Parts of speech and dependent clauses: A typological study

van Lier, E.H.

Publication date
2009

Sampling, Research Questions, Method

4

4.1 Introduction

The aims of this chapter are as follows: First, in section 4.2, it presents the sample of languages investigated in this study. Second, in section 4.3, the concrete research questions that will be addressed in this study are formulated and operationalized. Third, section 4.4 explains the method that will be used to answer these questions. Finally, section 4.5 provides a brief outlook on the remaining part of the study.

4.2 The language sample

The sample used for this study consists of 50 languages and is composed by means of the Diversity Value Technique (henceforth DVT; Rijkhoff et al. 1993, Bakker forthcoming). This technique can be used with any language classification. A sample composed with the DVT consists minimally of one language per family, according to the classification chosen. Absolute isolates (as opposed to isolates within a genetic grouping) are always part of any sample. Once such a minimal sample is put together, it can be expanded – depending on the researcher's wishes – on the basis of so-called diversity values. These values are assigned to all nodes in the language family tree. They express the complexity of the tree below the node in terms of the amount of daughter nodes and the way they are embedded. The diversity value determines the proportion of languages to be drawn from under the relevant node in relation to its sister nodes, given a certain desired sample

size. Once the number of languages to be selected from a particular family is determined, the linguist chooses those languages for which the best descriptive grammars (and/or other data sources) are available. Preferably, the languages should come from different subgroups within their respective families (Bakker, forthcoming).

In principle, the DVT composes so-called *variety* samples, which are used to explore the range of diversity in a relatively little-studied linguistic domain. Variety samples stand in contrast to *probability* samples, which are used to determine significant correlations between grammatical traits (Croft 2003: 23). In a sample of the latter type, the languages represent independently selected cases. Thus, this type of sample explicitly avoids the situation in which grammatical traits are shared between languages as a result of descent from a common ancestor or through language contact. Samples that are selected without controlling for these factors are said to be *genetically* and/or *areally biased*.

Although the DVT thus aims primarily at maximal genetic diversity, it can also be used for the composition of what Bakker (forthcoming) calls *pseudo-probability samples*. These are relatively small variety samples with a relatively high degree of independence between the selected cases. They can be composed by combining the DVT (applied to an up-to-date language classification) with an areal classification as a stratifying dimension. The rationale behind the use of small variety samples as pseudo-probability samples is that languages that belong to different families vary along certain structural parameters. Therefore, in the initial stage of the sampling process, the goal of covering maximal diversity (relevant for a variety sample) and the goal of avoiding genetic bias (relevant for a probability sample) can be reached by the same procedure: picking one language per family. Combining this method with an areal stratification allows one to avoid the second major source of bias for a probability sample: feature sharing as a result of language contact. Only when variety samples are expanded in order to cover a wider range of variation (which may be attested in specific genetic or areal groupings) do they become fundamentally ill-suited for probabilistic research, since the languages can no longer be regarded as independent cases[61].

Hengeveld et al. (2004) use a pseudo-probability sample for their study on the relation between PoS systems and word order constraints. This 50-language sample is composed by applying the DVT to Ruhlen's (1991) classification and then choosing languages that are spoken in non-contiguous areas, at least whenever the bibliographical situation would permit it.

Notably, Ruhlen's classification is controversial to the extent that it distinguishes a relatively small number of (large) language families: 19 families and a number of isolates[62]. This number is much smaller than for instance in the Ethnologue classification, which distinguishes 120 families (Gordon 2005). As mentioned above however, the DVT requires a minimum of one language per family (see Rijkhoff & Bakker 1998). In view of this requirement, Ruhlen's classification provides a suitable basis for composing a relatively small language sample, which is desirable in the case of a pseudo-probability sample, and also in practical terms of feasibility. The Ethnologue classification, in contrast, yields a sample of minimally 120 languages. Note also that the specific methodology of the DVT partly makes up for the more extensive 'lumping' in Ruhlen's classification by dictating the selection of more languages from families with larger internal diversity.

