FYI: theory and typology of information packaging

Smit, N.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
So far, this dissertation has dealt with theoretical matters. The conceptual domain of information structure was divided in three separate parts in chapter 2, the structure of and interaction between which have been modelled in chapter 3. Chapter 4 introduced the theoretical framework used in this study, and chapter 5 proposed a number of improvements to the framework to make it better suited to deal with the five categories of information packaging (the informational articulations) under investigation. Now, we will turn to the empirical side of things starting with a number of methodological issues that are discussed in this chapter.

The structure of the present chapter is as follows. In the first section, a brief survey will be presented of some previous work that is relevant to the research reported in this study. Some challenges for a typological study in the domain of information structuring will be discussed in section 6.2. The main research question is introduced in section 6.3. In section 6.4, the design for this study will be introduced. The type of data that is taken into consideration is explained in section 6.5. Finally, the sample on which the research is based is discussed briefly in section 6.6.

### 6.1 Previous work

A crude division of typological studies in two classes can be made. One is that of morphosyntactic typological studies, that try to shed light on the distribution of a particular structural phenomenon or cohesive class of phenomena across languages (constituent order; morphological type). To account for these distributions in surface structure, underlying structure (semantic; pragmatics) and cognitive principles
from outside the realm of linguistic competence proper may be invoked. The other class is that of semantic typological studies, which are interested in a semantic category or class of categories (a dimension, such as person, number, possession, negation), and attempt to shed light on its distribution in surface structure.

What these two classes have in common is the subordinate role pragmatic distinctions play in each of them: if considered at all, it is usually invoked in an auxiliary capacity. Pragmatics in a sense is the ideal ‘escape hatch’, to the influence of which one can attribute virtually all remaining inconsistencies in whatever distribution of whatever linguistic phenomenon. Studies that concentrate on the surface structure distribution of pragmatic categories are relatively rare. Of those, studies that take a cross-linguistic perspective are even fewer. And among those, studies that concentrate on the typology of an entire dimension (for instance, information packaging) instead of a single category (focus) are virtually non-existent. Nevertheless, a number of such studies does exist. They will be briefly discussed below.¹

In an excellent early contribution to pragmatic typology, Dooley (1987) argues for three basic configurations of pragmatic structuring, similar to the informational articulations invoked in this study. Using data from three genetically unrelated Brazilian languages (Xavanté, Hixkaryana and Mbyá Guaraní), he shows that “languages adopt a basic configuration from a certain small set of possibilities, all of which have plausible explanations in terms of human cognition and discourse-pragmatics”. Dooley’s work is significant for two reasons. He is the first to suggest that informational categories enter into syntagmatic relationships (his configurations) and that these configurations are cross-linguistically relevant. He is also the first to explore the relationship between these informational configurations and global morphosyntactic properties that influence their distribution. For instance, he points to a possible relationship between S-medial constituent order and the lack of a clear Topic function. The research presented in the subsequent chapters is to a large extent inspired by Dooley’s work.

The paper by Lambrecht and Polinsky (1997) and its more systematic follow-up (Lambrecht 2001b) concentrate on the cross-linguistic realization of assertions with a categorical versus a thetic articulation. Both papers base their observations on a number of genetically unrelated languages. They are significant because they revive seminal work by Sasse on the distinction between assertions in which a relevance relation is predicated and those in which this is not the case, and the impact of this distinction on surface structure (Sasse 1991; Matras and Sasse 1995). Lambrecht and Polinsky furthermore introduce the notion of a paradigmatic contrast: the inventory of informational articulations constitutes a paradigm,

¹The survey in this section skips over the numerous language-specific studies of a single informational category and its possible subdivisions, such as Watters (1979); De Rijk (1978); Vismans (1997). Such studies are not pertinent to the research at hand, since they do not consider the interaction between informational categories in the assertion (a notable exception being Maslova 1997; Bellver and Michaelis 1999). In addition, most studies that I know of do not report contextualized examples, which makes it impossible to assess whether the analyses correspond to the notions used here.
the members of which are to be expressed in such a way that they are minimally distinct from one another. This principle allows for a straightforward formulation of hypotheses about the coding of said articulations. For example, what distinguishes a categorical from a thetic articulation is the presence versus the absence of a Topic layer. Consequently, in the view of Lambrecht and Polinsky, this allows for the prediction that thetic statements are expressed in such a way that the interpretation of any of their referents as Topic is blocked.

Another study that is highly relevant is that of Van Valin (1999), because it introduces the notion of a trade-off between morphosyntax (in his paper, freedom of constituent ordering) and information structure (the question whether or not particular informational functions are tied to particular linear positions in the clause). In his cross-linguistic comparison of the encoding of various focus constructions (again comparable to my informational articulations), Van Valin proposes a four-way typology of the interaction between morphosyntax and information structure, classifying languages as either rigid or flexible on either dimension. He also notes interesting correlations between his typology and established typological parameters such as constituent order, in particular the fact that V-medial languages tend to exclude positions on either side of the inflected verb from particular informational interpretations.

A final study that deserves mention is that of Matić (2003), which to my knowledge is the first to conduct a thorough quantitative study in the domain of pragmatic typology. For two reasons, his research is less pertinent to the present discussion. Matić is not primarily concerned with charting a pragmatic dimension, but rather shed light on the distribution of a specific morphosyntactic coding strategy. Moreover, the number of pragmatic categories that he takes into consideration is far greater and more diverse than the articulations considered in the present study, as categories from the domain of referent activation and episodic coherence are also included.

