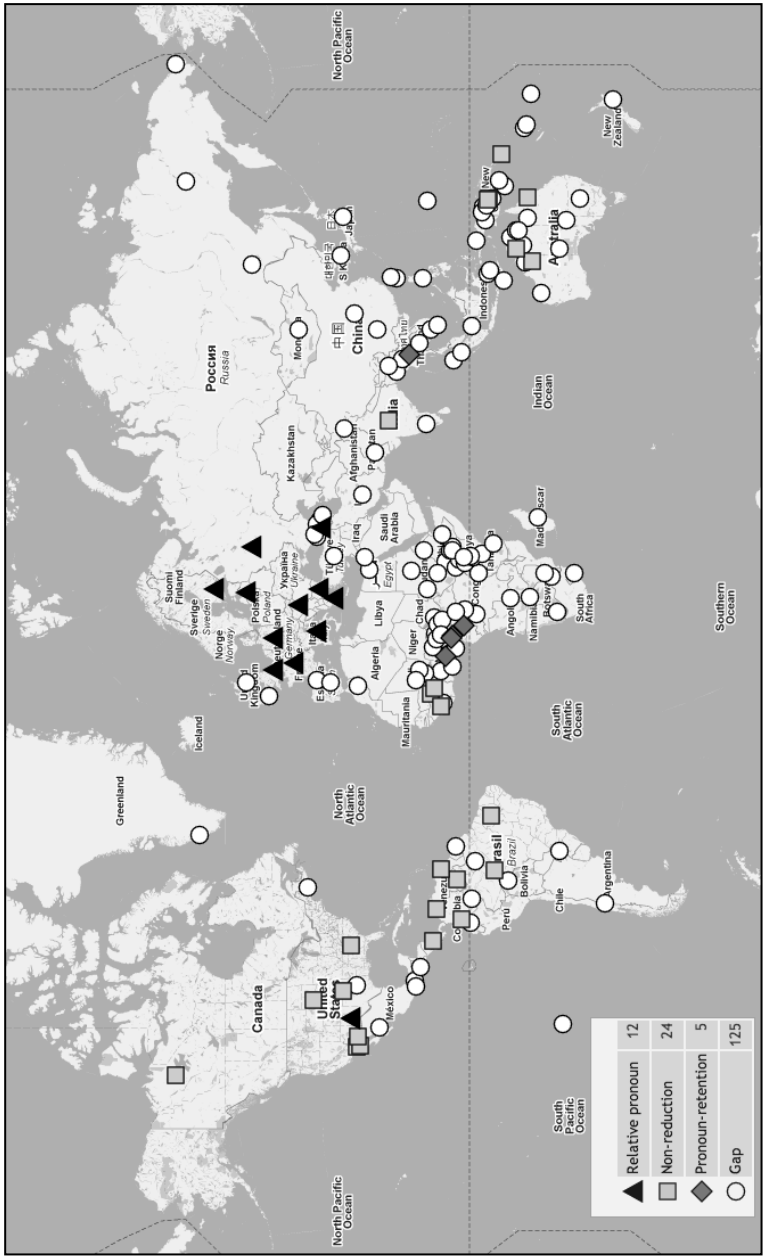
(6) Turkish (Comrie 1998: 82)

[kitab-ı	*al-an]*	*öğrenci*
book-ACC	buy-PTCP	student

'the student [who bought the book]'
(cf. The student bought the book.)

Comrie and Kuteva (2008) also looked at how the four strategies they identified are distributed across their sample of 166 languages.

Map 1. The distribution of four strategies for relativisation on subjects, in 166 languages (Comrie and Kuteva 2008



Each strategy is represented by a different coloured and shaped symbol. Notice that there are many white circles (125 instances in total) which represent the gap-strategy. The other strategies are less frequent. The non-reduction strategy seems to be scattered across the world map (grey squares, 24 instances), but the two remaining strategies are very restricted in their location. Perhaps the most striking and interesting pattern represented on this map is that the type of relativisation strategy familiar to us from English, the relative pronoun strategy (represented by black triangles), is only found in Europe, apart from one other example in North America (12 instances). The pronoun retention strategy (represented by dark grey diamonds) is even rarer, and restricted to languages in West/Central Africa and Southeast Asia (5 instances).

