Languages of Europe: an introduction to Linguistic Typology

A course valid for:

- Schlüsselkompetenzen
- Sprachen der Welt / Sprachen im Kontrast
- BA Language Science

Instructor: Luigi Talamo

Department of Language Science and Technology

UNIVERSITÄT DES SAARLANDES

Credits

The present introduction is based on

William Croft, Typology and Universals (2nd edition), Cambridge: Cambridge University Press, 2003.

We give Bill some credits for his excellent textbook (although a little bit difficult to follow)

What is typology

Two definitions:

- **typological classification**: a classification (a taxonomy) of structural types across languages;
- **typological generalization**: a generalization of linguistic patterns (a language universal) across languages.

Both definitions imply that the linguist should investigate more than one language**: cross-linguistic comparison**.

The first definition applies to 19th and 20th century Morphology, where scholars researched across language grammars for similar linguistic types, in the same way as botanists searched across plants for similar shape of leaves, flowers and seeds. The basic typology opposed isolate language i.e. languages that do not use morphology at all (no cases, no verbal inflections) to fusive/agglutinative languages i.e. languages using morphology (case, verbal inflection): Mandarin Chinese vs. German.

The second definition applies to a more modern version of linguistic typology i.e., the one started by Joseph Greenberg in the 1960s and takes the form of a implication correlating two or more linguistic features.

An implication is something like 'If X is true , then Y': the validity of X (its being true) depends on the validity of Y. This predicts the following cases:

1. X is true, Y is true;
2. X is false, Y is true;
3. X is false, Y is false

ruling out a fourth case:

- X is true, Y is false

Now think of X and Y as linguistic features and of cases as linguistic types.

Linguistic example: if a language has adjectives, then it also has verbs. The implication predicts the following linguistic types:

1. Language has adjectives and has verbs;
2. Language doesn't have adjectives, but it has verbs;
3. Language doesn't have adjectives and doesn't have verbs;

ruling out a fourth type:

- Language has adjectives, but it doesn't have verbs.

What is typology

A third definition:

- **an approach** or **a theoretical framework** to the study of language. Typology is then a methodology to analyse languages.

In this sense, typology is complemented by **functional linguistics:** a theory of language that seek the explanation of linguistic structure not in the language itself (as in **formal linguistics**) but in terms of linguistic function.

Functional-typological explanation

Since the 1960s, linguists have been somewhat divided into 'functionalist' and 'formalist'. A particular type of 'formalism' emerged since the late 1950s and was headed by Noam Chomsky: this school of linguistics is also called 'Generative Grammar' and its practitioners 'Chomskyans' or 'Generativists'. On the functionalist side, the happy marriage between functionalism and linguistic typology led to the functional-typological explanation of linguistic facts.

What is typology

The three definitions correspond to three different steps in the **scientific study of language**:

1. **Observation** and **classification**;
2. **Generalization** over previous observation;
3. **Explanation** over previous generalization.

Scientific study of language: linguistics. Each step is build on the previous one.

What is typology

Explanatory models in Linguistic Typology include (but are not limited to):

- **Iconicity**: language structure resembles language meaning;
- **Economy**: shorter language structures are preferred to longer ones;
- **Processing**: how the human brain processes the language structure;
- **Diachrony**: today's language structure is explained by yesterday's language structure.

Iconicity -> reduplication is a strategy used in many languages to convey an augmented meaning. Many times a word = many times its meaning. For instance, 'her house it's very very very big'
Economy -> think of greetings in German or in English. We can say just 'Morgen or Abend' instead of 'Guten Morgen or Guten Abend'.
Processing -> the linguistic discipline studying how the human mind processes (=elaborates) language is called 'Psycholinguistics'. Using psychological experiments, psycholinguists verify linguistic hypothesis. For instance: do very long German words are elaborated in different ways by French and German speakers, maybe decomposing the word or with different speed rates?
Diachrony -> in Linguistics we oppose diachrony i.e., a study of a language through its history, i.e. through its different temporal phases (the study of English from Modern era to the Contemporary Era) to synchrony i.e., a study of a specific temporal phase of language (the study of English during Queen Victoria). Sometimes, linguistic structure doesn't have a contemporary explanation but is explained by structure that appeared in older temporal stages. For instance, the future in Romance languages such as French *chanterai*, Spanish *cantaré* derives from a very early temporal stage in which the future was built using the infinitive plus the verb 'to have' = French *chantair ai, Spanish *cantar eo < Latin cantāre habeō

Typology, Universals and Generative Grammar

Typology aims to discover **Language Universals: generalizations that apply to each and every language.**

So does **Generative Grammar**: what is the difference? In the **explanation**:

- **Generative Grammar** will seek an internal explanation, stating that Universal are **genetically innate** in the human brain.
- **Linguistic Typology** will search for explanation in extra-linguistic facts, such as "**general cognitive, social-interactional, processing, perceptual** or **other abilities**." (Croft 2002:5)

As we have seen before, Greenberghian Typology and Chomskyan Generativism appeared quite at the same time, in the late 1950s

A generative linguist will explain languages by using the language structure itself (most notably, the syntax), which is thought to be a sort of programmable computer hardware encoded in the brain. This computer is programmed through external linguistic inputs, which are said to be very few since the new-born baby is only exposed to family/caregiver's language (poverty of stimulus).

For instance, in the generativist theory the class of adjectives needs the class of verbs since adjectives share a similar structure to verbs.