The sample used in the present study is very similar to the sample used by Hengeveld et al. (2004). The first reason for this is a practical one: It allowed me to take as a point of departure for most languages the PoS system classifications as proposed by Hengeveld et al. (2004). Nevertheless, I have adapted my sample in some respects. First, I have replaced a few languages from the sample of Hengeveld et al. for bibliographical reasons. The relevant cases are listed below:

- Hdi instead of Gude (Chadic, Afro-Asiatic)
- Dhaasanac instead of Oromo (Cushitic, Afro-Asiatic)

[61] Maslova and Nikitina (submitted) have recently challenged the assumption behind 'classical' probability sampling, namely that language family size is a 'historical accident', and that therefore every family should be represented in the sample with an equal number of languages, in order to avoid disproportial representation of certain language types. They argue that family size is determined by *transition probabilities*, which are in turn determined by *language constants*, i.e. universal properties of languages and their users. Therefore, they argue for the use of 'truly' random samples (so-called *R-samples*), rather than samples containing one randomly selected language per family (so-called *I-samples*). R-samples will contain a higher percentage of more stable language types. On the other hand, frequency distributions in R-samples may be due to an accidentally higher initial frequency of some source type, rather than to a systematically higher amount of transitions to this type. Therefore, R-samples and I-samples should be compared in order to find out whether the distribution attested in the modern language population is indeed the outcome of differences in transition probabilities. If yes, then it constitutes evidence for a statistical universal. If not, then the comparison can still indicate tendencies towards a linguistically meaningful equilibrium: If a more stable trait occurs less frequently synchronically, then there will be a gradual increase of languages with this trait. Conversely, if a less stable trait occurs more frequently, then the number of languages with this trait will decrease over time.

[62] In the second edition of Ruhlen, Korean-Japanese-Ainu and Kartvelian are distinguished as separate families, while in the first edition they were considered sub-branches of Altaic and Caucasian, respectively (Hengeveld et al. 2004: 528, note 4).

- Lavukaleve instead of Nasioi (East Papuan, Indo-Pacific)
- Abun instead of Tidore (West Papuan, Indo-Pacific)
- Ma'di instead of Ngiti (Central Sudanic, Nilo-Saharan)
- Slave instead of Navaho (Na-Dene)
- Gooniyandi instead of Ngalakan (non-Pama-Nyungan (Bunaban), Australian)
- Kharia instead of Mundari (Munda, Austro-Asiatic, Austric)

The second difference between my sample and that of Hengeveld et al. (2004) is that I have expanded mine with two languages with a flexible PoS system, since the most interesting data were obtained for this type of languages. The added languages are:

- Santali, which comes from the same sub-family as Kharia: Munda, Austro-Asiatic, Austric;
- Kambera, which belongs to the same family as Samoan and Tagalog: Malayo-Polynesian, Austronesian.

Obviously, adding these languages results in areal and genetic biasing of the sample[63]. For this reason, Santali and Kambera are not taken into account when calculating statistical dependencies between grammatical traits (see Chapter 7). Rather, they are used as a further back-up to the patterns found in the other flexible languages of the sample[64].

Finally, two further details should be noted. First, the present study does not include any data on the extinct languages that are part of the sample used by Hengeveld et al. (2004), namely Etruscan, Hurrian, Hittite, Nahali, and Sumerian. Furthermore, in two cases I use a different language name than Hengeveld et al. (2004): *Bukiyip* instead of *Mountain Arapesh*, and *Hmong Njua* instead of *Miao*.

[63] As Hengeveld et al. (2004: 529) note, the inclusion of Tagalog in their sample (and also in mine) represents a violation of the genetic criterion, since it does not belong to the Formosan branch of the Austric family, which would have been the right choice. It was included in Hengeveld et al. in order to represent a language with a maximally flexible PoS system. Even though this choice is relevant in view of the present study as well, it undermines the value of the sample as a pseudo-probability sample, since it involves deliberately choosing a case (i.e. a language) on the basis of the grammatical phenomenon being studied. I thank Elena Maslova for emphasizing this point.

[64] With the same goal in mind, some data from some other flexible languages have been and will be discussed, including Tongan, Maori, and Mundari (see Chapters 2 and 5).

The total number of languages in the sample is 50, as can be seen in Table 4.1 below. In Appendix 1, a more complete specification of the sample is provided, in which the languages are also classified for family and subfamily according to the Ethnologue, and for Genus according to the World Atlas of Language Structures (WALS)[65].

As this section makes clear, a language sample is composed with a specific typological research question in mind. Variety samples are in principle not appropriate to answer probabilistic research questions, since they are specifically designed to manipulate chances of occurrence: They intend to maximize the likelihood of capturing all the linguistic diversity for the phenomenon under study (Croft 2003: 21). However, small, genetically and areally balanced variety samples, of the type used in Hengeveld et al. (2004) and in the present study, can indeed be used as pseudo-probability samples, on which to test hypotheses about dependency relations between structural traits, in this case between various parameter values related to PoS and DCs. Bearing this in mind, in the following section I turn to the specific formulation and operationalization of the research questions of this study.