Map 1 and the methodology associated with creating it demonstrate that by conducting a typological survey of a particular function or structure, it is possible to examine its distribution and find that what we might be used to from a Euro-centric perspective is actually very unusual. The relative pronoun strategy is not restricted in terms of genetic units, but is in fact an **areal trait** in Europe. The map also reveals that the only strategy clearly distributed across all continents is the gap strategy. We can say with conviction that the gap strategy is by far the most common linguistic strategy for relativising subjects and that it is not restricted to certain areal or genetic groups.

2.2 The accessibility hierarchy

The examples we have seen so far have all been of relativisation of subjects; they are all of the type in (7a), where the **subject** of the relative clause is coreferential with the referent of the head noun. The underlined gap in (7a) indicates the normal position of the subject of *likes.* In English we can relativise on a variety of grammatical functions, as the following examples show[5]:

[5] English also has a relative pronoun strategy equivalent for these examples.

(7)

a.	the woman that ___ likes Mary	SUBJECT
b.	the woman (that) Mary likes ___	DIRECT OBJECT
c.	the woman (that) the boy gave a rose to ___	INDIRECT OBJECT
d.	the woman (that) Mary spoke with ___	OBLIQUE
e.	the woman (that) Mary knows the family of ___	POSSESSOR
f.	the woman (that) Mary is taller than ___	COMPARATIVE

However, not all languages exhibit the same possibilities. Keenan and Comrie (1977, 1979) looked at relativisation possibilities in a wide range of languages (50 in total) and came up with what they call the Accessibility Hierarchy[6].

Figure 1. Accessibility Hierarchy (Keenan and Comrie 1977, 1979)

subject > object > indirect object > oblique > genitive > object of comparison

Initially, the hierarchy was thought to indicate a universal stipulating that if a language can relativise a grammatical function to the right of the hierarchy, then it will also be able to relativise everything to the left of that point. For instance, if a language can relativise on obliques then it will also be able to relativise on subject, object and indirect object. Similarly, if a language can relativise on one grammatical function only, it will be the subject. Clearly, it is only through comparing languages that such a hierarchy could be devised.

Sometimes patterns and subsequent claims about language made by typologists need modification in light of additional language data. The two hierarchies presented in Figure 2 (based on Comrie 1989: 147-8) are proposed to represent the relativisation strategies of Persian. Like English, Persian can relativise on grammatical functions at all points on the Accessibility Hierarchy, but it uses two different strategies to do so. Figure 2 illustrates that the gap strategy is used in Persian to relativise on subjects and direct objects whereas the relative pronoun strategy is used to relativise on everything

[6] Sells (page 214, section 2.2) discusses a similar hierarchy of grammatical functions from a theoretical perspective.

except the subject (including the option to relativise on the direct object using the pronoun strategy).

Figure 2. Relative clause strategies in Persian (Iranian, Indo-European; Iran)

Gap: subject > object > indirect object > object of preposition > possessor

Pronoun: subject > object > indirect object > object of preposition > possessor

Data of the kind underlying this hierarchy indicate the initial universal claims associated with the hierarchy were too strong. Instead, we have to say that (i) the subject must always be relativisable[7], and (ii) that any given strategy will cover a contiguous portion of the hierarchy. The Accessibility Hierarchy stands as an example of how, by looking at different types of structures for a single situation type, we can come up with universal claims about language.

3. The mutual relationship between language documentation and typology

One very obvious connection between typology and language documentation and description is that typologists rely on high quality grammatical descriptions to be able to carry out their work. For instance, the data sample of 166 languages used for the relative clause study discussed above were not all collected by Comrie and Kuteva firsthand (although some undoubtedly were). In the most part, they have relied on quality descriptions produced by other linguists.