A functional-typological linguist will seek explanation outside the language i.e., in the human need to communicate something, for instance posing that languages having adjectives have also verbs since adjectives and verbs correspond to similar communicative concepts.

Typology, Universals and Generative Grammar

What's the difference between Typology and Generative Grammar? In the **analysis** of data.

- Generative grammar adopts a **rationalist** approach: the analysis proceeds with a top-down methodology (deductive). Hypotheses are first elaborated and then tested on linguistic data.
- Typology adopts an **empiricist** approach, working bottom-up (inductive). Hypotheses are elaborated on the basis of analysed linguistic data.

Deductive: from top of innate ideas to the down of real linguistic facts.
Inductive: from bottom of real facts to the up of verified hypotheses.

Typology, Universals and Generative Grammar

Finally, the difference lays in the **methodology**.

- In Generative Grammar, **few languages**, even just the linguist's native language, are enough to verify Language Universals.
- In Typology, the **more number of languages**, the better. In order to discover Language Universals, a cross-linguistic approach should be taken.

Since Language Universals are encoded in each and every human brain, the generative linguist can formulate and verify hypotheses just using her/his native language. On the other side, a language typologist should elaborate her/his theories analysing the biggest number of languages, virtually all languages. Since it's impossible (there are nowadays 6000-7000 languages and a lot of them are undocumented), the typologist should rely on sample of languages i.e., carefully-selected list of languages.

Cross-linguistic comparison

The methodological validity of cross-linguistic comparison can be advocated at two different levels:

- **Cross-linguistic level**: a typological generalization could not be formulated without comparing more than one language;
- **Language-specific (one language at a time approach) level**: the description of a particular language benefits from the descriptions of other languages.

Croft discusses a contrastive analysis of the articles in English and French. Even in languages so closely related, there are differences in the encoding of the same meaning, which are listed at the beginning of §1.3 and addressed as 'types of uses of the article'. The comparison opposes definite articles (English: the and French: le/la) vs. indefinite articles (English: a/an and French: un/une). The distribution across the examples is the same only in three cases out of eleven. This particular type of linguistic problem is far from being resolved (but join the last class of this course if you are interested!) i.e., no satisfactory generalization of the distribution of articles has been provided yet, but shows the following:

- generalization made on the basis of just one language never holds (even in comparison with very close languages);
- The knowledge of how other languages work helps the linguist to better describe a specific language (even her/his own language!)

Cross-linguistic comparison

Within a single language, one usually identifies a **grammatical category** by using the **distributional method**:

- The **occurrence** or **distribution** of a **grammatical category** is observed in different constructions;
- a unique **grammatical category** exists if it behaves in the same way in all of these different constructions.

The standard syntactic argumentation uses the distributional method to discover grammatical categories in a language. For instance, in order to discover the grammatical category of subject in English we take different constructions of what we suspect it may be a subject: nominative form of the pronoun, agreement, unexpressed argument of the infinitive, unexpressed argument in the imperative, unexpressed shared argument in a conjoined sentence. In all of these constructions, the same grammatical category appears as the preverbal noun phrase (NP), giving evidence for the existence of a subject in English.

Cross-linguistic comparison

Cross-linguistically, the **distributional method** can not be applied. Rather, in order to make typological generalisations, we have to invoke a **cross-linguistic comparison**:

- Does our language-specific generalisation apply to other languages?

Maybe, there is even something more and we discover connections between linguistic structures which are **explicit** in some languages but **hidden** or **apparently mysterious** in other languages.

- We can also discover that connections between certain linguistic structures are in fact **rarely observed outside our European languages**.

We should ask if what we have observed and generalised for a single language can be extended to other languages. As a bonus, we can discover connections between linguistic structures which weren't explicit in a single language, or they don't have an apparent reason.

As for hidden things in Euro-languages, in English (and in other Euro-languages) the connection between the protasis (the antecedent) of a conditional sentence (the if clause) and the topic (the as for preposition) is not manifested: these two constructions are formally different:

- If you eat that, you will get sick.
- As for Randy, he's staying here.

However, if we look at other languages, we find that in distant and unrelated languages the if-clause and topic are in fact encoded with the same formal structure. For instance, in Turkish and in Tagalog (the national language of the Philippines) the protasis and the topic are marked with the same formal structure.

As for mysterious linguistic structures observed at our latitudes, in English (and in other Euro-languages) we see that the preposition *with* is used in constructions with different meaning: comitative, instrument and manner. According to the

distributional methods, we make a generalisation for English: does this generalisation is observed in other languages as well? Yes, it does: very distant languages such as Hausa (a language spoken in Nigeria) and Classic Mongolian (an ancient language spoken in China) show the same distribution. What is the reason behind such a wide distribution? A metaphorical (causal) relation between the three participants: an animate companion (a comitative) is metaphorically similar to an inanimate companion (an instrument), and both they are similar to a way of doing something.

Finally, and we come to the central topic of our course, cross-linguistic comparison (and Typology more in general) helps us to see our familiar languages with a different eyes. With respect to things such as indefinite pronouns in predicate (copular) nominals (His brother has become a soldier, but cfr. Romance languages) or obligatory unstressed pronouns in the subject position (He waters plants and not *waters plants, but again cfr. Romance languages), European languages are very exotic when compared to the languages of the world.

The problem of cross-linguistic comparability

We need to identify the **same linguistic phenomenon** across languages; unfortunately, structural properties (**the form**) of linguistic phenomena are not sufficient when comparing languages:

- Languages heavily vary with respect to **phonetics/phonology**, **morphology** and **syntax**.