Family (Ruhlen 1991)	Language(s)
Afro-Asiatic	Hdi, Dhaasanac
Altaic	Turkish
Korean-Japanese	Japanese
Amerindian	Pipil, Hixkaryana, Tuscarora, Koasati, Guaraní (Paraguayan), Warao, Imbabura Quechua
Australian	Kayardild, Nunggubuyu, Gooniyandi
Austric	Tagalog, Samoan, Kambera, Paiwan, Garo, Thai, Santali, Kharia, Hmong Njua
Caucasian	Abkhaz
Kartvelian	Georgian
Ckukchi-Kamchatkan	Itelmen
Elamo-Dravidian	Tamil
Eskimo-Aleut	West Greenlandic
Indo-Hittite	Polish
Indo-Pacific	Wambon, Alamblak, Lavukaleve, Abun, Bukiyip
Isolates	Ket, Burushaski, Basque, Nivkh

[65] See *www.ethnologue.org* (or Gordon 2005) and *http://wals.info/* (or Haspelmath et al. 2005).

Family (Ruhlen 1991)	Language(s)
Khoisan	Nama
Na-Dene	Slave
Niger-Kordofanian	Babungo, Kisi, Bambara, Krongo
Nilo-Saharan	Lango, Ma'di
Pidgins and creoles	Berbice Dutch Creole
Sino-Tibetan	Nung, Mandarin Chinese
Uralic-Yukaghir	Hungarian

Table 4.1: The language sample

4.3 Research Questions

4.3.1 Introduction

In the previous chapters, it was shown that PoS classes and DCs can be defined as primary and secondary constructions, respectively, expressing the propositional functions of predication, reference, and/or modification. On the basis of this parallel, it can be hypothesized that there will be a match between the distributional patterns of PoS classes and DC constructions in any language. In particular, it is expected that the functional possibilities of the (secondary) DCs constructions in a language are dependent on those of the (primary) PoS categories available in that language. This hypothesis will be investigated from two converging perspectives:

(i) The perspective of *global matches*, i.e. in terms of the presence of functional flexibility versus rigidity in PoS systems as compared with systems of DCs constructions (section 4.3.2);
(ii) The perspective of *specific matches* between particular types of flexible and rigid PoS classes and DCs constructions, in terms of set of functions that can be expressed by them (section 4.3.4).

In addition, and cross-cutting these two perspectives, I will investigate the influence of another parameter, namely the internal morpho-syntactic characteristics of the DC construction(s) under scrutiny (see sections 4.3.3 and 4.3.5). I will take this parameter into account by testing each of the hypotheses to be formulated in the next subsections in two ways:

(i) Considering DCs as a single, undifferentiated construction type, and

(ii) Splitting up the hypothesis into several sub-hypotheses, each of which addresses a different structural DC type.

In what follows, a set of fully explicit hypotheses will be formulated and operationalized according to the general research design just outlined.

4.3.2 Global functional matching

First, it is predicted that a *global match* exists between flexibility or rigidity as attested in the PoS system of a particular language, and flexibility or rigidity in the set of DC constructions in that language. This hypothesis is operationalized in the form of the two-fold prediction A1/A2, as given in (1a, b):

Predictions A1/A2

(1) a. If a particular language has *one or more* flexible DCs, then it should also have *one or more* flexible PoS classes.
b. If a language has rigid DCs *only*, then it should also have rigid PoS classes *only*.

The phrases in italics in the above predictions require some explanation. Recall from Chapter 2 that PoS systems are called *flexible* when they include *one or more* flexible lexeme class(es). This means that a language with a flexible PoS system does not need to have flexible PoS classes *only*. In fact, Tagalog and Kharia are the only two languages in the sample that do not have any rigid PoS classes. All other languages with flexible PoS systems display a mixture of flexible and rigid PoS classes. In contrast, rigid PoS systems were defined as consisting of rigid PoS classes *only* (even though not all rigid systems have a rigid PoS class for every propositional function). These asymmetric definitions also apply to the predictions in (1) above, as the phrases in italics are meant to indicate: (1a) makes reference to languages with flexible PoS systems, and (1b) to languages with rigid PoS systems.