In turn then, we may ask what people documenting languages want from typologists. Initially we might say that language documenters look to typology to inform them of variation evident across languages in order to guide them about the concepts and terminology that they should use as part of their corpus annotation and grammatical description. If this were not the case, every time a linguist describes a language, they would simply be starting afresh: fieldworkers would not be using the same terminology across descriptions, and analyses would not be informed by similarities to or differences from

[7] Sells (this volume, page 234, examples (48)-(49)) shows that only subjects can relativise in Western Austronesian languages like Toba Batak.

other languages. Typology aids research on languages that have not previously been documented by making field linguists aware of what is 'out there', what is 'possible' and what is 'probable'. In short, typology can make language documenters informed of possible variation before they start working on a language. Such awareness is particularly important in fieldwork since many phenomena that might seem exotic in comparison to one's native language may actually be typologically common. In an extreme case at the other end of the spectrum, an aberrant construction may ultimately prove to be something typologists have thought not to exist. Although it is not necessary to be a typologist to document and describe a language, becoming an expert on the typological characteristics of a given language family is advisable. If linguists know what patterns are common in language – in general or in the languages of a particular family – they will be quick to recognise unexpected deviations from the 'norm' in the language(s) they are examining.

What else can typology do for language documenters? There seems to be a common misconception among linguists that typologists just raid grammars and descriptions for analysis and do not give anything back to those who collect the data. This is not true. Many tools created by typologist as part of their research are of direct use to field linguists.

Perhaps the most tangible tools that typologist create for field linguists are questionnaires and stimulus kits. Questionnaires are very useful for linguistic research, but stimulus kits are particularly important for language documentation because they remove some of the language bias that questionnaires introduce. There are plenty of questionnaires available for phonology, morphology, syntax and semantics. Stimulus kits are less freely available but lots of information is available at the following link (see also Lüpke, this volume, for some examples and illustrations):

http://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaires.php

Another tool of use to fieldworkers that was developed by typologists is the set of recommendations called the Leipzig Glossing Rules (LGR):

http://www.eva.mpg.de/lingua/resources/glossing-rules.php

These are a set of principles for the encoding of interlinear morpheme-by-morpheme glosses. The LGR are a standardised set of glossing conventions that provide guidance to annotators when glossing examples. The guidelines include suggestions for category abbreviations in glosses, and examples of how different morphological operations and properties (e.g. affixation,

infixation, reduplication, cliticisation, cumulative expression, etc) can be distinguished effectively within inter-linear gloss lines.

The LGR contain suggestions for category abbreviations (e.g. present tense = PRS) because some categories (with different core meanings) found in glosses are often given identical labels across descriptions that could ultimately be confused. A case in point is the similarity between the terms *perfective* and *perfect,* which are often glossed as PERF in language descriptions without further clarification. The LGR distinguish these as *perfect* = PRF, and *perfective* = PFV. In fact, to avoid this problem, many linguists involved in language description have abandoned the label *perfect* for the term *anterior* in order to remove any confusion, however it persists in older grammars.

The benefit of a standardised set of terms to use in description is that is creates the opportunity to make documentation and description outcomes more accessible to fellow linguists. However, use of the standard abbreviations provided in the LGR does not say anything specific about the semantics or grammatical properties of a labelled category. It simply makes it easier to identify what terminology is being used.