6.2 Challenges

The predominance of semantics as the preferred driving force behind typological distributions is hardly surprising. Typological studies in the domain of pragmatics are rare for a very good reason: the road towards them is strewn with practical and methodological complications. Excellent reviews of these problems can be found in various places in the literature, for example in Vallduví (1992). This section will highlight four of these challenges that have influenced the research design used in this study. These are related to the fundamental differences between the roles of semantics and pragmatics in language and language use; the consequent lack of systematicity in the encoding of pragmatics; the deeply rooted disagreement about the representation of pragmatic categories in theoretical linguistics, and related to that, the limited availability of cross-linguistic data of sufficient quality.
**Priority of semantics**  Most problems reside in the relationship between semantic, pragmatic and surface structure, and have to do with the fact that the tie between semantic structure and surface structure is intrinsically stronger, more direct and more salient than that between pragmatic structure and surface structure. Language users are keenly aware of the fact that they employ language to transfer representational meaning. Most of the categories in that domain are under their direct, conscious control, and manipulating them often has an immediate effect on the morphosyntactical structure of the message. Pragmatic meanings on the other hand are typically dependent on the presence of semantic structure. This is especially true for the informational categories in this study, which have no raison d’être if no representational structure is communicated. Finally, some of the social/interpersonal functions of language can typically also be fulfilled by non-verbal communicative behaviour. This is different for the representational function, to which language is much more indispensable: while communication about the here-and-now may rely on other channels besides language, the transfer of semantic structure – often relating to non-present or non-physical discourse referents – has no viable alternative. The exclusive dependence of representational transfer on language provides a plausible explanation why semantic structure seems to impact surface structure on such a larger scale than pragmatics does.

**Lack of systematicity**  Other factors that promote the predominance of semantically oriented typological research are related to this. As semantics claims center stage in the determination of surface structure, the expression of pragmatic categories by consequence is far less systematic: pragmatics in a sense has to ‘make do’ with what coding devices are left after semantics has finished shopping. As a result, things like informational categories tend not to occupy a particular locus in the morphosyntactic apparatus, within as well as across languages. This poses an obvious methodological problem for any attempt at informational typology. That is, it is virtually impossible to determine a priori the range of surface structure phenomena that must be taken into account and which ones can be omitted, and to determine what level of categorization captures all relevant distinctions while still being general enough to enable cross-linguistic comparison. On the contrary, the expression of information structure is pervasive, and tends to affect a large number of widely different aspects of morphosyntax and phonology. This is different from typological studies in the semantic domain: dimensions like person or possession typically affect only a limited part of surface structure. Which specific part that is may differ from one language to the next, but it seems fair to say that categories from a single functional dimension are generally expressed by means of a more or less coherent set of coding devices in surface structure; the expression of semantic structure is local and systematic.

---

2Hengeveld and Mackenzie (2008: 422ff) remark certain pragmatic categories – in particular, abstract illocutions – may affect the prosodic contour of the utterance directly, i.e. without interference from the Representational Level.
**Theoretical divergence** The limited systematicity in the coding of pragmatic categories provides a partial explanation for the lack of agreement in the theoretical literature as to what kinds of categories are necessary and sufficient to describe pragmatic domains such as information structure. A related matter concerns not the inventory of necessary and sufficient categories in a domain, but the relations between them: do all categories in it belong to a single dimension of information structure, and are they hence mutually exclusive? Are some categories hierarchically related to others as their sub- or supercategories? For instance, is topicality dependent on the state of activation of the corresponding referent in the discourse? Or are we really dealing with mutually independent dimensions? One particularly salient problem in this respect is the status of informational notions as universal or language-specific primitives, and the role of structural markedness. That is, individual languages tend not to reserve unambiguous structural coding devices for all categories in the informational domain, but typically choose to mark only a subset. For example, English basic SVO constituent order with two lexical arguments, as in *the statistician outsmarted the linguist*, is compatible with a number of different readings. In such cases, the question becomes relevant as to the status of structurally unmarked categories; are they irrelevant altogether for the language that does not mark them, or are they ‘lumped together’ in the expression with neighbouring categories? It will be clear that both views lead to different kinds of typologies with different kinds of explananda (cf. Bolkestein 1998; De Vries 1993). A ‘non-aprioristic’ account would ultimately need to explain why inventories of informational concepts are cross-linguistically different. On the other hand, a **universalist** account (which is the viewpoint defended in this study) needs to explain cross-linguistic differences in the neutralization of functional categories in surface structure instead.

**Data quality** The lack of a uniform, widely accepted theoretical framework in turn explains the uncertainty in descriptive work on individual languages as to which informational categories are to be expected, how they are related, and how they can be identified in surface structure encoding. Also, while a wide range of elicitation materials and techniques are available to document semantic categories, no such tools have yet been developed to describe informational distinctions. As a result, pragmatic categories are often not documented systematically and where they are, descriptive grammarians tend to use language-specific functional notions that are often not made sufficiently explicit. Furthermore, with regard to the non-apriorism – universalism distinction referred to earlier, it should be noted that most descriptive work is strongly biased towards the discussion of functional categories that are overtly marked, and seldom gives a complete description of those parts of functional categories that have no impact on surface structure in the language under consideration. This poses an additional problem for those who take a universalist perspective, since they cannot just content themselves with the information given in the grammar: the chance that this does not cover all relevant categories is significant.