4.3.3 Global matching including differentiation for structural DC type

The predictions in (1) can be made more specific by taking into account the internal morpho-syntactic characteristics of the DC construction(s) under study. In particular, this is done by adding a parameter that differentiates between the three structural types of DCs defined in Chapter 3, section 3.4. They are summarized once more in Table 4.2 (which is the same as Table 3.1):

DC type		Argument expression		TAM/Person	DET/CASE
number	label	1	2		
Type 1	B	SENT/Ø	SENT/Ø	+	+/–
Type 2	D-SENT	SENT/Ø	SENT/Ø	–	+/–
Type 3	D-ALT	ALT	SENT/ALT/Ø	–	+

Table 4.2: DC types and their internal formal properties

Considering the morpho-syntactic properties of the three DC types in Table 4.2, we can say that:

(i) Type 1 DCs (Balanced) do not show any formal reflection of de-categorization (i.e. TAM/Person is expressed as in independent clauses), nor of re-categorization (arguments, if overt, are SENT, i.e. expressed as in independent clauses)[66].
(ii) Type 2 DCs (D-SENT) show formal reflections of de-categorization (i.e. TAM/Person is (partially) lost), but not of re-categorization (arguments, if overt, are SENT, i.e. expressed as in independent clauses).
(iii) Type 3 DCs (D-ALT) show formal reflections of both de-categorization (i.e. TAM/Person is (partially) lost) *and* re-categorization (at least one argument is expressed with an ALT strategy, i.e. different from independent clauses).

A larger amount of formal de-/re-categorization of a DC construction implies that it has less morpho-syntactic characteristics of a clause and more of a lexical construction. It is hypothesized is that the more *formally* similar a DC construction is to a lexical expression, the more *functionally* similar the DC will be to its lexical counterpart. This means that deranked DCs, i.e. type 2 (D-SENT) and 3 (D-ALT), are expected to show more functional similarity with lexical categories than balanced DCs, i.e. type 1 (B). Furthermore, within the group of deranked DCs (types 2 and 3) it is expected that the distributional patterns of more deranked DCs (type 3) will be functionally more similar to lexical expressions than those of less deranked DCs (type 2). Thus, we can formulate the two sub-predictions B1 and B2, as in (2a) and (2b):

[66] Recall that determiners and/or case-markers are not regarded as features reflecting re-categorization, because they do not affect the internal structure of the construction on which they operate.

Prediction B1/B2:

(2) a. The functional possibilities of deranked DCs (type 2/3) are more similar to the functional possibilities of PoS than those of balanced DCs (type 1).
b. Within the group of deranked DCs, the functional possibilities of type 3 DCs are more similar to the functional possibilities of PoS than those of type 2 DCs.

Note that, even though it has never been tested empirically on a larger scale, the hypothesis of a functional connection between lexeme classes and deranked DC constructions is not new. In fact, deranked complement clauses (nominalizations/infinitives), relative clauses (participial clauses), and adverbial clauses (converbal constructions) are traditionally characterized as the syntactically derived clausal counterparts of nouns, adjectives and adverbs, respectively (Croft 1991, Koptjevskaja-Tamm 1993, Haspelmath 1994, 1995). These connections between PoS and deranked DCs can be summarized as in Table 4.3 (adapted from: Haspelmath 1995: 3-4)[67].

Forms	Functions		
	Ref Head	Ref Mod	Pred Mod
PoS	noun	adjective	(manner) adverb
deranked DC type	nominalization/infinitive construction	participle construction	converb construction

Table 4.3: Functional connection between PoS and deranked DCs

We can now combine predictions B1 and B2 in (2) above with predictions A1 and A2 in (1) above. For *flexible* constructions, this yields the four sub-hypotheses given in (3a-d):

(3) a. If a language has one or more flexible balanced DC(s) of type 1, then it should also have one or more flexible PoS class(es).
b. If a language has one or more flexible deranked DC(s) of type 2/3, then it should also have one or more flexible PoS class(es).

[67] Notably, the PoS classes and DCs defined in Chapters 2 and 3 make reference to 4 propositional functions: in Table 4.3 the function of head of a predicate phrase is not taken into account. In Chapter 6 it will become clear that this functional slot is irrelevant for the present study to the extent that there are no (or hardly any) DC constructions that express it.

c. If a language has one or more flexible deranked DC(s) of type 2, then it should also have one or more flexible PoS class(es).
d. If a language has one or more flexible deranked DC(s) of type 3, then it should also have one or more flexible PoS class(es).

On the basis of the structural differences between the DC types, it is expected that the prediction in (3a), involving balanced clauses, is less likely to receive empirical support than the one in (3b), concerning deranked clauses. Further, the prediction in (3c), involving less deranked clauses of type 2, is expected to be less probable than the one in (3d), with concerns more deranked clauses of type 3.

Note that, due to the asymmetrical definitions of languages with flexible versus rigid PoS systems, it does not make sense to formulate specific predictions about implicational relations between the presence of rigid PoS only and the presence of rigid DCs *of a specific behavioural potential type only*. This is because, on the basis of the available typological evidence discussed in Chapter 3, one would not predict that languages express all subordination relations with rigid DCs of just a single structural type. Rather, semantically different types of subordination relations are likely to be expressed by morpho-syntactically different DC constructions. Therefore, no separate sub-hypotheses are formulated that make reference to and rigid DCs of type 1/2/3 only.