The next typological tool of use to those creating a detailed description of a language is the Universals Archive:

http://typo.uni-konstanz.de/archive/intro/

Since the early 2000s, researchers at the University of Konstanz have been collecting published universals found in the typological literature, especially those of an implicational kind (e.g. *if a language has X then it will also have Y* or *if a language has X and Y it will have Z).*

This archive of universals, which includes hierarchies and semantic maps, can provide predictions to be tested in field data and thus is of interest to documenters and describers. This is particularly pertinent for language documentation work involving a large corpus of spontaneous speech and little elicitation. By looking through the universals archive it may be possible to pinpoint areas of research where targeted elicitation and analysis is necessary to create a richer description. With this in mind, one way to utilise the Universals Archive would be to look through the universals listed and devise testable hypotheses. For instance, when working on a language that has dominant VSO order and a universal within the archive refers to this characteristic, consider whether that universal stands true for the language under investigation. If it does not, it provides a reason to investigate how the language diverges from the regular pattern and why.

Finally, typologists also create cross-linguistic databases that provide an insight into which variables will be useful in an analysis of data in a language corpus. When browsing an online database, users are typically able to search through the data based on a number of different variables. For instance, in an agreement database, one might be interested to check if any language permits agreement in subordinate clauses with second-person dual subjects. In a sophisticated database it is possible to search on multiple parameters. By investigating what types of parameters are encoded in databases, it is possible for fieldworkers to become better informed about what sorts of variables will be important in their own documentation and descriptive work. There are many open-access typological databases available; some major ones can be accessed via the following links:

http://www.hum.uva.nl/TDS/

http://www.smg.surrey.ac.uk/

http://wals.info/

Although many of the outcomes of typological work can be beneficial to field-linguists, language documenters do not need to please typologists: by being aware of the benefits typology brings to corpus construction and related analyses, language documenters provide the best resources for typologists as a by product, not a goal.

4. Fine-grained variables for description and typology

Having outlined some of the things that typology can do for the language documenter, I should also point out what typology cannot do for you. There is increasing recognition in typology that linguistic categories are language-specific not universal (e.g. Croft 2001, Haspelmath 2007), and that the linguistic categories posited in a given description are language-specific descriptive categories (cf. Haspelmath 2008). For this reason, anguishing over finding an appropriate label for a category is in many ways moot. Essentially, this is because **assigning** a label to a category **does not describe** it: the variables which are important for cross-linguistic comparison are actually much more fine-grained than category labels conventionally indicate. This does not mean that linguists should abandon all existing terminology, or that every label should be non-descriptive (e.g. calling categories X, Y, and Z when you mean verbs, nouns and adjectives, cf. Garvin 1948). While it is still appropriate to use the terms *past tense*, *perfective* or *imperfective*, etc. as a guide to a general meaning, it is increasingly popular in typology to use uppercase labels for these (language specific) categories and to refer to them,

at least in typological work, with the language name. When talking about past tense in English in this way, it is referred to as the *English Past* not just *Past Tense* because it is a specific category relevant to English. Similarly, if you were to look further afield to Eleme (spoken in southeast Nigeria), the *Eleme Continuous* is a language specific category found in Eleme alone (although it may be similar to categories labelled *Continuous* in other languages). Variation of this kind indicates that descriptions need to be very thorough even with categories that might otherwise be taken for granted. In principle we are free to label a category with any language-specific term deemed appropriate, however there is an onus on the language documenter to increase the transparency of the descriptive content of such terms, and not to assume the existence of pre-established categories (e.g. from the Latin grammar tradition). Along with the augmented need for detail and clarity in language descriptions, the realisation that categories are language-specific calls for a new honesty in assessing the scientific credentials of the methodologies typologists use in comparing grammatical categories cross-linguistically. Fine-grained quantitative distributional analyses with description give a complex but potentially useful basis on which to compare languages.

Haspelmath (2007:125) proposes that:

> instead of fitting observed phenomena into the mould of currently popular categories, the linguist's job should be to describe the phenomena in as much detail as possible. Language describers have to create language-particular structural categories for their language, rather than being able to "take them off the shelf". This means that they have both more freedom and more work than is often thought.

