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: http://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.
Epilogue

9.1 Introduction

The first chapters of this study have stated that the job of information packaging is to shape the expression of a piece of information in such a way that it is ‘properly dressed’ for the occasion, i.e. the communicative exchange in which it is invoked. What we found is that languages differ as to the scenarios of information packaging they care to distinguish, and that these differences seem to be systematic.

In chapter 5, a paradigm consisting of five such occasions – operationalized as informational articulations – has been identified. Some of these articulations form syntagmatic constructs, consisting of multiple units. All of them are two-dimensional, in that they unite categories from two orthogonal dimensions of information packaging: addressation and actualization. As is pointed out in chapter 6, they can be grouped along three parameters; predicationality, constitution and locus of Focus. Thetic articulations form non-predicational structures since they consist of a single informational unit, while categorical articulations consist of two units which engage in a predicational configuration. Topic-based articulations contain a Topic layer, while Comment-based articulations contain a Comment layer. Finally, articulations may differ as to the locus of the Focus operator. Address-based articulations place a Focus operator on the Topic layer, while Entry-based articulations place a Focus operator on the Comment layer.

In chapter 7 it was shown that the parameter-based groupings that were hypothesized in chapter 6 manifest themselves in the coding strategies that languages adopt to convey the informational articulations. That is, hardly any coding strategies are attested in the languages of the sample that are capable of conveying articulations in a combination that is not governed by one of the three parameters...
above. In chapter 8 the same overall tendency was shown to exist for the entire constellation of articulations and coding strategies per language: almost without exception, the quantified similarity of clusters of articulations governed by a parameter is greater than that of clusters of articulations for which this is not the case. Both the qualitative, per-strategy analysis and the quantitative analysis of the languages’ entire constellations of form-function correspondences in the domain of information packaging suggest the same thing: the parameters along which the domain of information packaging was operationalized are indeed relevant to the form-function interaction in a broad range of languages, and the idea that addressation and actualization are dimensions that should be considered in their own right yet in interaction with one another, is reflected in their grammars.

In this chapter, I will revisit some of the conclusions that were reached in this dissertation. Also, a number of prospects for further research will be discussed.

9.2 Conclusions

9.2.1 Theory

This study has presented a ‘mechanistic’ theory that relates information packaging in linguistic expressions to a structured representation of discourse knowledge that exists outside grammar: in Functional Discourse Grammar, this ‘module’ is placed in the Contextual Component. Speakers maintain a model of their interlocutors’ state of discourse knowledge, which they aim to manipulate by making assertions that contain information as well as a set of instructions to allocate the assertion and to actualize the Addressee’s state of discourse knowledge.

By characterising the nature of the linguistic categories Topic, Comment and Focus in terms of what they do rather than in terms of what they mean, much clearer definitions can be given that do not depend on ill-understood and hard-to-operationalize interpretive notions like aboutness (in the case of Topic) or importance or saliency (in the case of Focus). Also, rooting the linguistic categories that deal with discourse knowledge management in a model of discourse knowledge itself leads to a coherent picture of the relations between the various notions. It makes clear, for instance, why multiple Topics are unlikely to occur, whereas multiple Foci are quite possible.\(^1\) Finally, some sort of clearly defined extralinguistic motivation for informational categories is an essential requirement for a strictly function-oriented investigation as has been carried out in this study. In order to explore the structural manifestation of informational categories, one must be able to identify these independently from their expression. The theory that has been proposed here is a good candidate to do so, because it couples a clearly defined

\(^1\)Obviously, the supposed uniqueness of Topic is restricted to the functional category by that name. It may have a formal correlate that can occur more than once in the same utterance: in most cases, it will be possible to explain such distributions by the fact that the formal category at hand correlates to other functional categories besides Topic (such as Given) which can occur multiple times in the same utterance.
architecture of discourse knowledge with a deeply interpersonal understanding of communication. This provides all the ingredients that are necessary to identify the manifestation of informational categories on the basis of the context in which they occur rather than the morphosyntactic means that are used to express them.