How do we find a **cross-linguistic valid definition** for our linguistic phenomenon?

- On **semantic** (and **pragmatic**) basis **(the meaning):** "I fully realize that in identifying such phenomena in languages of differing structure, one is basically employing semantic criteria." (Greenberg 1966:74)

Before starting looking at grammars, asking native speakers or querying corpora, we need a valid cross-linguistic definition for our phenomenon i.e., something that apply to all languages. Since the same phenomenon is encoded with a very wide range of structural properties across languages, we cannot formulate valid definitions on the basis of morphological, syntactic or phonological properties. For instance, the relative clause appears cross-linguistically encoded by several strategies, ranging from syntax to morphology. Even in European languages we have both

- Syntactic strategies: German 'Der Mann, der in seinem Büro, arbeitet';
- Morphological constructions: German 'Der in seinem Büro arbeitende Mann '.

We then need to seek criteria outside the structural properties of languages: 'semantic' criteria, which, in a broad sense, include pragmatic criteria such as discourse and conversational facts.

For instance, in the seminal paper by Keenan and Comrie on the accessibility of the NP, the following definition is proposed for relative clause: "a relative clause is a syntactic object specifying a set of objects in two steps: a larger set is specified, called the domain of relativization, and then restricted to some subset of which a certain

sentence, the restricting sentence, is true". In the example above from German we identify the domain of relativization as "man" and the we restrict the domain to the subset of "working in his office".

The problem of cross-linguistic comparability

Typological research strategy:

- Identify a particular **semantic** or **pragmatic** structure (the **function/meaning** of a linguistic phenomenon);
- Examine the **morpho-syntactic strategies (**the **form** of a linguistic phenomenon **) encoding** that structure;
- Search for **connections** between these **strategies/constructions** and other **linguistic factors,** both **internal (grammatical)** and **external (functional).**

The typological research strategy can be defined as circular, since the identification of the semantic/pragmatic structure of a linguistic phenomenon starts from the observation of formal constructions across languages and such identification is then refined by searching for more formal constructions and for connections between these different constructions and other grammatical or functional categories.
The difficulty of finding a suitable definition for cross-linguistic research shouldn't be overstated: many concepts are quite easy to identify without too many controversies, such as notions like tense, gender, number, aspect. Problems arise when we try to provide definitions for major grammatical categories such as parts of speech, syntactic roles, head/modifiers or sentence parts.

The problem of cross-linguistic comparability

We find **semantic/pragmatic notions** at **two different levels** of the language architecture:

- At the **linguistic phenomenon level (the whole linguistic sign)**: for instance, the subject;
- At the **strategy level encoding the linguistic phenomenon (the formal/structural part of the linguistic sign)**: for instance, case-marking/adpositions, agreement markers encoding the subject.

Morpho-syntactic strategies have their own **semantic notion**:

- Case-marking and adpositions: bounded (affixes) or free particles denoting a semantic relation between the noun phrase and the verb;
- Agreement markers: bounded (affixes) or free particles denoting the argument itself.

In the complex architecture of language the semantic/pragmatic structure is present both at the linguistic phenomenon level and in the strategy encoding the phenomenon (double-articulation of language, see double_articulation.pdf):

- Semantic/pragmatic notion of subject;
- Semantic/pragmatic notion of case-marking/adpositions or agreement markers encoding the notion of subject.

Morpho-syntactic strategies encoding subject across languages are as follows: case-marking / adpositions, agreement, word-order, or a combination of both of these.

Let's focus first on the morpho-syntactic strategies, which we will analyse in more detail in the next class:

- Case-marking or adpositions: they can be found attached to noun (affixes: Russian), as independent particles after or before the noun (adpositions: Romanian) or attached to the verb they refer to (Mokilese: Micronesia);
- Agreement markers (indexation): attached to verb (affixes: Hungarian), as independent particles after or before the verb (adpositions: Woleaian:

Micronesia), attached to other constituents (affixes: Ute, an indigenous language of North America) or to any constituents (affixes: Bartangi, a language spoken in Tajikistan).

By analysing the examples, let's try to provide a semantic notion for these strategies.

The problem of cross-linguistic comparability

What's a subject, anyway?

- Standard Average European (Eurocentric) perspective: "semantic relation between the event denoted by the verb and the participant denoted by the English subject";
- Cross-linguistic perspective: something "involving **external function**, including some notion of **agent of an action** and **topic of the sentence**".

We have a cross-linguistic valid definition for morpho-syntactic strategies encoding subject. Now let's focus on the notion of subject: if we try to apply these morpho-syntactic strategies across languages we may find that what we found doesn't correspond to the Eurocentric definition of subject.

For instance, in Chechen-Ingush, a language spoken in the Chechen Republic, we find that the agreement marker corresponds to our definition of subject only in the first example.

As with morpho-syntactic strategies, the solution is to give a notion based on external function and then find examples accordingly. For instance, good examples of sentences containing subject are those highlighting the animacy of the agent and her commitment to the action denoted by the verb. For instance: She broke the pencil and not She listens to the music or, worst, She smells perfume.

The problem of cross-linguistic comparability

- **External definitions**: cross-linguistic definitions based only on **external basis**;
- **Derived structural definitions**: cross-linguistic definitions made by **putting together structural categories** (which we have already defined on **external basis**).

Which is the best choice?

- It depends on the purposes of our typological investigations.