To summarise, descriptive work that will be useful for the (contemporary) typologist will involve explicitly defined, fine-grained variables. In corpus-based work within a documentation project, this will also be quantified in some way. To illustrate the first point, we will look at case studies of word classes in Jaminjung, a language of northern Australia, and agreement morphology in Eleme, spoken in southeast Nigeria.

4.1 Jaminjung word classes

As linguists, we are used to the idea that 'nouns' and 'verbs' are major categories in language and many linguists would claim they are universal properties of language. Yet, if a language has a verb, what semantic domain will it cover? If we look at Jaminjung (Djaminjungan, Australia) we find two word classes broadly described as verbs (Schultze-Berndt 2000, 2003). Inflecting Verbs (IVs) and Uninflecting Verbs (UVs) constitute two major

word classes that are distinct from Nominals (N). IVs include items glossed as: 'go', 'come', 'eat', 'put', 'hit', 'poke', while UVs include items glossed as: 'go down', 'smash', 'walk', 'drink', 'light (a fire)'. With this in mind, we might wonder which of these two verb-like categories is the 'real' verb. In investigating parts-of-speech in a language like this (or indeed any language for that matter), a fine-grained distributional analysis is essential. Although IVs and UVs are conventionally glossed using the names for English Verbs, they do not have the same morphosyntactic properties as one another, indicating that they belong to different distributional classes: IVs can take tense-aspect-mood (TAM) and person marking, while UVs and Ns cannot. UVs also differ from IVs in that the former can be reduplicated to signal repetition, duration or intensity. IVs can be used in independent predication as in (8), while UVs require the presence of an IV, as seen in (9). For clarity, the glosses of inflecting verbs are underlined, while uninflecting verbs are not:

(8) Jaminjung (Schultze-Berndt 2000: 118)

gagawurli-wu	*yirr-ijga:::-ny,*	*manamba*
long.yam-DAT	1PL.EXCL-go-PST	upstream

'We went for long yam, upstream.'

(9) Jaminjung (Schultze-Berndt 2003: 150)

nganthan	*wij-wij*	*ngath-angga-m?*
what	RDP:scrape	2SG:3SG-get/handle-PRS

'What are you scraping?' (the addressee was scraping a carrot)

In (8) the IV *ijga* 'go' is the only 'verb' in the predicate. In contrast, (9) contains a reduplicated UV *wij* 'scrape' together with an IV *angga* 'get' or 'handle'. In Jaminjung one class of items which we would call verbs in English can occur in independent predicates, like 'go' in (8), but the other class are dependent - they have to occur with an IV, either in the same clause or as a clause dependent on another clause. Semantically, UVs and IVs are similar, but morphosyntactically UVs share a lot of characteristics with Jaminjung Nouns. For instance, UVs can take a subset of case markers such as the Dative, making them more noun-like. Example (10) shows a UV *wirrigaya* 'cook' marked with the dative case, while (11) illustrates a Noun *guyung* 'fire' marked with the same case:

(10) Jaminjung (Schultze-Berndt 2003: 154)

guyug=biyang	*nganji-bili=rrgu*
fire=now	2SG:3SG-POT:handle=1SG.OBL

[wujuwuju	*wirrigaja-wu]*
small	cook-DAT

'You should get fire(wood) for me now, for cooking the small (fish).'

(11) Jaminjung (Schultze-Berndt 2003: 157)

ga-jga-ny	*yina-wurla*	*guyug-gu::*
3SG-go-PST	DEM-DIR	fire-DAT

'She went over there for firewood…'

Under this analysis, 'fire' is a Noun because it can occur with the full range of case markers, whereas the UV 'cook' can only occur with certain case markers like the Dative. In both instances, the Dative has the same semantic interpretation of purposive.