The postulation of articulations encourages the study of informational categories in interaction with one another and their joint impact on surface structure, thereby improving FDG’s descriptive adequacy. For one, there is a crucial difference between a categorical assertion of which the Topic instruction is not instantiated by a Subact of Reference, and a thetic assertion in which simply no Topic instruction is imparted at all. This is a distinction that is reflected in the surface structure of many languages, but that can only be identified if informational categories are considered as parts of more complex structures. In this sense, the informational units are on a par with arguments in a predication, whose semantic functions can only be assessed when they are considered as part of the predication in which they occur. Just like a predication and its constituent elements, the interaction between the informational units contributes to the meaning of the complex structure they are part of.

Incorporating a frame-based approach to information packaging requires several changes to the treatment that these categories are given in Functional Discourse Grammar, where they are canonically modelled as pragmatic functions attached to Subacts of evocation at the Interpersonal Level. It has been argued that on grounds of theory-internal consistency and descriptive adequacy, this treatment ought to be abandoned. The proposed alternative is to model informational articulations by means of frames that are predefined in the Fund, parallel to what has become the standard treatment in FDG for predication, illocutions and other complex exocentric Layers. An important advantage of this modification is that typological variation in the domain of information packaging with these changes can be located easily in the Fund rather than in the combinatorial machinery of Grammar, which is the preferred solution in Functional (Discourse) Grammar.

9.2.2 Methodology

This study was set up as a strictly function-oriented investigation. Informational articulations were identified on the basis of preceding and following context, and not on the basis of their morphosyntactic manifestation. In this way, the expression of the entire domain of information packaging, operationalized in five basic articulations, can be subjected to a comprehensive survey for all languages in the sample, including categories whose typical mode of realization is non-expression (continued Topics) and categories which are not expressed by means of a dedicated coding strategy but only a functionally underspecified – or default – one.

For each articulation, a list was compiled of coding strategies that can be used to convey it. In this process, the language-specific morphosyntactic means that the strategy employs are disregarded. Instead, we concentrated on the coding potential of the strategy: the question which articulations can be conveyed by each of the strategies. Since coding strategies are often multifunctional and articulations can
often be conveyed by multiple strategies, the entire constellation of form-function correspondences for each language takes the shape of a complex network of many-to-many relationships. Except for the ‘aprioristic’ decision to only investigate the encoding of five articulations, no other a priori assumptions are made. In particular, no lower or upper limit has been formulated as to the number of relevant coding strategies that should be considered per language. This aspect of the methodology reflects the assumption that no utterance is void of information structure: consequently, coding strategies whose most obvious function is something else besides the conveyance of informational distinctions may nevertheless be informationally conditioned.

Haiman’s principle of iconicity dictates that recurrent similarity in form is indicative of proximity in meaning. This principle is central to the methodology proposed in this study, which is in essence a semantic maps approach. If two basic articulations in a language consistently make use of the same coding strategy, they can be said to be close to each other in the language user’s ‘cognitive space’ of information packaging. If they are consistently expressed by means of different coding strategies, they will be far apart. By plotting the articulations according to the extent to which they belong to a single coding potential, an informational map arises. This study has concentrated on the construction of such maps per language, with specific attention to a number of classificational aspects. The logical next step is to move away from mere classification towards ‘typology proper’, and subject the maps of a larger number of languages to cross-linguistic comparison. In such an exercise, two classes of issues will be at stake, the first of which is the robustness of the classification. The results of this study show that the restrictions on the possible orientation of articulations on the informational map can be described by means of the three parameters that define the paradigm of information packaging: predicationality, constitution and locus of Focus. It would be interesting to see whether this is also the case for an adequately representative sample and if so, whether the result would tell us something about the relative impact of each parameter. Second, this study has concentrated on the establishment of ‘informational type’. It has refrained from relating this characteristic of a language to other features, such as its genealogical or areal position, or any of the better-known morphosyntactic typological parameters such as constituent order, semantic alignment and morphological type. More comments on the topic will be made below.

**Quantifiable typology** It has been stated above that this study is classificational in that it aims to group the informational articulations into maps based on the total set of coding strategies, group these maps in terms of alignment patterns, and group languages in terms of the extent to which they conform to these alignment patterns. An important aspect of the various classificatory techniques that are proposed is their emphasis on quantitative rather than qualitative classification.