The problems for language description and theory development in the field of
information structure are two sides of the same coin. The lack of reliable data continues to impede theory in making significant progress, while precisely the lack of a cross-linguistically usable theoretical-descriptive framework is responsible for this situation. The predictable result is a mushrooming of notions, terminologies and theoretical frameworks, some of which use the same terms with diametrically opposing meanings. For instance, while Centering Theory (Grosz et al. 1995) defines Topic as the most salient element of the clause (i.e., that which is at the forefront of the language user’s attention), many other theories use saliency as a key component of their definition of Focus instead. Still other informational categories are completely ideosyncratic (such as spectrum in Rising 1992), and are either designed to fit the description of a single language or language family or a particular theoretical framework. A practical complication that this confusion entails for typologists is that authors of descriptive grammars are often vague about which particular theoretical framework they have taken their terminology from, which makes it hard to assess whether the gloss used by the grammarian is still appropriate under the theoretical assumptions held by the typologist.

6.3 Research question

This study is primarily concerned with the structure of information packaging that is part of the underlying functional representation of the utterances that speakers use in verbal communication. The means of surface structure encoding that are used to convey informational categories are treated exclusively as a ‘window’ on this underlying structure. They are seen as instrumental, and are not investigated in their own right.

Articulations In chapter 5, some modifications were proposed to the way in which Functional Discourse Grammar treats information packaging. The aim of these modifications was to arrive at a representation in which information packaging instructions can be represented as primitives, referred to as informational articulations. They are inserted in the language user’s assertion in accordance with the Speaker-driven and Addressee-oriented information packaging considerations as present in the communicative intention. It has been stressed throughout the previous chapters that the informational articulations are not unitary primitives, but are configurational/syntagmatic constructs. This composite nature of the articulations derives immediately from their dual function in discourse knowledge management: addressee and actualization of the propositional structure of the interlocutor’s discourse knowledge. Although the approach proposed in chapter 5 allows for the definition of many more such informational configurations, five of these – referred to as core informational articulations below – are assumed to have universal status. Table 6.1 gives an overview of their names and the abbreviations used for them, as well as their formal representation in FDG.
Paradigmaticity in information packaging. The interaction between the dimensions of addressation and actualization endows the domain of information packaging with a paradigmatic structure. Three parameters describe the paradigm. The first of these, [predicationality], distinguishes between articulations that predicate relevance and those that do not. The second, [constitution], is concerned with the type of informational layers that an assertion consists of, and distinguishes articulations with and without a Topic and Comment Layer, respectively. The third, [locus of Focus], is concerned with the layer the Focus operator is attached to, which can be the Topic or Comment layer in the articulation, or a Subact of evocation.

The parameters are not fully independent. In particular, [constitution] and [locus of Focus] are not completely separable, in that assignment of a Focus operator to a Topic or a Comment layer entails the presence of such a layer in the constitution of the articulation. Also, for a single-layer articulation whose constitution is known, the locus of the Focus operator is trivial. Nevertheless, it appears that all three are required to describe the five articulations that are distinguished in this study. This is done in Table 6.2. Note that, with the exception of [predicationality], the parameters are not binary. [Constitution] is ternary, and is best split in two binary parameters [±Top] and [±Cm] for classificatory purposes. [Locus of Focus] is not binary either, as the locus of the Focus operator may be the Topic layer, the Comment layer or neither of these, as is the case with the Identificational articulation.3

3As has been noted before, the parameters in Table 6.2 allow for other, more complex articulations than the five discussed in this study. For example, it is conceivable that certain languages allow categorical assertions in which both the Topic and Comment layer bear a Focus marker, or languages that make a distinction between Identificational Focus as part of a categorical or a thetic articulation. Also, [predicationality] may have more values than the ones distinguished here, for instance to accommodate the pragmatic relation that involves a secondary Topic as proposed in Nikolaeva (2001).
(B and A) are dissimilar on two of the three parameters (constitution and locus of Focus), and similar on [predicationality]. The Address-central and Entry-central Categorical articulations (D and C) are dissimilar on [locus of Focus], but similar on [predicationality] and [constitution]. The Entry-central Thetic and Categorical articulations (B and C) are dissimilar on the parameter [predicationality], but similar on the parameter [locus of Focus]; the same is true for the Address-central Thetic and Categorical articulations (A and D). The dotted lines indicate that A, C and D are similar on [constitution], as well as B, C and D. Finally, E is – somewhat inconsistently – positioned in the centre of the figure, to reflect the fact that it remains agnostic to any particular value on the parameters [constitution] and [predicationality], and has its Focus operator on an evocational layer instead of an informational one.

**Table 6.2** Articulations decomposed as binary feature sets

<table>
<thead>
<tr>
<th></th>
<th>predicates relevance</th>
<th>constitution</th>
<th>locus of Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Topic</td>
<td>Comment</td>
</tr>
<tr>
<td>A</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>B</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>C</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>D</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E</td>
<td>±</td>
<td>±</td>
<td>±</td>
</tr>
</tbody>
</table>

**Figure 6.1** The paradigmatic structure of information packaging
Research question  The research question pursued in this study derives from what has been said above, and can be formulated as follows:

Main research question. How does the paradigmatic structure of the domain of information packaging affect the way that natural languages treat informational articulations in surface structure?

In the following chapters, more specific predictions will be formulated that are derived from this question.

6.4  Design

It was noted in passing in the previous section that this study employs a mapping design, meaning that it is interested in the correspondences between discrete surface structure categories and discrete categories from underlying structure. The question is how these functional categories, which form a conceptual space (Croft 2003: 134), ‘map onto’ the available surface structure categories. The data points in such a mapping design consist of information about the correspondence between points in the conceptual domain and means of surface structure encoding. The design proposed here can be seen as a member of the family of semantic map approaches (Haspelmath 2003).

6.4.1  Terminology

In order to avoid confusion further down the road, a number of basic notions and notational conventions will be highlighted in this section.