Table 1. Morphosyntactic properties distinguishing IVs, UVs, and Ns in Jaminjung (Schultze-Berndt 2000, 2003)

	IVs	UVs	Ns
TAM/person marking	✓	×	×
Independent predication	✓	×	(✓)†
Determination	×	×	✓
Case marking	×	(✓)§	✓
Class size	closed*	open	open

Key: † in verbless equative or ascriptive clauses
§ with a subset of case markers in subordinating function
* around 30 members

Schultze-Berndt's analysis of parts-of-speech in Jaminjung is very explicit about the variables that are used to demonstrate to which class lexical items belong. Table 1 summarises the characteristics of IVs, UVs and Ns. Perhaps the most interesting and perplexing property of these classes, concerns their relative sizes. The IV class is a closed class and only has around 30 members while the UV class is an open class, so loan words (or more specifically,

borrowed verbs) are UVs. It is striking from a typological perspective that what we might assume to be an open class (of Verbs) is actually closed.

The use of fine-grained variables like these is paramount in the best descriptions of a language, and useful for typologists too: such descriptions provide a great deal of information about how the classes differ.

In the Australian languages that have similar systems to Jaminjung, a variety of terminology has been used for UVs, including Preverb, Coverb, Verbal Particle, Participle, Base and (Main) Verb (Schultze Bernt 2003: 146). In fact, Schultze-Berndt herself refers to the UV as Uninflected Verb, Preverb and Uninflected Particle throughout her various publications. Notice that the label itself does not tell us much about a word class; a fine-grained distributional analysis of the data does.

4.2 Eleme agreement morphology

The analysis of agreement morphology is another area of description where fine-grained variables are useful. The following paradigm of subject affixes in Eleme (Ogonoid, Niger-Congo; Nigeria) is representative of a wide range of variation in the indexation of grammatical roles in the language:

(12) Eleme (Bond 2010: 5)

(a) *ǹ-ʔerá*
1SG-stop
'I stopped.'

(b) *rɛ̃̀-ʔerá*
1PL-stop
'We stopped.'

(c) *ò-ʔerá*
2-stop
'You (SG) stopped.'

(d) *ò-ʔerá-i*
2-stop-**2PL**
'You (PL) stopped.'

(e) *è-ʔerá*
3-stop
'S/he stopped.'

(f) *è-ʔerá-ri*
3-stop-**3PL**
'They stopped.'

First-person singular (12a) and plural (12b) are indicated by the use of prefixes, *ǹ-* and *rɛ̃̀-* respectively. In both cases, a distinct prefix is used to indicate a specific person and number combination. In contrast, second-person singular (12c) and third-person singular (12e) are indicated through the use of a prefix that is specified for person, but not number. The plural counterparts to these in (12d) and (12f) share the same prefixes *ò-* (2) and *è-* (3) but there are

also suffixes that indicate second-person plural *-i* (12d) and third-person plural *-ri* (12f). Table 2 provides the various allomorphs of the affixes, which are determined by consonantal assimilation (1SG), vowel harmony (2, 3) and alternate forms determined by apparent free variation (1PL).

Table 2. Variables affecting the distribution of affixes in Eleme (Bond 2010)

default subject affix		permits overt controller	required by overt controller	requires pronominal in subject position
1SG	*m̀-/ǹ-/ŋ̀-/ŋ̀m̀-*	✓	×	×
1PL	*rɛ̃-/nɛ-*	✓	✓	×
2	*ò-/ɔ̀-*	×	×	×
3	*è-/ɛ̀-*	×/?	×	×
2PL	*-i*	✓	✓/?	✓
3PL	*-ri*	✓	✓	×

In describing this paradigm, I have separated the various subject affixes and looked at them in terms of three different variables, summarised in Table 2. The first variable concerns whether it is possible for an independent pronoun or noun phrase that expresses the same features as the affix to co-occur with it. For instance, can the first-person singular prefix co-occur with an independent pronoun that is first-person singular? Table 2 demonstrates that first-person singular and first-person plural can occur with an independent pronoun, as can the second-person plural and third-person plural suffixes. In contrast, the second-person and third-person prefixes cannot. Already, we can see that individual affixes in the paradigm do not behave in the same way with respect to this parameter. Although it is sometimes assumed that affixes in the same paradigm will have similar behaviour, they frequently do not because their actual use is based on various semantic and pragmatic/discourse properties (e.g. the distinction between addressee, speaker reference, non-participant reference).

