UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

# Transparency in language: a typological study

Leufkens, S.C.

**Publication date**
2015
**Document Version**
Final published version

Link to publication

**Citation for published version (APA):**
Leufkens, S. C. (2015). *Transparency in language: a typological study*. [Thesis, fully internal, Universiteit van Amsterdam]. LOT.

**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, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)*

Download date:11 Oct 2026

Sterre Leufkens

# Transparency in language

## A typological study

In this thesis I investigate the degree of transparency of 22 languages. Transparency is defined as the extent to which a language maintains one-to-one relations between units at different levels of organisation, i.e. pragmatics, semantics, morphosyntax and phonology. Languages are compared not only as regards the amount of non-transparent phenomena in their grammars, so as to rank them from relatively transparent to relatively non-transparent, but also to uncover an implicational pattern in the distribution of non-transparent features across languages.

Transparency is discussed in connection to the notions of simplicity and learnability. Simplicity and transparency are shown to be different concepts that are both relevant in accounting for acquisition data. Whereas most definitions of simplicity take it to apply to particular domains of grammar, transparency is defined as a property of interfaces between levels within the grammar. The term is further operationalised using the framework of Functional Discourse Grammar.

It turns out that all languages have at least some non-transparent features, most notably of the redundancy type. Fusion and domain disintegration are less commonly attested, but the non-transparent features that are found only in the least transparent languages are so-called form-based forms: highly syntacticised forms and structures, that have no pragmatic or semantic motivation. Furthermore, non-transparent relations at the interfaces of the phonological and pragmatic levels are found in many languages, whereas it is less common to violate transparency at morphosyntactic and semantic interfaces. The attested implicational hierarchy of transparency proves that transparency is a relevant typological notion.

: LOT
Netherlands Graduate School of Linguistics

ISBN 978-94-6093-162-8

ACLC

: LOT 380

Sterre Leufkens

Transparency in language

Sterre Leufkens

# Transparency in language

## A typological study

ACLC

: LOT
Netherlands Graduate School of Linguistics

Landelijke Onderzoekschool Taalwetenschap

# Transparency in language

## A typological study

Published by

LOT

Trans 10

3512 JK Utrecht

The Netherlands

phone: +31 30 253 6111

e-mail: lot@uu.nl

http://www.lotschool.nl

Cover illustration © 2011: Sanne Leufkens – image from the performance ‘Celebration’

ISBN: 978-94-6093-162-8

NUR 616

Copyright © 2015: Sterre Leufkens. All rights reserved.

# Transparency in language

## A typological study

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. D.C. van den Boom

ten overstaan van een door het college voor promoties ingestelde

commissie, in het openbaar te verdedigen in de Agnietenkapel

op vrijdag 23 januari 2015, te 10.00 uur

door Sterre Cécile Leufkens

geboren te Delft

Promotiecommissie

| | |
|---|---|
| Promotor: | Prof. dr. P.C. Hengeveld |
| Copromotor: | Dr. N.S.H. Smith |
| Overige leden: | Prof. dr. E.O. Aboh |
| | Dr. J. Audring |
| | Prof. dr. Ö. Dahl |
| | Prof. dr. M.E. Keizer |
| | Prof. dr. F.P. Weerman |

Faculteit der Geesteswetenschappen

# Acknowledgments

When I speak about my PhD project, it appears to cover a time-span of four years, in which I performed a number of actions that resulted in this book. In fact, the limits of the project are not so clear. It started when I first heard about linguistics, and it will end when we all stop thinking about transparency, which hopefully will not be the case any time soon. Moreover, even though I might have spent most time and effort to 'complete' this project, it is definitely not just my work. Many people have contributed directly or indirectly, by thinking about transparency, or thinking about me. I am very grateful to all of them, including everyone I do not thank in person below. It's true, I could not have done this without you!

First of all, I want to thank the Netherlands Organisation for Scientific Research (NWO), for generously funding my PhD project, also called research programme 322-70-003.