Articulation  Informational articulations are the functional units pertinent to the present study. They are syntagmatic constructs at the Interpersonal Level of Grammar, where they reflect the manipulations that a Speaker instructs an Addressee to perform on his discourse knowledge. Articulations are Speaker-driven, because they are in accordance with the Speaker’s communicative intention. They are also Addressee-oriented, in that Speakers’ instructions conform to the assumed state of knowledge of their interlocutors.

Informational articulation (I). A syntagmatic configuration at the Interpersonal Level of Grammar that conveys instructions to the Addressee concerning the tasks of addressation and actualization of discourse knowledge that the Speaker intends him to perform.

4The term mapping originates in early work of Chomsky, where it refers to the stages in the derivation at which syntax interacts with LF. Similar notions occur in other transformational frameworks, such as LFG (mapping theory, cf. Bresnan 2001), RRG (linking algorithm, cf. Van Valin 2005) and FG (expression rules, cf. Dik 1997a; Hengeveld and Mackenzie 2008).
Given their strongly interpersonal nature, it will prove useful to be able to
distinguish between intended and interpreted articulations. Articulations as in-
tended by the Speaker are referred to as $I_S$. Articulations as interpreted by the
Addressee are referred to as $I_A$. For example, $A_S$ refers to an Address-central
Thetic articulation intended by the Speaker, while $D_A$ refers to an Address-central
Categorical articulation as interpreted by the Addressee, et cetera.

Articulations are identified by means of external evidence (Croft 2003: 14ff),
namely on the basis of the context that surrounds them. That is, it is assumed that
the information packaging instructions that the Speaker imparts to the Addressee as
part of a Communicated Content $C_t$ can be inferred from the instructions imparted
in $C_{t-1}$ and $C_{t+1}$, whether the communicative exchange is monologic or dialogic.

**Coding strategy** At the level of surface structure, the constructions that are em-
ployed to convey the informational articulations will be referred to as coding
strategies. A coding strategy is defined as a unique surface structure pattern
(consisting of morphosyntactic and/or phonological features) that can consistently
be associated with a (cluster of) semantic and/or pragmatic properties. The notion
is kept extremely vague intentionally, so that it can easily be applied to categorize
surface structure phenomena of whatever kind, in whatever combination, at whate-
ver level of granularity. Its vagueness is not problematic for the research design at
hand, because the interest of the mapping design is not in the morphosyntactic
properties of the coding strategy.

**Coding strategy** ($S$). A unique language-specific surface structure
pattern, associated with a (cluster of) functional distinction(s) in a
consistent manner.

It is important to stress that coding strategies, unlike the informational articulations
they are used to express, are language-specific. This assumption follows from a rigid
interpretation of the view that universal syntactic categories do not exist, found both
in Functional Discourse Grammar (Hengeveld and Mackenzie 2008) and functional
typology. As Haspelmath (2007: 123) puts it, “not only are similar categories in
two languages never identical, but languages also often exhibit categories that
are not even particularly similar to categories in other languages.” An important
consequence of a strict interpretation of this non-apriorism is that no lower or
upper limit can be set a priori for the number of strategies that is necessary and
sufficient to describe the encoding of the informational domain in a given language.
Instead, such an inventory is identified for each language separately. In order to
do this, contrastive analysis of a large corpus of utterances is required, in which
the identification of a coding strategy is based on the fact that it is contrastively
distinct from alternative strategies available in the same Grammar. In this study,
the particular morphosyntactic properties and behaviour that cause this contrastive
distinctiveness are not at issue. The general abbreviation for coding strategies
will be italic capital $S$. Specific coding strategies are referred to by means of a
label mnemonic of their prototypical morphosyntactic features, again printed in
italics. These labels are nothing more than mere labels; no theoretical significance is implied by their names.\footnote{The contrastive analyses can only be appreciated if the relevant coding strategies of one language are considered collectively. For the languages of the sample, such analyses are provided in the appendix.}

It cannot be emphasized enough that the coding strategies distinguished in this study by no means give an exhaustive account of the morphosyntax of the languages in the sample. On the contrary, following the definition given above, only those strategies are taken into account where the morphosyntactic construct can be related to a (cluster of) informational articulation(s) in a consistent fashion. These can be seemingly haphazard selections and recombinations of syntactic categories described in the grammars of these languages, but that is not problematic: all that the coding strategies are used for in this study is to describe whether and in what combinations informational articulations may cluster in surface structure. The grammatical properties of the expression that these articulations take are beyond its scope.

**Coding potential** However vague the definition of coding strategies given in the previous section, it should be clear that the basis for their identification is the consistent association of form and function. This entails that coding strategies can be described in two ways: in terms of their morphosyntactic and/or phonological properties and behaviour, and in terms of the functional categories they are used to express. The latter aspect will be referred to as the **coding potential** of the coding strategy. It will be abbreviated as $CP_S$:

**Coding potential ($CP_S$).** The subset of informational articulations \( \{I\} \) that a given coding strategy $S$ may be used to express.

The coding potential is defined as the subset of articulations that the coding strategy may be used to express. Because we are dealing with the surface structure side of things, the articulations when used to define $CP_S$ are printed in italics. An example is given in (1):

\[
(1) \quad CP_{\text{postverbal}} = \{B, C, E\}
\]

‘The coding potential of the postverbal strategy consists of the Thetic, Categorical and Identificational articulations’

The notion of coding potential has a quantitative as well as a qualitative side, as both the number and type of articulation(s) that $S$ can be used to express may differ. Adopting terminology introduced in Hengeveld (1992), a coding strategy with a coding potential that comprises a single informational articulation will be called **differentiated**, whereas a strategy that can be used to express more than one articulation will be called **flexible**.