The second variable approaches the same issue from a different angle. It considers whether the affix is *required* if an overt controller is present (i.e. if there is also a subject NP or independent pronoun). In this case the first-person singular prefix and the first-person plural prefix differ in their distribution. The first-person singular prefix can occur with an independent pronoun but it is not required to do so. In contrast, the plural form is always required if there is an independent pronoun. This type of distinction might be missed in an analysis that does not look at individual parameters

systematically. Table 2 demonstrates that the rest of the affixes behave in the same way for this variable as they did for the first parameter.

The final variable concerns whether the affix requires a pronominal to be in subject position or not. This is most pertinent for the suffixes. In (12d) and (12f) the suffixes *-i* and *-ri* co-occur with their respective prefixes. However, constructions where there is no prefix, but where an independent pronoun is in subject position instead are also possible. This suggests that the prefix is occupying the argument slot that an independent pronoun otherwise occupies. The second-person plural suffix requires some element in that slot; either a prefix or an independent pronoun. In contrast, the third-person plural suffix does not require anything in cases where the referent of the suffix is retrievable from the discourse structure and context.

These variables demonstrate the distribution of the various prefixes and suffixes in a paradigm which appears to be fairly unproblematic at face value. By looking at individual affixes across fine-grained variables a much more sophisticated and revealing analysis is permitted. This is exactly the kind of analysis that a contemporary typologist might find useful, but is also the kind of analysis that characterises good quality descriptions. Bickel (2007: 247) claims:

> such variables allow capturing rather than ignoring diversity, and they stand a greater chance to be codable in replicable ways across many languages. Fine-grained variables form just the right input for research on how structures distribute in the world, and, at the same time, they provide just the right tools for analyzing individual structures beyond futile naming exercises.

Effectively, it is a waste of time to argue whether a particular linguistic feature belongs to a particular category or not, rather, it is better to demonstrate its characteristics with a fine-grained distributional analysis. The best descriptions, then, look at multiple fine-grained variables in establishing category or class membership and explicitly identify which variables have been used to establish that class.

5. Conclusion

Contemporary typology is concerned with the diversity found in the world's languages and aims to answer the question 'what is why where?' rather than 'what is possible?' Typology has its own empirically-based methodologies. The standard typological method involves selecting a semantic-pragmatic domain, identifying the strategies used by languages to express that domain and searching for generalities across the domain, such as dependencies or other functions associated with those structures. Typologists do not just take: they are willing to give as well. The resources that result from cross-linguistic

comparison are useful in language documentation in terms of annotation tools, questionnaires and stimulus kits, hypotheses to test, and variables to investigate.

Cross-linguistic research has raised doubt about the idea that there are a small number of innate or universal categories. Rather, it suggests that there are many language-specific categories, some of which do not have parallels across languages. As a consequence, the best descriptive and documentation research will (i) use quantification from the corpus to support the distribution of forms and (ii) use fine-grained variables to account for variation.

References

Bickel, Balthasar. 2007. Typology in the 21st Century: Major current developments. *Linguistic Typology* 11, 239-251.

Bond, Oliver. 2010. Intra-paradigmatic variation in Eleme verbal agreement. *Studies in Language* 34, 1-35.

Comrie, Bernard. 1998. Rethinking the typology of relative clauses. *Language Design* 1, 59-86

Comrie, Bernard & Kuteva, Tania. 2008. Relativization on Subjects. In Martin Haspelmath, Matthew S Dryer, David Gil, & Bernard Comrie, (eds.). *The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 122.* (Available online at http://wals.info/feature/122. Accessed 2009-07-01).

Croft, William. 2001. *Radical Construction Grammar*. Oxford: Oxford University Press.