By using structural categories as building bricks we can develop derived structural definitions. For instance, a cross-linguistic definition of passive may be developed by putting together subject, verb, object and the active construction, assuming that these notions are described on external basis:

Passive construction: the subject of the passive verb is the object of the counterpart active verb. Mary eats the apple vs. The apple is eaten by Mary

There is no best choice, only cases in which one of the two definitions suits better the data or the purposes of our investigation. It's useful to compare the two definitions on the same object of investigation. Let's take the subjunctive as found in the German Konjunktiv:

Wenn ich viel Geld hätte, würde ich eine Weltreise machen

- External definition: a subjunctive clause denotes a situation that doesn't take place in the reality, but it's only presupposed or imagined;
- Derived structural definition: "a clause which expresses the subject and the object of the clause in the same way as an ordinary declarative main clause does, but

whose verb inflections differ from those of the verb in an ordinary declarative main clause" (p. 18).

If our research focuses more on the nature of this particular type of modality i.e., what's the nature of the subjunctive mood we will perhaps choose a purely external definition, while if we decide to see modality in the context of complex sentences (such as the example above) maybe a derived structural definition works better.

Language sampling for cross-linguistic research

There are about 6000 to 7000 languages spoken in the world today: since many of them are poorly documented or not documented at all, we need to build a **sample** for our research. The sample should be as diverse as possible, in order to

- cover all the possible realisations of the object of our research;
- avoid 'false typological friends', i.e. grammatical features that seem connected but in fact they are not.

The need for a good sampling is explained by Croft through the following objects of research:

- the passive: by looking at European languages only, we may think that the passive "involve the presence of an auxiliary and/or a preposition governing the agent phrase". Examples from Lummi (a Salish language once spoken in North America) and Bambara (a Niger-Congo language spoken in Mali) show that there may be other strategies to express the passive;
- If a language is pro-drop, then it will have agreement (and the other way round): again, by looking at our Euro-languages only, we'll see that in languages in which the independent subject pronoun may not be expressed (Spanish, Italian) have rich verbal inflections (indexical markers), while in languages with compulsory subject pronouns (German, English) there's little verbal inflections. In fact, this doesn't hold cross-linguistically: two important languages spoken in Asia (Mandarin Chinese and Japanese) do "not have obligatory independent subject pronouns and also do not have indexation"

Language sampling for cross-linguistic research

Two types of sample:

- **Variety sample**: "selects a subset that is intended to maximize the likelihood of capturing all the linguistic diversity for the phenomenon under study";

- **Probability sample**: "selects a sample from the set of languages whose probability of being chosen over another sample is known in advance."

A variety sample is aimed at capturing all the linguistic diversity, thus including as many diverse languages as possible. It is better suited for general questions, as it may not cover in every details the object of our research. With a probability sample we achieve results that are more precise, but we have to know in advance and in more details which linguistic features we are searching for, in order to select a set of languages over another: the probability sample is better suited for specific questions.

Language sampling for cross-linguistic research

Variety sample -> Genetic difference

- **Across families**: "the greater time depth from divergence, the greater likelihood of diversity". We choose languages from different families.
- **Within the same family**: we choose languages from branches that are more distant as possible. The greater the number of branches and sub-branches, the greater likelihood of diversity.

The general idea behind sampling languages that belong to different families is that languages evolve through the time, diverging from a common ancestor. The highest classification in language taxonomy is family: since many languages are undocumented, we must look into the same language family as well, in order to reach a reasonable size for our sample. Since the 19th century, a language family is represented with a tree-like form, with branches representing language groups, which in turn may have sub-groups, and so on.
genetic-trees_Dunn2011.pdf is a modern representation of four genetic trees for four different families: it is obtained by automatically computing a lot of lexical data and it doesn't necessarily cover all the existing languages in a family. As for the computation of diversity, for instance we can see that the Indo-European family starts with two major branches: Hittite vs. Rest of IE languages. In turn, Rest of IE languages is divided into two branches: Tocharian vs. Rest of the rest of IE languages, and so on.

Language sampling for cross-linguistic research

Variety sample -> Problems

- **Representation:** the representation of a language family (or even of a language group) is a matter of debate (and of research) in itself.
- **Research:** some language families (IE, Austronesian) are very well established, while others are not (Australian?)
- **Time**: Language families vary with respect to the time-depth of their branches: the branch of one language family may be much older than the branch of another.
- **Space**: Languages influence each other, so it's better to sample geographically distant languages.
- **Granularity:** a variety sample is not suitable for fine-grained analysis, as it only captures the extremes of a linguistic phenomenon.

For instance, there are at least two major proposals for the IE family tree: Indo-Hittite with two initial branches (as in genetic-trees_Dunn2011.pdf) and Indo-European with ten initial branches.
There is a general consensus on some language families such as Indo-European, Austronesian and Afro-Asiatic, while others language families such as Australian are still debated.
The time problem can be solved by "calibrating the branch by time depth", as it is graphically rendered in genetic-tres_Dunn2011.pdf, in which the branch length corresponds to the time-depth. Across language families or groups, we will sample languages with similar branch length.
The geographical problem is probably the best known problem and the easiest to visualise. English and French belong to different groups, Germanic and Romance, but the English lexicon was heavily influenced by French due to historical (and geographical) reasons. In turn, French was heavily influenced by another Germanic language, German, due to the very close contact between the two languages. So, English, French and German are not the best candidates for representing diversity in a sample!
Finally, a variety sample is defected by default, in the sense that it is designed to capture the broadest range of linguistic diversity and doesn't capture the

intermediate types of a linguistic phenomenon, but only its extremes. Intermediate types are found in genetically-related or geographically-closed languages, which are exactly the languages we are trying to avoid!
For instance, Salishan languages are a small family of languages spoken in North America. In building a variety sample, we will probably choose just one language from this family. However, Salishan languages display a great diversity in the encoding of passive:

- Lummi: verbal inflection plus adposition marking the agent;
- Upriver Halkomelem: no dedicated verbal inflection, agreement with passive subject (indexical markers), no adposition marking the agent;
- Bella Coola: no dedicated verbal inflection adposition marking the agent.