4.3.4 Specific functional matching

I now turn to the second perspective on the relation between PoS and DCs, which involves the investigation of specific matches between the functional possibilities of particular PoS classes and DC constructions. We have seen in Chapters 2 and 3 that there are 10 predicted types of flexible constructions (lexical and clausal), depending on how many and which propositional functions they can express. These flexible construction types are listed once more under (4):

(4) a. Lexical and clausal constructions that can be used in all four propositional functions: *contentives* and *contentive clauses*;
b. Lexical and clausal constructions that can be used in all propositional functions except the head of a predicate phrase: *non-verbs* and *multi-functional clauses*;

c. Lexical and clausal constructions that can be used as the head and modifier in a referential phrase: *nominals* and *nominal clauses*;
d. Lexical and clausal constructions that can be used as modifiers in referential and predicate phrases: *modifiers* and *modifier clauses*;
e. Lexical and clausal constructions that can be used as the head and modifier in a predicate phrase: *predicatives* and *predicative clauses*;
f. Lexical and clausal constructions that can be used as heads of referential and predicate phrases: *heads* and *head clauses*;
g. Lexical and clausal constructions that can be used as heads of referential and predicate phrases and as modifiers in referential phrases: *Flex PoS A* and *Flex clause A*.
h. Lexical and clausal constructions that can be used as heads of referential and predicate phrases and as modifiers in predicate phrases: *Flex PoS B* and *Flex clause B*.
i. Lexical and clausal constructions that can be used as the head of a referential phrase and as a modifier in a predicate phrase: *Flex PoS C* and *Flex clause C*.
j. Lexical and clausal constructions that can be used as the head of a predicate phrase and as a modifier in a referential phrase: *Flex PoS D* and *Flex clause D*.

As regards rigid constructions, four construction types were predicted, each of them specialized for the expression of a single propositional function. They are listed in (5):

(5) a. Lexical and clausal constructions that are specialized for the function of head of a predicate phrase: *verbs* and *predicate clauses*.
b. Lexical and clausal constructions that are specialized for the function of head of a referential phrase: *nouns* and *complement clauses*;
c. Lexical and clausal constructions that are specialized for the function of modifier in a referential phrase: *adjectives* and *relative clauses*;
d. Lexical and clausal constructions that are specialized for the function of modifier in a predicate phrase: *manner adverbs* and *adverbial manner clauses*.

It is hypothesized that each DC of a specific flexible/rigid type, being a secondary construction, will have a PoS class of the same flexible/rigid type as its primary counterpart. This hypothesis is operationalized in the form of prediction C in (6):

Prediction C:

(6) If a language has a DC construction of a flexible or rigid type X, then it should also have a PoS class of type X.

Applying Prediction C to each of the specific flexible constructions listed in (4), we arrive at the set of predictions in (7):

(7)
a. If a language has contentive clauses, then it should also have lexical contentives.
b. If a language has multi-functional clauses, then it should also have lexical non-verbs.
c. If a language has nominal clauses, then it should also have lexical nominals.
d. If a language has modifier clauses, then it should also have lexical modifiers.
e. If a language has predicative clauses, then it should also have lexical predicatives.
f. If a language has head clauses, then it should also have lexical heads.
g. If a language has Flex clause A, then it should also have Flex PoS A.
h. If a language has Flex clause B, then it should also have Flex PoS B.
i. If a language has Flex clause C, then it should also have Flex PoS C.
j. If a language has Flex clause D, then it should also have Flex PoS D.

A parallel list of predictions for the specific rigid constructions listed in (5) appears in (8):

(8)
a. If a language has predicate clauses, then it should also have lexical verbs.

b. If a language has complement clauses, then it should also have lexical nouns.
c. If a language has relative clauses, then it should also have lexical adjectives.
d. If a language has adverbial manner clauses, then it should also have lexical manner adverbs.

The predictions in (7) and (8) can be further fine-tuned by taking into account the parameter of structural DC types. This is done in the next section.

4.3.5 Specific matching including differentiation for structural DC type

The predictions in (7) and (8) concerning specific flexible and rigid construction types are now combined with predictions concerning different structural DC types. Starting with maximally flexible constructions, i.e. lexical contentives and contentive clauses, this yields the set of predictions listed in (9a-d). The expected likelihood that each of the implicational relations in (9) will actually hold, is the same as for the set of implications in (3) above. The implication in (9a), concerning balanced DCs, is less likely to receive empirical support than the one in (9b), concerning deranked DCs. Further, the implication in (9c), which makes reference to less deranked DCs of type 2, is less likely to hold than the one in (9d), involving more deranked DCs of type 3.