As every utterance in discourse is considered to have a single informational purpose and consequently will be represented as a single informational articulation at the Interpersonal Level of Grammar, it will be clear that each data point only
illustrates a single element of CPs. Whether or not S is a specialized or a flexible coding strategy can only be determined on the basis of multiple data points.

**Informational configuration, informatogram** In chapter 8, the necessity will arise to refer to the mapping between coding strategies and informational articulations of a language in its entirety, rather than the mapping per coding strategy in terms of its coding potential. This ‘full picture’ will be called the informational configuration of the language. The informational configuration can be represented graphically as an informatogram, of which an example is given in Figure 6.2. The circular nodes in the informatogram represent the informational articulations, abbreviated A thru E as explained in Table 6.1 above. The rectangular nodes represent the coding strategies, the number of which cannot be established a priori, but which may differ from one language to the next. Whether the S-nodes are printed above or below the I-nodes has no theoretical significance, but is an arbitrary choice made to preserve the readability of the figures. The edges between the nodes, which together reflect the coding potential of the S they connect to, represent the data points.\(^6\)

![Figure 6.2 Example informatogram](image)

### 6.4.2 Implementation

The data format used during the stage of collection follows the basic principles of the mapping design outlined above. Each data point consists of a contextualized example illustrating the association of a coding strategy and an informational articulation, whereby the identification of the latter is aided by the surrounding context.

Data points give partial information about the coding potential of the coding strategy involved. Consider the following example from Udihe. Udihe is a V-final language, but can express (Subject) referents in postverbal position under

\(^6\)In the appendix, the numbers on the edges correspond to the relevant example numbers. This information is not included in the informatograms in the chapters.
certain circumstances, one of which is the evacuation of such a referent from a position that is canonically associated with its interpretation as Topic. In (2), the unexpected event of ‘Crow’ diving into the kettle is presented in its entirety, and g’ai is expressed in postverbal position to prevent a topical interpretation of that referent. It can be thus determined from the surrounding context that the utterance has a Entry-central Thetic articulation (B).

(2) (the kettle began to boil, while Otter sat opposite of it) B
   joxo culi xuiq-ku-niee g’ai. pot through dive-pst-3sg crow
   ‘Crow dove into the kettle.’ (Nikolaeva and Tolskaya 2001: 899)

Example (2) illustrates that the Entry-central Thetic articulation belongs to the coding potential of postverbal referent expression in Udihe. However, the coding potential of this strategy is not limited to conveying this articulation. As the next example illustrates, it can also be employed to express an Entry-central Categorical articulation (C). In (3), ‘Otter’ is already established as the active Topic of the conversation, and the event of him putting the pot on the fire is appended to the knowledge that the Addressee possesses already.

(3) ('come over to my house, Crow, then I’ll make you dinner') C
   kawa do-lo-ni joxo-i toxolo-gi-e z’oto. house inside-loc-3sg pot-rfl put_on_fire-iter-pst.part otter
   ‘Inside the house, Otter put the pot on the fire.’ (Nikolaeva and Tolskaya 2001: 897)

Finally, the strategy is also used to convey the Identificational articulation (E). In (4), the intended actualization of the Addressee’s discourse knowledge is realized by instantiating the Agent slot of the (otherwise presupposed) event of X PUTTING THE POT ON THE FIRE with the (also presupposed) referent ‘Crow’. The utterance has an identificational articulation, a fact of which the presence of xaisi is a further indication.7

(4) (Otter came to Crow’s) E
   eme-mie joxo-i toxolo-gi-e-ni g’ai xaisi. come-inf pot-rfl put_on_fire-iter-pst-3sg crow too
   ‘When it [Otter] came, Crow also put the pot on the fire.’

7Particles like xaisi are considered Focus operators (König 1991), which unambiguously indicate that the elements in their scope constitute an actualization of presupposed knowledge. The examples clearly show that there are different motivations to exploit the postverbal slot. In (2), postverbal Subject expression is in a sense epiphenomenal because the main concern of the language user is to prevent a Topic interpretation of the postverbal referent. In (4), on the other hand, the prominence of the target position is exploited. Whether and how such motivations are related is not at issue in this study. The reason why (3) has a postverbal Subject is not clear: the slot seems to be associated with other kinds of prominence besides Focus (see also Nikolaeva and Tolskaya 2001: 844ff).
Setting the Stage

(Nikolaeva and Tolskaya 2001: 898)

On the basis of the evidence above, the coding potential of the Udihe strategy *Postverbal* can be said to range over the Categorical, Thetic and Identificational articulations. This is graphically represented in Figure 6.3.

![Figure 6.3](image1)

**Figure 6.3** Coding potential of the *Postverbal* strategy in Udihe

In the same way, the other coding strategies that are pertinent to the expression of informational articulations can be plotted as well. The resulting informatogram that reflects the entire informational configuration of Udihe is given in Figure 6.4. It shows that four coding strategies can be distinguished in Udihe that are relevant in the domain of information structure expression. The coding potential of the *Postverbal* strategy comprises the Entry-central Thetic (B) and Categorical (C) articulations, as well as the Identificational articulation (E); the coding potential of the *Preverbal* strategy ranges over the Address-central Thetic and Categorical articulations (A and D); the *S-initial* strategy can express Entry-central Categorical and Address-central Thetic articulations (C and A); and finally the *Default* strategy has a coding potential that comprises all articulations except the Address-central Categorical (D). Note again, that the morphosyntactic properties and behaviour of the coding strategies are not at issue, and that the labels used for the strategies are mere mnemonic references to much more complicated morphosyntactic processes.