Language sampling for cross-linguistic research

Probability sample -> independent occurrences of the combinations of traits.

Problems:

- **Genetic difference**: languages may share a combinations of traits due to common ancestor
- **Areal contact**: languages borrow traits from geographically-close languages

For a probability sample, we have to choose languages showing independent occurrences of the phenomenon we are investigating. We again encounter the issues of genetic difference and areal contact.

As for genetic difference, we find for instance in Russian and Czech (two Slavic languages) the two constructions:

- Preposition na + accusative: motion;
- Preposition na + locative: location.

The combinations between preposition and case is identical in both languages, so we might to propose the following implication: "if a language uses both adpositions and case affixes for indicating grammatical relations, then the case affixes will be used to distinguish motion from location. " (p. 23) However, the pattern is inherited by the common ancestor of Russian and Czech, Common Slavic, as attested in Old Church Slavonic.

As for areal contact, let's take for instance Romanian, Albanian and Bulgarian, three IE languages belonging to three different branches, which share a certain amounts of morpho-syntactic constructions such as: postposed definite article in the form of

suffix and lack/avoidance of infinitive forms. These are not tracts inherited by the common PIE language nor independent traits, but areal features due to the long areal contact between these languages.

Language sampling for cross-linguistic research

Probability sample -> "wide areal and genetic distribution is neither a necessary nor a sufficient condition" for independent occurrences of traits. (p. 24)

- **Cognate languages** may have **independently** developed the same occurrence of traits;
- **Languages** that are **genetically** and **geographically distant** may still retain traits from the common ancestor.

This is particularly true for **stable linguistic phenomena**.

Not Necessary: Spanish and Russian are cognate languages that are quite close in the IE genetic tree and not too far from a spatial perspective. In both languages we find that the reflexive marker is also used as the marker of middle voice of some verbs: (Spanish: enclitic particle **se,** Russian suffix **–sya**)

- Reflexive: Sp. *Lucia **se** lava* and Ru. *Lyusiya moyet**sya** '*Lucia wash herself';
- Middle voice: *Sp. La puerta **se** abre* Ru. *Dver' otkryvayet**sya***. 'The door opens'

Since the two strategies traces back to the same IE reconstructed morphemes, we might be tempted to suppose that Sp. and Ru. either retain a common IE pattern or have borrowed the pattern due to a contact. However, this is not the case: in the two languages, the pattern was developed independently and at different times.

Not Sufficient: Fula and Kinyarwanda are two Niger-Congo languages, however very far from each other both in genetic and spatial perspective. Both languages have a SVO order, which is probably a retention of the order reconstructed for Proto Niger-Congo, the ancestor language.

The identity between the marker of middle voice and the reflexive marker is an

example of an unstable phenomenon, which is less likely to be inherited by a common ancestor. Word orders are examples of more stable phenomena, which can be transmitted from the mother languages to daughters.

Language sampling for cross-linguistic research

Probability sample -> **proportion** of languages.

The proportion may reflect:

- the actual **number** of the world languages;
- the **areal** and **genetic** distribution: Dryer's pooling technique.

The problem is that the **probability sample** assumes a **stationary distribution** of the world languages, which in turn

"implies that enough time has passed for a language to have possibly passed through all the relevant types" (p. 28).

A possible solution is represented "by sampling **language changes**, not languages themselves" (Diachrony Typology).

Along with genetic and areal issues, a third problem is represented by the proportion of languages in the sample. How many languages shall we take from different families? And within the same family, how many languages from different groups?
A first solution may consist in building a sample reflecting the actual distribution of the world languages; for instance, if we have a sample of 20 languages and IE languages represent the 15% of the world languages, we should take into account the 15% of 20 i.e., 3 languages.
A second, more elegant solution is to divide the world languages into continent-sized areas, as proposed by Matthew Dryer: if a geographic area contains languages with the same occurrence of traits, say, all languages are SVO then we'll treat this area as a single datapoint, pooling together all the languages in a genus. If the area is divided into languages with SVO and languages with SOV, then we will have two datapoints for this area, and so on. Dryer's proposal essentially reflects the distribution of world languages at about 1000BC, where today's languages were represented by few proto-family or proto-group languages, for instance, Romance languages by Latin, Germanic languages by proto-Germanic and Austronesian languages by proto-Austronesian. However, a more stable linguistic situation was probably attested at about 4000BC, where only proto-family and stock (a level superior to the family) existed.
If we turn back to the current linguistic situation, we are still far from having reached

a stationary distribution of the world languages, which corresponds to having experienced all the possible grammatical traits, for instance, the two relative positions of Adjective and Noun.
It has been suggested that since the actual state of languages don't cover all the possible realisations of a phenomenon, we should take a look to previous stages of the languages.

Data sources

Typological research relies on the following **data sources**:

- **native speaker elicitation**;
- **texts**;
- **descriptive grammars**.