Linguistic research, especially typological research, relies heavily on the knowledge of native speakers and language experts. Without it, this dissertation would be built on sand. Therefore, I want to thank Enoch Aboh, Daniël Boeke, Leston Buell, Michael Dunn, Helen Eaton, Nick Evans, David Foris, Michael Fortescue, Mona Hegazy, Anne-Christie Hellenthal, George Hewitt, Shoichi Iwasaki, Zaira Khalilova, Mohammed Jafar, Baris Kabak, Marian Klamer, Elena Maslova, René van Munster, Sebastian Nordhoff, Çisem Özkurt, John Peterson, Caroline Roset, Ian Smith, Miriam van Staden, and Manfred Woidich for their answers to all my questions. I hope I have done justice to your languages.

My study also strongly builds on the work of linguists whose expertise lies in the fields of typology, theoretical linguistics, or linguistic complexity. In this respect, I am grateful to Paul Boersma, Suzanne Aalberse, and Marc van Oostendorp for sharing their thoughts, and thinking with me to solve our shared questions. For the same reason, I want to thank Jenny Audring, Fred Weerman, Enoch Aboh, Evelien Keizer, and Östen Dahl. I additionally thank them for agreeing to be in my committee. I have learned a lot from all of you.

Lachlan Mackenzie and Eva van Lier were very kind to read my texts and give their constructive feedback, which was helpful, inspirational and reassuring. Those qualifications equally apply to my time in Leipzig, thanks to Martin Haspelmath and Susanne Michaelis, who gave me, apart from new insights into my research topic, a very warm welcome.

After having thanked colleagues for their linguistic contributions, I also want to thank them for their personal support and friendship. Many, many people from inside and outside the Bungehuis have helped me through the more difficult days of this project. The Functional Discourse Grammar community has helped me tremendously in becoming a true linguist and showing that there is more to that than competition. Thanks for this, Kees, Lachlan, Hella, Evelien, Wim, Daniel, Lucia, Freek, Arok, Marize, Gareth, Joceli, and Flavia. Furthermore, I am thankful to Anne, Akke, Angela, Anne Maren, Roland, and Rosanne from the board of the Alumnikring Taalwetenschap, for our effective collaboration and the fun meetings (well, dinners). I am grateful to the Unlearnable and Learnable Language group as well, for all the instructive and inspiring talks and discussions, and for providing the opportunity to present my own work and gain organisational experience.

I have spent the last ten years studying and working in the Bungehuis and I am very sorry to leave it, as it feels like a second home to me. My gratitude goes to all Amsterdam colleagues, fellow PhDs, and especially roomies, who have made it feel this way. Thank you, Rob, Jan Hulstijn, Jan de Jong, Ingrid, Cecilia, Gerdien, Elly, Katja Chládková, Karin, Titia, Lucia, Lissan, Elin, Konrad, Sophie, Iris, Bibi, Margot, Jan-Willem, Margreet, Joke, Marlou, Katja Bobyleva, Mark, Vadim, Tessa, and all others. I would have been lost without your understanding, kind words, coffee, and especially your jokes.

Obviously, one of the most determining factors in everyday PhD life is supervision, and I am happy that my supervisors took my work seriously and were always helpful, attentive and kind. I am grateful to Norval Smith for pointing out shortcomings in things out of my expertise, like English and creoles – always with great detail and an even greater sense of humour. Norval, thank you for all your helpful feedback,

always on point, even from Vienna. Moreover, I feel blessed that Kees was always just a door away, ready to answer small questions, but also bigger ones. From the beginning of our collaboration, when I was a student, I have always felt that I went into Kees' office with a problem, and came out of it relieved and full of inspiration. Thank you for that, Kees.

Finally, my biggest, warmest and soggiest thanks go to the people who were there for me outside office hours, no matter what. Thanks to all the members of the unsurpassed Ricciotti Ensemble, for making so many people happy, and bringing out the best in me. Sara de Vi, Kim, Lisa, Krzysztof, Rosa, Sofie, Marten, Floortje, Maurice, Sara de Vr, Jildou, Emmie, I am very proud and grateful to have you as my fabulous best friends. Klaas and Sanne, my sweet nymphs, you are the funniest people I know and that has brought much comfort in difficult times. Papa, mama, Sanne, and Eva, I can't thank you enough. Ik ben zo blij dat jullie er zijn.

Enough with the cuddling, now enjoy my book!