![Figure 6.4](image2)

**Figure 6.4** Udihe informatogram
The informatogram in Figure 6.4 highlights a general property of the data set, namely the abundant presence of many-to-many relationships between forms and functions. It shows that Udihe has no fully specialized coding strategies; instead, each strategy can be used to convey more than one informational articulation. Moreover, each informational articulation can be expressed by means of more than one strategy. While the complete absence of differentiated coding strategies is rare – Udihe and Greenlandic Inuktitut are the only languages in the sample that have no differentiated means of expression for informational distinctions at all – multifunctionality and redundancy in the expression of information packaging appear widespread in the languages of the sample.

6.4.3 Universalization of coding strategies

In section 6.4.1, I stressed that coding strategies are necessarily defined at the level of individual languages, since no universal syntactic categories can be formulated. The lack of such universal categories makes it impossible to give cross-linguistically valid definitions of coding strategies in terms of their morphosyntactic behaviour. However, the research design proposed in this study provides an elegant and straightforward way to circumvent this problem and thus enable cross-linguistic comparison after all. That is, rather than characterising strategies in terms of their morphosyntactic behaviour, they can be defined and compared in terms of their coding potential.

The number of logically possible coding potentials can be established independently of particular languages. It equals the number of subsets of whatever size that can be drawn from the set of five informational articulations, whereby no articulation can occur more than once in each subset and the order of elements in each subset is irrelevant. This number is calculated as the sum of binomial coefficients $\sum_{n=1}^{4} \binom{5}{n}$, where $n$ is the number of articulations in the coding potential. This means that there are thirty logically possible coding potentials, which together are sufficient to give an exhaustive, universal characterization of all coding strategies in all languages.\(^8\)

The full list of coding potentials is given in Table 6.3 below. Since the definition of coding potentials makes reference to universal categories (informational articulations) rather than to language-specific ones (morphosyntactic categories), the characterization of coding strategies in terms of coding potentials enables cross-linguistic comparison between them. Simply put, two strategies from different languages that differ radically in terms of their morphosyntactic behaviour, but which can both be used to convey the Entry-central (A) and Address-central Categorical (D) articulations, for instance, can be said to be identical for the purposes of the present study.

\(^8\)Namely, $\sum_{n=1}^{4} \binom{5}{n} = \sum_{n=1}^{4} \frac{5!}{n!(5-n)!} = 30$. This excludes the ‘empty’ and ‘all-inclusive’ coding potentials, which can convey no articulation at all and all articulations, respectively. These two patterns are excluded because they would require that the entire array of coding devices in the Grammar is taken into account, irrespective of their relevance to information packaging.
Table 6.3 Logically possible coding potentials, grouped by size

<table>
<thead>
<tr>
<th>n = 1</th>
<th>n = 2</th>
<th>n = 3</th>
<th>n = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A AB BD</td>
<td>ABC ADE</td>
<td>ABCD</td>
<td></td>
</tr>
<tr>
<td>B AC BE</td>
<td>ABD BCD</td>
<td>ABCE</td>
<td></td>
</tr>
<tr>
<td>C AD CD</td>
<td>ABE BCE</td>
<td>ABDE</td>
<td></td>
</tr>
<tr>
<td>D AE CE</td>
<td>ACD BDE</td>
<td>ACDE</td>
<td></td>
</tr>
<tr>
<td>E BC DE</td>
<td>ACE CDE</td>
<td>BCDE</td>
<td></td>
</tr>
</tbody>
</table>

The cross-linguistically valid characterization of coding strategies in terms of their coding potential is illustrated in the adapted version of the Udihe informatogram in Figure 6.5. Similar informatograms of the other languages in the sample can be found in the appendix and as illustrations in the next chapters.

The advantage of this adaptation is that the informatograms in this shape are directly amenable to cross-linguistic comparison.⁹

6.5 Data

In order to identify the informational articulation associated with the assertion that a Speaker makes, this study relies on the context of the assertion under consideration. Based on the expression of the Discourse Acts surrounding Aₜ, an estimate can be made of the Addressee’s Mₜ, or at least of the Speaker’s

⁹A number of languages in the sample have multiple different coding strategies with identical coding potentials; see for instance the informatogram of Ma’di in the appendix, which features three coding strategies with coding potential B, three strategies with coding potential C and three strategies with coding potential E. Where this situation obtains, the coding strategies concerned are conflated to a single rectangle in the informatogram, which is connected to the articulation(s) at issue by means of more than one edge.
assumptions about $M^N_W$. This reliance on external evidence from context is crucial: if we were to identify the information packaging instructions imparted by the Speaker on the basis of the morphosyntactic shape of the expression, this would result in circularity.

Therefore, this study exclusively considers contextualized data. The remarks made earlier notwithstanding, recent grammatical descriptions often include a selection of transcripts of spoken discourse. Of these, a practical choice has been made to consider primarily narrative monologues. This has two advantages. First, narrative monologue is by far the best-documented genre cross-linguistically. Second, narrative monologue typically makes limited deictic reference to the physical world that surrounds the speech situation, and features a relatively high amount of non-phoric third-person reference. This makes the analyses somewhat easier, because phoric elements, in particular reference to first and second persons, tend to behave cross-linguistically different from independent phrases of reference: expression of the former tends to be conditioned more strongly by the morphosyntactic environment these elements occur in, whereas expression of the latter retains a certain amount of independence from such considerations. Another reason why the use of narrative discourse is preferable in the context of this research is that the amount of presuppositional structure that is transferred in such settings is much higher than in other communicative interactions, where the social functions of language often take precedence over its informative functions.