According to Perkins 1989, the number of languages is a sample is in the order of hundred, with a minimum of forty-fifty languages. Of course, no one can reach even a fair knowledge of all the languages in her/his sample: linguists should rely on data sources in order to conduct typological research.

Data sources

Native speaker elicitation:

- **first-hand** data; ✅
- **impractical** with a **large sample;** ❌
- data is filtered by the **perception** of the consultant. ❌

A refined method is represented by the **questionnaire**:

- can cover a **large sample**; ✅
- **difficult** to **design**. ❌

Gathering linguistic data from native speakers allows the linguist to analyse first-hand data. However, most of the time it is impractical to interview dozen and dozen of native speakers; moreover, first-hand data may seem of higher quality, but unfortunately elicited data is often biased by the perception of the native speaker towards her/his language and even by “desire to give the interrogator an agreeable answer”.

We can gather linguistic data faster and from a much higher number of languages using the method of the questionnaire, which consists in a series of questions on the investigated phenomenon. However, reliable questionnaires are difficult to design, can be filled out only by language experts and, even if mitigated, have the same problems of biased data of language elicitation.

Data sources

Texts:

- **unfiltered** data of actual language use; ✅
- **quantitative** data; ✅
- data can be biased by the **genre**; ❌
- data can be filtered by **morpheme-by-morpheme gloss** and **free translation;** ❌
- **rare phenomena** can be scarcely attested or not attested at all. ❌

Data sources

Descriptive grammars:

- (relatively) comprehensive description of the grammatical system: ✅

"you gotta muck around in grammars; you can't just pick a fact out from a grammar" (Joseph Greenberg)

If the **grammar author** is

- a **linguist**: data can be biased by **theoretical assumptions**; ❌

a **native speaker** / a **language expert**: data can be biased by linguistic attitudes. ❌

Descriptive grammars are different from prescriptive grammars i.e., those grammars we had at schools teaching us how to write or speak correctly.
One of the great advantages of descriptive grammars over the other two data sources is that the grammatical system is presented in a comprehensive way, not just focusing on single phenomena or grammatical traits. However, grammar's comprehensiveness should be exploited: one of the skills of the typologist is to compare different grammatical traits from a grammar, not just cherry-picking them.

If the grammar is written by a linguist, there's the risk that it's not theory-free i.e., linguistic data can be biased by theoretical assumptions. A generative linguist will focus perhaps more on syntactic constructions, someone trained in Morphology will discuss derivation and inflection at length, and so on.

On the other side, if the grammar is written by a native speaker or a language expert (a translator, someone who has proficiency with the language, …), the description of language may reflect the linguistic usage and attitudes of the writer, for instance describing only some registries of the language.

Typological classification

A grammatical phenomenon can be instantiated by different **strategies** within and across languages.

A typological research starts with a survey of the **strategies** instantiating the phenomenon.

Strategies have a meaningful part, which is defined on **external basis** (**meaning and function**) and a formal part, of which we will try to provide a typology.

As we have seen in the previous slides (12 to 17), we can distinguish between the formal part of a grammatical phenomenon, the strategy, and its meaningful part, which ultimately has an external motivation. (We leave aside for the moment the fact the strategy is a linguistic sign in itself and is articulated into a meaningful part, again with an external motivation, and a formal part. We will analyse here morpho-syntactic strategies i.e., strategies whose formal structure corresponds to the domain of morphology and syntax.)

A cross-linguistically valid description of morpho-syntactic structures

- **simple strategies**: **no additional morphemes** are used to signal the function;
- **relational** and **indexical strategies**: **additional morphemes** are used to signal the function.

Examples will focus on **possessive construction**, but the following types of strategies applies to the vast majority of **linguistic phenomena**.

- **Possessor**: modifier;
- **Possessum (possessed item)**: head.

More precisely, we can propose a list of strategies encoding a given function; to exemplify the matter, let's take the example of the possessive construction:

- Grammatical phenomenon: possessive construction;
- Meaningful part (function): "the semantic relationship of ownership as used when the speaker intends to refer to the possessum (possessed item); i.e. the possessum is the head of the possessive noun phrase and the possessor is a modifier" (p. 32);
- Formal part (morpho-syntactic structure): various strategies.

A first distinction can be made between strategies that do not employ additional morphemes in order to encode the function of the grammatical phenomenon and strategies and strategies which make use of additional morphemes; the latter type is divided into relational and indexical strategies, a distinction based (again!) on external motivation.

How to read linguistic data

- **first line: original text (L1);**
- **second line: interlinear morphemic glosses,** i.e**. grammatical** (in **Leipzig glosses**) and **lexical** (in **L2**) **morphemes;**
- **third line: translation** in **L2.**

Example: (possessive construction in Russian)

kniga Ivan-a

Book Ivan-GEN

'Ivan's book'

A common practice in Linguistics is to present linguistic data with glosses and translation. Linguistic data is then organised on three lines: the original text, the interlinear morphemic glosses and a translation. The original text is of course given in the original language (L1), while translation and lexical morphemes are given in an auxiliary language, L2, which is usually English. Grammatical morphemes are described according to the 'List of Standard Abbreviations' described in the Leipzig Glossing Rules (leipzig-glosses.pdf).

Simple strategies

No **additional morpheme** beyond possessum and possessor:

- **juxtaposition**: possessum and possessor are **simply juxtaposed**;
- **morphological concatenation**: the possessor is **attached as an affix** to the possessum;
- **morphological compounding**: the possessor is **attached as a lexical root** to the possessum;
- **morphological fusion**: possessor and possessum **are fused** into one unit.