Quantitative classification offers a major advantage, which is its gradient nature. It is a truism to say that typological classes are never absolute: for a typological study of any phenomenon $x$, the data available for each object language will contain
token expressions that constitute counter-evidence against its classification. The usual solution to this problem is to look at tendencies in the available material. Simply put, a single counter-example in the face of a hundred confirming examples is less problematic than if the number of confirming examples is ten or less. However, this step relies on the representativeness of the available corpora, which is often questionable. That is, if phenomenon $x$ has a $1:100$ ratio of occurrence in the data, this finding cannot be safely extrapolated to the language in actual use. The ultimate risk of discrete classification based on a decision to neglect apparent counterexamples on account of a perceived tendency in imperfect data is that we pursue a typology of corpora rather than languages. Part of this problem is overcome if a classificatory method is devised that is capable of taking into account both confirming examples and counter-examples, and combining them to produce a somewhat balanced judgement. This is exactly what we need to do justice to the high dimensionality of the informational configurations, even though the data points in this study are coding potentials rather than linguistic expressions.

Given that an articulation may be in the coding potential of more than one coding strategy, a technique is needed to integrate the individual contributions that each coding strategy makes to the proximity of an $n$-tuple of articulations. To this end, a distance metric is proposed that capitalises on the idea that communication is a matter of coding and decoding. This insight is used to arrive at a matrix that expresses the probability of a Speaker’s intention to convey articulation $I_S$ being identified by the Addressee as articulation $I_A$. These probability distributions constitute sets which can be compared yielding a percentage of similarity that can be expressed as a value between .000 (no similarity) and 1.000 (full similarity). The calculation of the probabilities is ‘parametric’, which enables the researcher to include known skewings in the analysis (this aspect of the technique has not been exploited in this study).

The metric that is proposed is a fairly straightforward tool with a much wider potential domain of application. It could be used in any domain of linguistic typology where the researcher seeks to quantify the conceptual affinity between a set of underlying categories, but where sufficient data to use a canonical statistical technique to calculate similarity (such as Correspondence Analysis and related methods, or Multidimensional Scaling) is lacking. Moreover, the technique yields a quantified measure of similarity that has the same scale across languages, irrespective of the complexity of the network of form-function correspondences. Therefore, the results are valid both within and across languages, and very small differences in similarity may still be considered meaningful.

The language-specific informational maps can be classified in terms of the relative distances between the articulations, as is explained in section 8.4; the definitions of these alignment patterns are also gradient rather than absolute, as they depend on the quantified similarity between each pair of articulations. Finally, a simple measure is proposed to express the extent to which each alignment pattern is representative of the informational configuration of each object language. This last step is especially useful, because it makes it possible to characterize a language in terms of more than one pattern: it can be most representative of pattern $x$,
somewhat less representative of pattern y, not quite so representative of pattern z, et cetera. This is superior to discrete classification, where a language is put in a single class but much information is lost in the process.

In conclusion, this study hopefully has shown that typological classification by means of quantitative rather than qualitative criteria is not just reserved for those studies that have access to massive amounts of data. Even though such conditions are obviously preferable to small-scale exercises of the kind reported in chapters 7 and 8, it is a mistake to think that quantification is crucially reliant on the amount of input. On the contrary, even with limited data a successful attempt can still be made to merge conflicting qualitative observations into a single judgement, and to apply scalar rather than absolute classification.

9.3 Prospects for further research

The classificatory approach pursued in this dissertation has its limits. It convincingly shows that the paradigmatic structure of the informational domain is reflected in the way in which languages go about the encoding of informational distinctions in surface structure. It also illustrates that languages attach different importance to the various parameters that regulate information packaging: some languages make much of the distinction between thetic and categorical, whereas others care more for the distinction between focal and non-focal. However, the results of the classification do nothing to deepen our understanding of the interaction between form and function, which was the original aim of the overarching research project.

That said, I would like to argue that this study constitutes the necessary first step in such an endeavour. Without a clear picture of the expected band width of informational distinctions, an investigation into their interdependencies with surface structure is not possible. Therefore, the most important conclusion of this dissertation is the classifiability per se of information packaging systems, rather than the actual classification of the languages of the sample. Nevertheless, a number of prospects for future research are conceivable. They will be discussed below.