Most of the third-person narrative documented in grammars concerns storytelling, rather than the communication of ‘actual’ events that constitute new knowledge for the interlocutor. Obviously, an important function of story-telling is the strengthening of social ties in the community: many authors note that the stories they report are well-known in the community, and that the ability to tell them well comes with increased social status. Nevertheless, I believe that a distinction must be made between the social function of a genre and the social functions of a Discourse Act. Therefore, the consecutive assertions made in the context of story-telling will still be assumed to have primarily an informative function, notwithstanding the fact that most of interlocutors know quite well how the story will evolve. One could argue that their ‘actual’ knowledge is suspended in the context of the communicative exchange.

The inherent risk of limiting the investigation to third-person narrative discourse is that the results of this study cannot be interpreted as knowledge of the way in which languages deal with information packaging, but merely of the way in which languages deal with information packaging in narrative discourse. Therefore, one conclusion that can be drawn before we have even started is that this research should be repeated on the basis of a more diverse collection of transcripts. However, given the sporadic availability of transcribed discourse from other genres, it will be difficult to compile a sample with sufficient coverage.
6.6 Sample

It has been emphasized throughout this dissertation that the research it is based on is exploratory rather than confirmatory. The study does not aim to test pre-existing hypotheses concerning the distribution and encoding of informational categories, but is interested merely in exploring the domain of information packaging, identifying promising areas of future research and formulating hypotheses to be tested in confirmatory follow-up studies. The sampling of subject languages was done with these aims in mind. A sincere attempt was made to balance empirical and methodological requirements on the one hand with practical limitations on the other. The resulting fifteen-language sample is given in Table 6.4; an overview of its areal distribution is given in Figure 6.6. Despite its small size and dramatically incomplete genealogical coverage, I would still uphold that the sample is a variety sample, albeit hardly a very adequate one.

Variety sampling is widely considered the most appropriate technique when conducting exploratory research (Song 2001: 34) in which optimal coverage of all potential manifestations of a phenomenon prevails over the proportionate representation of previously established groups. Variety is usually operationalized in terms of genealogical and/or areal spread. Most methods of variety sampling require a minimal sample size equal to the number of superordinate phylogenetic groups in a genealogical classification of the researcher’s choice, which by definition includes all language isolates. One elegant method to decide on the inclusion of languages beyond that number is proposed by Bakker and Rijkhoff (1998), who provide an algorithm that determines the proportional representation of each phylogenetic group in the sample as a function of its internal complexity rather than its overall size. Depending on which genealogical classification is used, the minimal number of languages in a variety sample using this method is somewhere between twenty-nine (when using the genealogical classification proposed in Ruhlen 1987)\textsuperscript{10} and 252 (when using the genera proposed in Dryer 1992). Still, such minimal samples are considered to be a poor approximation of the probable diversity in the linguistic universe. An adequate sample in fact “can never be too large” (Bakker fc), and it seems that most typological studies nowadays are conducted on the basis of samples of fifty languages and up.

However, sample size is a function of the availability of data and research time. Compiling a variety sample large enough to withstand methodological criticism is a time-consuming task, especially if the necessary data is not readily accessible for the researcher. As was explained in section 6.5, the data used for this research consist entirely of contextualized utterances taken from running text. In addition, in order to assess the coding potential of individual coding strategies large quantities of data

\textsuperscript{10}That is, one representative of each of the nineteen superordinate phylogenetic groups, one representative from the ‘pseudo-phylum’ of contact languages, and nine language isolates. The size of any \textit{actual} minimal variety sample based on Ruhlen’s classification is twenty-three, because for six of his language isolates – Sumerian, Etruscan, Hurrian, Meroitic, Nahali and Gilyak – so little data is available that their inclusion is meaningful in only a handful of very specific studies.
<table>
<thead>
<tr>
<th>ISO</th>
<th>Language name</th>
<th>Genealogical classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ruhlen (1987)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lewis (2009)</td>
</tr>
<tr>
<td>mzp</td>
<td>Movima</td>
<td>Amerind</td>
</tr>
<tr>
<td>gyd</td>
<td>Kayardild</td>
<td>Australian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Australian &gt; Pama-Nyungan</td>
</tr>
<tr>
<td>laj</td>
<td>Lango</td>
<td>Nilo-Saharan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nilo-Saharan &gt; Eastern Sudanic</td>
</tr>
<tr>
<td>tvo</td>
<td>Tidore</td>
<td>Indo-Pacific</td>
</tr>
<tr>
<td></td>
<td></td>
<td>West-Papuan &gt; North-Halmahera</td>
</tr>
<tr>
<td>dbj</td>
<td>Begak</td>
<td>Austric</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Austronesian &gt; Malayo-Polynesian</td>
</tr>
<tr>
<td>kgo</td>
<td>Krongo</td>
<td>Niger-Kordofanian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nilo-Saharan</td>
</tr>
<tr>
<td>tae</td>
<td>Tariano</td>
<td>Amerind</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arawakan &gt; Maipuran</td>
</tr>
<tr>
<td>sci</td>
<td>Sri Lankan Creole Malay</td>
<td>Dravidian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Creole &gt; Malay-based</td>
</tr>
<tr>
<td>kal</td>
<td>Greenlandic Inuksitut</td>
<td>Eskimo-Aleut</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eskimo-Aleut &gt; Eskimo</td>
</tr>
<tr>
<td>ude</td>
<td>Udihe</td>
<td>Altaic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Altaic &gt; Tungus</td>
</tr>
<tr>
<td>scs</td>
<td>North Slave</td>
<td>Na-Dene</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nuclear Na-Dene</td>
</tr>
<tr>
<td>xbr</td>
<td>Kambera</td>
<td>Austric</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Austronesian &gt; Malayo-Polynesian</td>
</tr>
<tr>
<td>hix</td>
<td>Hixkaryána</td>
<td>Amerind</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Carib &gt; Southern Carib</td>
</tr>
<tr>
<td>mhi</td>
<td>Ma’di</td>
<td>Nilo-Saharan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nilo-Saharan &gt; Central Sudanic</td>
</tr>
<tr>
<td>lez</td>
<td>Lezgian</td>
<td>Caucasian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caucasian &gt; North-Caucasian</td>
</tr>
</tbody>
</table>

Notes. Sri Lankan Creole Malay has been classified according to its substrate (Tamil).