Simple strategies are not attested at our latitudes, but they are quite frequent worldwide.

The first type of simple strategy consist in put together (juxtaposing) the possessed item and the owner of the item.. Cfr. WALS map no. 24: https://wals.info/feature/24A#2/26.1/153.1, feature: no marking. Cfr. Yoruba and Kobon.

In the second and third type of simple strategy the possessor is attached to the possessum, either as an affix (pronoun) or as a lexical item (lexical root).

Finally, the fourth type is quite rare and is attested mostly for inalienable possessor (kin terms): the possessor is fused into the possessum, and the two items cannot be recognised as standalone units in synchrony. For instance, in Lakhota we have three different words for 'my mother', 'your mother' and 'his/her mother'.

Relational strategies

Relational: the additional morpheme relates to the semantic relation that holds between the two items.

Bound morphemes: **case affixes**, which can be attached to:

- the **modifier** (possessor: dependent marking);
- the **head** (possessum: head marking).

Free morphemes: **adpositions**.

Case markers are **relational strategies.**

Recall the two definitions given at pages 16-17 on Relational vs. Indexical strategy. The first type of relational strategy (bound to the modifier: possessor) is common across Europe. For instance, in Russian (ex. 6), a case marker in the form of suffix –a is attached to the possessor to denote the relation of ownership with the possessum. Cfr. WALS map no. 24: https://wals.info/feature/24A#2/26.1/153.1, feature: Possessor is modifier-marked. The case affix can be also bound to the head i.e., to the possessed item (possessum), as in Fijan. Cfr. WALS map no. 24: https://wals.info/feature/24A#2/26.1/153.1, feature: Possessor is head-marked. This is uncommon in Europe.

Finally, case markers can also take the form of prepositions or postpositions. This is for instance quite common in Romance, Germanic and – to some extent – Slavic languages, such as the example reported by Croft for Bulgarian (ex. 7). Question: which strategy is employed in English? And in German?

Indexical strategies

Indexical: the additional, bound morpheme (affix) denotes the argument itself.

- Also called '**agreement markers**': they **agree** with a **controller,** which however **is not always present**.

Two types of indexical strategies:

- **person**: category of person;
- **nonperson**: other categories (gender, number, case, …).

Possessive indexical strategy: it indexes the possessor on the possessum.

An index is a semiotic sign that stands for another sign, showing evidences of it: examples of indexes are the smoke of a fire, the symptoms for a disease, the footprint for an animal, and so on.
Indexical markers largely coincide (but do not entirely overlap) with agreement markers i.e., morphemes signalling agreement with a controller, which corresponds to the head. For instance, in the German sentence *Wir sehen uns in Paris*, the morpheme *–en* agrees with the controller *Wir*: in some sense, it represents in the verb the evidence of the subject *Wir* .

We can distinguish between two types of indexical strategies on the basis of what's indexed.

If the index denotes the category of person, then we speak of person indexical strategy. For instance, in WALS map no. 24 we have a feature called 'Double marking', which is found in some south-eastern European languages, such as Greek and Turkish and in one northern European language, Finnish. Let's take a closer look to an example from Turkish (Turkish_possessive.pdf)

Turkish:

Hafta-nīn gün-ler-i
Week-POSS day-PL-3SG
'Days of the week'

The suffix –i is an indexical suffix indicating that the owner is a third person, while -nīn is a relational suffix indicating the possession. The Turkish indexical suffix –i is identical to the strategy used in Mam (p. 35).

If the index denotes a non-person category, we call the indexical strategy 'nonperson'. Nice examples of nonperson indexical markers are adfixes found on adjectives; for instance, in French the suffix –a indicates that the owned object is of female gender.

French
M-a soeur
1SG.POSS-F.SG sister.(F).SG
'my sister'

French and Russian then use the same nonperson indexical strategy: cfr. the suffix –ja in Croft's example (16).

Classifiers: indexical or relational?

Classifiers denote an item for one of its **property**: shape, type of referent (human, animal, plant, object, …).

Similarly to relational morphemes they can be either bound or free morphemes. Depending on the construction, we have:

- **possessive** classifiers;
- **numeral** classifiers;
- **verbal** classifiers.

A type of strategies that is very rare in European languages is classifiers, which denotes an item (usually an object) for one of its property: its shape (round, elongated), type of referents (liquid or concrete, human or animal or plant, gender, age). It is difficult to classify classifiers as either indexical or relational, since they probably cover both roles:

- They index, as they indicate how the object itself look like;
- They relate, as they act as an interface between the items involved in a construction.

As for the possessive construction, a possessive classifier can be for instance found on the possessor in order to refer to (index) the possessed object. For instance, in Kosraean (an Austronesian language spoken in Micronesia), the classifier *SAnA* indicates that the possessed item is a plant. (Croft's example no. 23)

In fact, possessive classifiers are not so widespread. Numeral classifiers are by far the most attested type of classifiers: let's take a look to Map no. 55A 'Numeral classifiers'. On a sample of 400 languages, 140 show numeral classifiers, and in 78 of these languages numeral classifiers are mandatory. Just three languages spoken in Europe have numeral classifiers and they are only optional strategies. Let's see an example

from Turkish: (Turkish_numerals.pdf)

Turkish
bir el tabanca attı
One CLF pistol-shot fired
He fired one pistol-shot

In which the classifier *el*, which means ‘hand’, is used with certain abstract objects, such as ‘shots of firearm’ or ‘deals of cards’.