Role of semantics This study was interested in the impact of information packaging on surface structure. A moderately successful attempt has been made to keep the impact of semantic structure out of the equation, by taking into account only assertions that denote canonical events (Langacker 1999: 24ff): happenings that, representationally speaking, are remarkable to the least extent possible.

Even though precautions have been taken, the influence of semantics cannot be neutralized completely. It seems that semantics has a particularly important role to play in two areas. First, the acceptability of referents as the Topic of the assertion is partly dependent on semantic factors, as was argued in chapter 3. Likewise, certain semantic distinctions affect the possibility of imposing a thetic articulation onto the assertion, as is discussed in detail in Jäger (1997) for the distinction between Stage-Level and Individual-Level predications. It would be worthwhile to pursue both issues and investigate whether the interdependencies between informational
and representational structure can somehow be incorporated in the methodology proposed here.

**Cross-linguistic comparison**  It has been mentioned several times that this study is classificatory rather than typological. It shows that classification is possible and that the attested systems can be related to the paradigmatic structure of the informational domain, but it does not engage in a thorough cross-linguistic comparison of the various possible configurations. It would be very interesting to see whether the similarity between certain informational articulations is consistently greater than that between others cross-linguistically, and whether such distributions can be made sense of in terms of genealogical or areal relations between languages. Ideally, such a study would need to be conducted on the basis of larger data sets per language, so as to take into account the relative frequency of occurrence of the relevant coding strategies in the calculation of similarity. If a larger number of languages were considered, this could also help to gain better insight into the way in which features are ranked as proposed in section 8.5.

**Information packaging and morphosyntax**  The most interesting topics for further research concern the relation between information packaging and morphosyntax. In particular, it would be worthwhile to see how known typological parameters correlate with the relative impact of predicationality, constitution and locus of Focus on the informational configuration.

A number of such parameters may prove especially relevant. One of these may be the position of the verb in the basic constituent order. That is, it has often been noted that the inflected verb functions as an important boundary in the clause which for instance is exploited to mark off the Topic from the Comment. For languages with Verb-initial basic constituent order, it could then be expected that unambiguous marking of the Topical status of a referent is something which requires a special operation. Since the presence of a Topical referent is a property that is shared between Address-central Thetic and Categorical articulations, it may be the case that Verb-initiality correlates with $[\text{AD BC}]$ alignment. Conversely, languages that have an S-initial basic constituent order may need dedicated operations to ‘detopify’ the initial Subject. The result may be a strong correlation of S-initiality with $[\text{B ACD}]$ alignment. A similar influence may come from the syntactic property of cross-reference marking. It is conceivable that languages which employ cross-reference marking care less for the Focality of the Subject/Topic, since it is doubly evoked anyway. Hence, it is plausible to expect a correlation between cross-reference marking and $[\text{AB CD}]$ marking.

**Controlled experiments are needed**  One thing is needed more than anything else, however: controlled experiment. That is, in the absolute ideal situation, the mapping approach proposed for this study is implemented by changing the conversational context of a very limited controlled set of semantic structures in such a way that they match the informational articulations, and then see how this affects
the surface structure of the expression. Various experiments are conceivable. The easiest one to carry out is probably to have a non-verbal stimulus like the Pear Story or Frog Story that provides contexts that warrant the use of all five articulations and let a large number of speakers select the expressions that are contextually most appropriate to convey particular chunks of the story. If the story has sufficient length and the number of test subjects is sufficiently large, this should give a fairly reliable impression of the coding potential of that language’s coding strategies in the domain of information packaging. A less ‘aprioristic’ approach, however, would be not to constrain the response, but to place the same event in different contexts and see how speaker care to express it. In order for this approach to be sufficiently reliable, however, a massive number of test subjects is necessary.

As an intermediate solution, reliability of the analyses would improve greatly if it were carried out by experts in the various languages. If this study has illustrated anything, it should be the fact that a consistent approach towards information packaging is needed. If grammars discussed information packaging in terms of articulations, and if the informational coding potential of coding strategies were surveyed, our understanding of the relation between information structure and surface structure would much improve.