**Table 6.4** The sample
Figure 6.6 Geographical spread of the languages of the sample.
had to be surveyed per language, and large portions of the individual Grammars had to be studied in considerable detail in order to determine whether apparent instances of surface structure variation could really be ascribed to a pragmatic distinction, or were conditioned by semantics or other parts of morphosyntax instead. Given the time limits imposed on the research, it simply proved impossible to collect data on more than the fifteen languages currently in the sample.

Three criteria were adopted in the compilation of the present sample: data availability, genealogical diversity and ‘morphosyntactic interestingness’ of the languages concerned. These criteria determined the order in which the sample languages were selected from a larger ‘proto-sample’; the time available dictated that the inclusion of further languages should stop at the eventual fifteen. The resulting sample suffers from two very noticeable shortcomings: the over-representation of Ruhlen’s Amerind and Austric phylums, and especially the absence of a large number of other superordinate groups.

As regards the former problem, the inclusion of the three Amazonian Amerind languages at issue was motivated entirely by their respective ‘interestingness’: Movima was included on account of its peculiar interaction between word order and alignment; Hixkaryana because of its Subject-final canonical constituent order, and Tariana because of the **Topic advancement** construction and the apparent reliance of cross-reference marking on informational considerations. The expectation that, on account of their morphosyntactic traits, the inclusion of these languages in the sample would help to gain a better insight into the interaction between information packaging and surface structure prevailed over the ambition to include languages from as many different genealogical backgrounds as possible. In addition, it should be noted that the status of Amerind as a unitary group is heavily contested in the literature, and that alternative genealogies like the *Ethnologue* (Lewis 2009) classify Movima, Tariana and Hixkaryana as representative of three distinct superordinate groups. This last argument is not available to downplay the over-representation of the Austric phylum in the sample: two of the sample languages – Begak and Kambera – belong to the Malayo-Polynesian subgroup, as does the lexifier language of a third (Sri Lankan Malay). The latter is included as a representative of the group of contact languages; it can also be regarded a Dravidian language on account of the fact that Tamil is its most important substrate influence. As to the inclusion of both other languages, the interestingness of the Begak (which exhibits a system of **Focus markers** similar to Tagalog) is what motivated its inclusion (Kambera had already been included at that stage).

At this point, it should be remarked that diversity samples based on genealogical diversity suffer from a coverage problem of their own. As is observed in Croft (2003: 22), a sample that consists of languages that are maximally dissimilar genealogically may be likely to capture all **discrete types** in a typology, but is considerably less likely to give insight into the existence of **intermediate types**, if these are thought to arise as a consequence of diachronic change and language contact. Therefore, Croft argues, a study carried out on the basis of a diversity sample should ideally be complemented by a study of closely related languages to gain insight into the transition patterns that may exist between the discrete states.
in the typology. While the present sample does not begin to come close to being an adequate representative of the approach Croft advocates, the over-representation of the Amerind and Austric groups may be seen to address this concern.

The real problem with this sample is the severe under-representation of a large number of superordinate genealogical groups: eight of Ruhlen’s phyla are not included in the sample at all, and none of his language isolates is. For a number of absentees, data availability was the primary reason why they were not among the primary candidates for inclusion in the sample. This is particularly true for the isolates – Burushaski, Ket and Basque – for which transcribed discourse data is known to exist, but was difficult to get a hold of. Similar data problems were at hand for some of the absent phyla: Kartvelian, Chuckchi-Kamchatkan, Khoisan and Afro-Asiatic were excluded because no representative of these families could be found for which a sufficient amount of transcribed discourse data was readily available. This is especially frustrating for the Afro-Asiatic phylum, as it is well known from the literature that one of its branches (Cushitic) would definitely need to be included in this study on account of the ‘interestingness’ of its behaviour in the domain of (pragmatically conditioned) noun incorporation (cf. among others Svolacchia et al. 1995). For the other four absent families – Sino-Tibetan, Korean-Japanese, Indo-Hittite and Uralic-Yukaghir – ample data is available from several of their member languages that could have been used: these groups are excluded simply for reasons of time.11

In sum, parts of the linguistic universe may be too prominent in the sample, roughly half of the potential cross-linguistic variation that is genealogically induced is omitted from this study. Therefore, any cross-linguistic generalizations derived from the results reported here have to be treated with the greatest possible suspicion. Nevertheless, it should be recalled that the eventual aim of this study is to gather empirical support for a typology of information packaging, rather than to account for the distribution of classes of this typology across the world’s languages. As we will see, even the severely impaired sample that is used here provides the necessary support for the classes I set out to find. Therefore, even though there is an urgent need to repeat the present study on the basis of a better sample, the results that are obtained are not rendered invalid because of its lacking genealogical coverage.

11 Accidentally, it should be noted that languages from these four groups are among the best-studied cases of the influence of information structuring on morphosyntax: many influential theoretical insights have been inspired by work on Japanese, Chinese and the European languages. While this was not the ‘official’ reason why they are not among the primary candidates for inclusion in the sample, the fact remains that a lot of conflicting statements about information packaging are made for these languages, precisely because they are so well-documented.