Numeral classifiers are mandatory in many languages spoken in Asia, such as Chinese and Japanese. Croft’s example (24) is from Chrau, an Austroasiatic language spoken in Vietnam; in order to count a crossbow, the classifier for ‘long objects’ must be employed.

Finally, the last type of classifiers is used with verbal constructions: the classifier marks one of the verbal arguments, looking very similar to an indexation marker. However, here the marker is specified according to one of the property of the argument. In Croft’s example (25) the prefix dân- is attached to the verb to denote that the object to be cooked is of granular nature.

More grammaticalized strategy

Linkers are markers that cannot be classified as relational or indexical.

- "There is just one invariant morpheme that is used to code the dependency" (p. 38): the morpheme is not paradigmatically opposed to other morphemes in the language.

Croft's classification of morpho-syntactic strategies adopts both a synchronic and a diachronic perspective.
As for the synchronic perspective, we have seen that case affixes and agreement markers are found in a paradigm i.e., we have a paradigm for cases (for instance German –n vs. –s) and a paradigm for agreement markers (for instance German –en vs. –e). We can also talk of a paradigm for classifiers, since in many languages spoken in Asia we have to select the right classifier among dozen of classifiers, but it is less clear if classifiers denote a relation or stands for something else.
Linkers do not form a paradigm, as they are found without any opposing morphemes. For instance, the English strategy for the possessive construction, the linker 's, is not opposed to any other strategies i.e., we don't have other strategies denoting different constructions. Other examples include the Persian *ezafe linker –é-* and the Moroccan Arabic *dyal*; both linkers mark again the possessive constructions in those languages. (examples 32 and 33).

Summary

Three **grammatical properties** describing strategies:

- **additional morpheme**: none, relational, indexical, linker;
- **degree of fusion of elements**: none, concatenation, fusion;
- **order of elements**.

Through **grammaticalization** processes, morpho-syntactic strategies vary over time, originating from **different linguistic items**.

We have seen that strategies are described according to two grammatical properties: type of additional morpheme (if any) and degree of fusion of elements. For instance, we can define affix markers as 'concatenated relational strategies' and adpositions as 'juxtaposed relational strategies'. We have also seen that the order of elements is also meaningful, leading to the possibility of marking the head or the modifier.
The interesting thing is that strategies vary over time, originating from different linguistic items: the box (figure 2.1) summarizes the different grammaticalization processes from which morpho-syntactic originate.
Here are some examples.

Lexical items to adpositions (independent word to concatenation)
On page 34, Croft discusses the case of the Tzutujil adposition *majk 'because of'* which derives from a lexical item meaning *'sin'.*

Case affixes to linker
Let's take the example of the English linker 's. Similarly to other German languages, Old English (Anglo-Saxon English) was an inflected language like German, with a case suffix –s denoting the genitive. Over time, English lost all its case markers except for the genitive case, which, however, cannot be longer considered a case marker since

it’s not paradigmatically opposed to other case markers.

What's being classified?

Modern Linguistic Typology is not concerned anymore with the description of **language type** but with the description of **linguistic type**.

- **language type: languages can only have one type of strategy.** For instance, they show adpositions or case markers in every grammatical construction;
- **linguistic type: languages may have more than one type of strategy.** For instance, they can have two types of word orders, say, VO an OV.

The idea that languages belong to just one type dates back to the 19th century, where languages were classified according to four morphological types: isolating, agglutinative, fusive and incorporating, i.e. types based on the shape that words assume.

As always, WALS gives a nice exemplification of this topic. Let's take Map no. 20 https://wals.info/feature/20A#2/26.7/156.6 and discussion in https://wals.info/chapter/20. Please note that the classification is based ONLY on case and tense-aspect markers, i.e. relational strategies denoting syntactic relations and two verbal features.

Isolating languages are languages in which morphological strategies are not employed i.e, words do not have affixes, only juxtapositions. The most popular example is Chinese, but many languages are isolating, such as Boumaa Fijian. The relevant datapoint in WALS no 20 is exclusively isolating.

In agglutinative languages we have morphological strategies marking just one function. Here, the canonical example is Turkish. Inflectional languages are similar to agglutinative languages, but here morphological strategies may carry more than one function. WALS no.20 conflates these two features in the 'exclusively concatenative' datapoint.

Finally, words in incorporating languages correspond to sentences in our inflectional

languages.

What's being classified?

A linguistic type is NOT basic is:

- It is **restricted** to certain grammatical categories;
- It has **semantic or pragmatic specializations**;
- It is **structurally unusual**;
- It is **less frequent** with respect to other linguistic types.

Let's go back to modern Linguistic Typology: once recognized that languages do not belong to just one type, showing only linguistic type, we can still classify languages according to the BASIC linguistic type.

Let's take the word order in German: what is the basic linguistic type?

1. SOV is syntactically restricted to subordinate clauses:
2. OVS is pragmatically and semantically restricted to topicalization constructions: *Diesen Mann habe ich lange nicht mehr gesehen*
3. SOV is found in constructions with additional structure, such as the cleft construction: *Es ist Hans, dem ich einen Briefe geschrieben habe,* which uses '*Es ist*';
4. If we count clauses in a collections of text, we probably find that clauses with SVO are more frequent than clauses with SOV and OVS.

Then the basic word order in German is SVO.