(9) a. If a language has balanced contentive clauses of type 1, then it should also have lexical contentives.
b. If a language has deranked contentive clauses of type 2/3, then it should also have lexical contentives.
c. If a language has deranked contentive clauses of type 2, then it should also have lexical contentives.
d. If a language has deranked contentive clauses of type 3, then it should also have lexical contentives.

Analogous sets of predictions can be set up for the remaining nine types of flexible constructions listed in (4)/(7) above, by means of substituting lexical contentives for non-verbs, nominals, modifiers, etcetera, and substituting contentive clauses for multi-functional clauses, nominal clauses, etcetera.

Similarly, the parameter of structural DC type is expected to interact with the predictions concerning different types of rigid constructions: For each of the rigid constructions listed in (5)/(8) a set of four sub-predictions can be set up along the lines of (9) above. Taking as an example lexical and clausal constructions specialized for the function of head of a referential phrase, i.e. nouns and complement clauses, we get the set of implicational relations in (10a-d). Again, the expectation is that the implication in (10a) is less likely to be born out than the one in (10b), and that the implication in (10c) is less probable than the one in (10d).

(10) a. If a language has balanced complement clauses of type 1, then it should also have lexical nouns.
b. If a language has deranked complement clauses of type 2/3, then it should also have lexical nouns.
c. If a language has deranked complement clauses of type 2, then it should also have lexical nouns.
d. If a language has deranked complement clauses of type 3, it should also have lexical nouns.

Analogous lists of testable implications can be set up for the other types of rigid constructions, by means of substituting nouns with verbs, adjectives, or manner adverbs, and by substituting complement clauses with predicate clauses, relative clauses, and adverbial manner clauses.

4.3.6 Summary

In sum, the general research question posed in this study is whether there is a match between the functional possibilities of lexical constructions (PoS) in a language and clausal constructions (DCs) in that language. This general question is approached from two converging perspectives: First, global matches are investigated between flexibility/rigidity as displayed by the PoS systems and the dependent clause constructions of particular languages. Second, specific matches are investigated between the functional patterns of particular types of flexible and rigid PoS classes and DC constructions. Across both perspectives cuts the additional parameter of structural DC type, defined in terms of behavioural potential. The specific predictions formulated above will be tested using a statistical method that is presented in the following section.

4.4 Method

The predictions formulated in the previous section take the form of classic Greenbergian implicational universals, namely: *(With more than chance frequency), if a language has structural characteristic A, then it has structural characteristic B* (Greenberg 1963; Cysouw 2005: 564). An implicational universal can thus be construed as a hypothesis about a particular type of dependency relation between A and B, such that one value of A (the positive one) constrains the value of B, whereas the other value of A (the negative one) does not (Maslova 2003: 103). This means that, in order to claim an implicational universal, two methodological steps are required, as described in (11a-b):

(11) a. Establishing a dependency relation between A and B, i.e. a statistically significant correlation between the co-occurrence in languages of characteristic A and characteristic B.
b. Establishing an asymmetrical dependency relation between A and B, such that the distribution of characteristic B among languages with characteristic A is more strongly skewed than the distribution of B among languages without A.

Consider first (11a): Establishing a dependency between A and B requires rejection of the *hypothesis of independence*, which states that the probability of any combination of values of A and B is just the product of the probabilities of these values taken in isolation. Statistical tests can be used to reject this hypothesis (Maslova 2003: 102).

In the present study, each of the predictions formulated in the previous section involves 2 binary parameters, one concerning PoS, the other concerning DCs. Each parameter can have two values: Either the particular type of PoS or DC is attested in the language ([+PoS], [+DC]), or it is not attested ([-PoS], [-DC]). For every prediction this yields a 2x2 contingency table with four cells. As is shown in Table 4.4 below, each of the cells in such a table represents one of the four possible combinations of the two binary parameters under investigation:

(i) [+PoS, + DC] = both the PoS and the DC feature are attested;
(ii) [+ PoS, - DC] = the PoS feature is attested, but the DC feature is not attested;

(iii) [- PoS, + DC] = the PoS feature is not attested, but the DC feature is attested;

(iv) [- PoS, - DC] = neither the PoS feature nor the DC feature is attested.