Croft, William. 2003. *Typology and universals* (second edition). Cambridge: Cambridge University Press.

Dryer, Matthew S. 2008a. Order of Subject, Object and Verb. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) *The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 81.* (Available online at http://wals.info/feature/81. Accessed on 2010-01-24).

Dryer, Matthew S. 2008b. Order of Adposition and Noun phrase. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) *The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 85.* (Available online at http://wals.info/feature/85. Accessed on 2010-01-24).

Garvin, Paul L. 1948. Kutenai III: morpheme distributions (prefix, theme, suffix). *International Journal of American Linguistics* 14, 171-178.

Gordon, Lynn. 1986. *Maricopa morphology and syntax.* Berkeley: University of California Press.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of language: report of a conference held at Dobbs Ferry, New York*, April 13-15, 1961. Cambridge, MA: MIT Press.

Greenberg, Joseph H. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of grammar* (second edition), 73-133. Cambridge, MA: MIT Press.

Haspelmath, Martin. 2007. Pre-established categories don't exist: consequences for language description and typology. *Linguistic Typology* 11, 119-132

Haspelmath, Martin. 2008. Comparative concepts and descriptive categories in cross-linguistic studies. Ms. MPI-EVA, Leipzig. (Available online at http://email.eva.mpg.de/~haspelmt/CompConcepts.pdf. Accessed on 2009-07-01).

Keenan, Edward L. and Comrie, Bernard. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8, 63-99.

Keenan Edward L. and Comrie Bernard. 1979. Data on the Noun Phrase Accessibility Hierarchy. *Language*, 55: 333-51.

Lüpke, Friederike. 2010. Research methods in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7,* 55-104. London: SOAS.

Schaub, Willi. 1985. *Babungo*. London: Croom Helm.

Schultze-Berndt, Eva. 2000. *Simple and complex verbs in Jaminjung. A study of event categorisation in an Australian language*. MPI Series in Psycholinguistics, 14. Nijmegen: University of Nijmegen.

Schultze-Berndt, Eva. 2003. Preverbs as an open word class in Northern Australian languages: synchronic and diachronic correlates. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2003*, 145-177. Dordrecht: Kluwer.

Sells, Peter. 2010. Language documentation and linguistic theory. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7,* 209-237. London: SOAS.

Comrie, Bernard. 1989. *Language universals and linguistic typology* (second edition). Oxford: Blackwell.

Croft, William. 2003. *Typology and universals* (second edition). Cambridge: Cambridge University Press.

Epps, Patience. In press. Linguistic Typology and Language Documentation. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology.* Oxford: Oxford University Press.

Shopen, Timothy (ed.). 2007. *Language typology and syntactic description* (Vols 1-3). Cambridge: Cambridge University Press.

Song, Jae Jung. *Linguistic typology: morphology and syntax*. Harlow: Longman.

Whaley, Lindsay J. Introduction to typology: the unity and diversity of language. London: Sage.

Discussion questions

1. Visit the WALS database online: http://wals.info. In the Complex Sentences section Feature section, open up the chapters on *Relativizing on Subjects* (Chapter 122) and *Relativizing on Obliques* (Chapter 123) by Bernard Comrie and Tania Kuteva. Compare the map for subjects with the map for obliques.

   a. What discrepancies do you see between the two maps?
   b. How are the different strategies distributed?
   c. Can you identify any languages that have different strategies for relativisation on subjects and obliques? If so, are any tendencies identifiable in the sample?
   d. What methodological problems exist in comparing these two maps directly?

2. In what ways can the use of a corpus formed as part of language documentation be employed to good effect in a typological study? Consider the benefits and difficulties of using corpora over printed grammars in cross-linguistic research.

3. *'Language documenters should ensure that their corpora are designed with typologists in mind.'* Do you agree or disagree with this statement? In what ways can the relationship between typology and language documentation and description be nurtured?