		DCs	
		+	–
PoS	+	(i)	(ii)
	–	(iii)	(iv)

Table 4.4: A 2x2 contingency table

The observed frequencies, i.e. the numbers of languages in the sample that display a particular feature combination (as mentioned in (i)-(iv) above), can be compared with the frequencies that would be expected if the co-occurrence of the PoS and the DC features would be purely coincidental. If this is not the case, i.e. if there is indeed a dependency relation between a particular PoS feature and a particular DC feature, then the observed frequencies in cells (i) and (iv) will be higher than the frequencies that are expected on the basis of chance, whereas the observed frequencies in cells (ii) and (iii) will be lower than the expected frequencies. A Fischer's Exact test can be used to calculate whether the deviation between observed and expected frequencies is statistically significant. This test produces a p-value between 0 and 1, which specifies how likely it is for the observed distribution to be the result of chance (Cysouw 2003: 91). The critical value used to identify a statistically significant correlation between two grammatical traits is $p<0.05$, meaning that there is a 0.5% chance that the observed frequencies are coincidental.

As Maslova (2003: 102) notes, if Fischer's Exact fails to reject the hypothesis of independence, i.e. if no significant correlation between the parameters can be established, this does not automatically mean that they really are independent: it can also be the case that the sample is too small to reveal the dependency. Following the guidelines of Cohen (1995), a sample size of 26 is needed to have an 80% chance of detecting a large effect with a test like Fisher's Exact. For the detection of a medium effect one needs a sample of 84, while small effects are detectable only with a sample as large as 785 cases. This means that the sample used in the present study suffices for the detection of medium to large effects. However, in the case of small effects, the sample will be too small for Fischer's Exact to be able to reject the hypothesis of independence.

Since Fisher's Exact is thus highly sensitive to sample size, I will report, in those cases where Fisher's Exact is statistically significant, the contingency coefficient (CC). This CC is a value between 0 and 1, which estimates the effect size independent of the sample size. A CC value of approximately 0.10 indicates a small effect, around 0.30 is a medium effect, and from 0.45 upwards is a large effect (Everitt 1977).

As mentioned in (11b) above, if Fischer's Exact yields a p-value below 0.05, then a second step must be taken, namely the identification of an *asymmetrical dependency relation*, such that the positive value of one parameter (in this case the PoS parameter) constrains the value of the other one (in this case the DC parameter). Maslova (2003) proposes a statistical method to establish such an asymmetrical dependency relation[68]. The basic idea is to correlate each of the two parameters, in this case the PoS parameter and the DC parameter, to a third, derived parameter: PoS = DC, which refers to the event of the PoS parameter and the DC parameter having the same value. This third parameter thus contrasts languages that confirm the correlation, i.e. languages in which PoS and DC have the same value (PoS = DC: either [+ PoS, + DC] or [- PoS, -DC]), with languages that disconfirm the correlation, i.e. languages in which PoS and DC do not have the same value (PoS ≠ DC: either [+ PoS, - DC] or [- PoS, + DC]). This involves two extra contingency tables, as illustrated with Tables 4.5a and 4.5b:

	PoS +	PoS –
PoS = DC		
PoS ≠ DC		

Table 4.5a: Maslova test no. 1: Change DC parameter to PoS=DC

	DC +	DC –
PoS = DC		
PoS ≠ DC		

Table 4.5b: Maslova test no. 2: Change PoS parameter to PoS=DC

Submitting the observed frequencies in two extra tables of this kind to Fischer's Exact tests may yield three possible results, listed in (12a-c)

[68] See also Cysouw (2005) for a brief assessment of this method.

(12) a. A *symmetrical* dependency: None of the tests shows a significant interaction, i.e. in both cases $p > 0.05$.
b. A *one-sided asymmetrical* dependency: One of the two tests shows a significant interaction, but not the other, i.e. in one case $p < 0.05$, and in the other $p > 0.05$.
c. A *two-sided asymmetrical* dependency: Both tests show a significant interaction, i.e. in both cases $p < 0.05$.

An example of (12a), a *symmetrical dependency*, is given in Tables 4.6 and 4.7a-b below. The original distribution appears in Table 4.6[69]. Applying Fischer's Exact to these frequencies yields a significant correlation, i.e. there is a dependency relation between the PoS and DC parameters. The two additional tables are *4.7*a and 4.7b, both of which yield a non-significant p-value. This means that neither the positive not the negative value of the PoS or the DC parameter imposes a constraint on the event of the two parameters having the same value.

Table 4.6: A symmetrical dependency; original distribution

	DC +	DC –
PoS +	12	40
PoS –	60	30

p < 0.05 (significant)

Table 4.7a: Distribution for PoS and PoS = DC (Maslova-test no. 1)

	Pos +	Pos –
PoS = DC	12	30
PoS ≠ DC	40	60

p < 0.05 (not significant)

Table 4.7b: Distribution for DC and PoS = DC (Maslova-test no. 2)

	Pos +	Pos –
PoS = DC	12	40
PoS ≠ DC	60	40

p < 0.05 (not significant)

[69] These 'observed' frequencies are just illustrations. They are taken from examples in Maslova (2003:105-106).

Second, Tables 4.8 and 4.9a-b illustrate the situation in (12b) above, i.e. a one-sided asymmetrical dependency. Table 4.8 shows the original distribution, which again yields a significant p-value, so that a dependency relation is established. Tables 4.9a and 4.9b represent the two additional tests. Table 4.9a yields a significant p-value, while 4.9b does not. More specifically, Table 4.9a shows that the positive value of the PoS parameter poses a strong constraint on the event of PoS and DC having the same value. In contrast, Table 4.8 makes clear that the event of (non-)attestation of the DC characteristic and the event of the PoS and the DC parameters having the same value are independent of each other.

Table 4.8: A one-sided asymmetrical dependency; original distribution

	DC +	DC –
PoS +	90	10
PoS –	35	28

p < 0.05 (significant)

Table 4.9a: Distribution for PoS and PoS = DC (Maslova-test no. 1)

	Pos +	Pos –
PoS = DC	90	28
PoS ≠ DC	10	35

p < 0.05 (significant)

Table 4.9b: Distribution for DC and PoS = DC (Maslova-test no. 2)

	DC +	DC –
PoS = DC	90	28
PoS ≠ DC	35	10

p < 0.05 (not significant)

This type of result may be interpreted as a statistical basis to formulate a weak unidirectional implicational universal (Maslova 2003: 106; Dryer 2003: 111) of the following form:

[+ PoS] ➔ [+ DC]
Languages with a positive PoS value significantly more often have a positive DC value (*regardless* of whether languages with a negative PoS value also tend to have a positive DC value).

Finally, Tables 4.10 and 4.11a-b illustrate the third type of result, a two-sided asymmetrical dependency (see (12c) above). The original distribution appears in Table 4.10, which again reveals a significant dependency relation. Moreover, the two extra tests, applied to the figures in Tables 4.11a and 4.11b, both yield a significant p-value.

Table 4.10: A two-sided asymmetrical dependency; original distribution

	DC +	DC –
PoS +	40	6
PoS –	50	60

p < 0.05 (significant)

Table 4.11a: Distribution for PoS and PoS = DC (Maslova-test no. 1)

	Pos +	Pos –
PoS = DC	40	60
PoS ≠ DC	6	50

p < 0.05 (significant)

Table 4.11b: Distribution for DC and PoS = DC (Maslova-test no. 2)

	DC +	DC –
PoS = DC	40	60
PoS ≠ DC	50	6

p < 0.05 (significant)

As in the previous case, the positive value of the PoS parameter poses a constraint on the event of PoS and DC having the same value. The significant correlation in Table 4.11a reflects this. In addition, the original distribution in Table 4.10 shows that the distribution of PoS for the negative value of the DC parameter is clearly more skewed than for the positive value. In other words, it is not only the case that [+PoS] statistically implies [+ DC],

but also that [-DC] implies [-PoS]. The latter dependency is shown by the significant correlation in Table 4.11b.

This type of result may be interpreted as a statistical basis to formulate a *strong* unidirectional implicational universal (Maslova 2003: 106; Dryer 2003: 111) of the following form:

> [+ PoS] ➔ [+ DC] *and* [-DC] ➔ [-PoS]
> Languages with a positive PoS value significantly more often have a positive DC value, and the tendency for languages with a positive PoS value to have a positive DC value is significantly stronger than the tendency for languages with a negative PoS value to have a positive DC value.

In sum, the method used in this study involves observing frequencies of co-occurrence of formal linguistic traits. These frequencies are used to detect different types of dependency relations between various pairs of formal parameters, related to PoS classes and DC constructions, respectively, in terms of the hypotheses formulated in the previous section.

4.5 Outlook

This chapter rounds off the first, theoretical part of the book. In the second part, the data will be presented. Chapter 5 presents the classification of PoS classes and PoS systems attested in the languages of the sample. Chapter 6 presents the typology of DC constructions in the sample languages. In Chapter 7, these two typological data sets are combined to test the hypotheses formulated in the present chapter. Chapter 8 provides a more descriptive discussion of the results obtained in Chapter 7, explaining them from a functionalist perspective. Finally, Chapter 9 presents the study's overall conclusions.