# Table of contents


Acknowledgments i
Table of contents v
List of figures and tables ix
Glossing conventions and list of abbreviations xi

1 Introduction 1
1.1 Topic and aims 1
1.2 Theoretical embedding 2
1.3 Methodology 3
1.4 Outline 3

2 Transparency in Functional Discourse Grammar 5
2.1 Functional Discourse Grammar 5
2.2 Defining transparency in FDG 12
2.3 Non-transparency in the lexicon 15
2.4 Categories of opacity 16

3 Transparency 21
3.1 Other interpretations of transparency: delimitation of the concept 21
3.1.1 Counterbleeding and counterfeeding 21
3.1.2 Iconicity 22
3.1.3 Homomorphism and isomorphism 24
3.1.4 Transparency of compounds and derivations 25
3.1.5 Simplicity 26
3.2 Earlier studies on one-to-one correspondence 26
3.2.1 Theoretical linguistics 27
3.2.2 Language acquisition 30
3.2.3 Creole studies 31

3.3 Simplicity 35
3.3.1 Absolute simplicity 36
3.3.2 Relative simplicity 40
3.4 Directionality: a continuum of transparency 42
3.4.1 Implicational hierarchy: typology of transparency 42
3.4.2 Diachrony: the transparency of creoles 43
3.4.3 Learnability: the difficulty of opacity 46

4 A list of non-transparent features 49
4.1 Redundancy 50
4.1.1 Multiple expressions of pragmatic information 50
4.1.2 Nominal apposition 50
4.1.3 Clausal agreement or cross-reference 51
4.1.4 Phrasal agreement 55
4.1.5 Plural concord in noun phrases containing a numeral 56
4.1.6 Negative concord 57
4.1.7 Modal concord 60
4.1.8 Temporal concord and tense copying 61
4.1.9 Spatial concord 62
4.1.10 Summary redundancy features 62
4.2 Discontinuity 63
4.2.1 Extraction and/or extraposition 63
4.2.2 Raising 65
4.2.3 Circumfixes 66
4.2.4 Infixes 66
4.2.5 Non-parallel alignment 67
4.2.6 Summary discontinuity features 67
4.3 Fusion 68
4.3.1 Cumulation of TAME and case 68
4.3.2 Morphologically conditioned stem alternation: suppletion 71

4.3.3 Morphologically conditioned affix alternation: irregular stem formation 72
4.3.4 Summary fusion features 74
4.4 Form-based form 75
4.4.1 Grammatical gender 75
4.4.2 Nominal expletives 76
4.4.3 Syntactic functions 78
4.4.4 Influence of complexity on word order 84
4.4.5 Function marking is predominantly head-marking 87
4.4.6 Morphophonologically conditioned stem alternation 89
4.4.7 Morphologically and/or morphophonologically conditioned affix alternation 89
4.4.8 Phonologically conditioned stem alternation 92
4.4.9 Phonologically conditioned affix alternation 94
4.4.10 Summary form-based form features 94
4.5 Summary of the list of non-transparent features 94

5 Methodology 99
5.1 Research questions 99
5.2 The sample 102
5.3 Methodology: implicational hierarchies 108
5.4 Hypotheses and expected outcomes 111

6 Results and discussion 117
6.1 How are non-transparent features distributed cross-linguistically? 117
6.1.1 An overview of the data 118
6.1.2 An implicational hierarchy of transparency 126
6.1.3 Explanations for transparency and opacity 131
6.1.4 Features not fitting the hierarchy 135

6.2 How are redundancy, fusion, domain disintegration and form-based form features distributed cross-linguistically? 140
6.2.1 An implicational hierarchy between categories of opacity 140
6.2.2 Implicational hierarchies within categories of opacity 145
6.3 How are non-transparent features at different interfaces distributed cross-linguistically? 152

7 Conclusions 157

References 165
In-text references 165
Descriptions of sample languages 176

Summary in English 181
Samenvatting in het Nederlands (Summary in Dutch) 189
Curriculum Vitae 197

# List of figures and tables


| | | |
|---|---|---|
| Figure 2.1 | The general architecture of FDG | 5 |
| Figure 2.2 | The general architecture of FDG, including levels of representation and primitives | 6 |
| Figure 2.3 | The layers of the Interpersonal Level | 8 |
| Figure 2.4 | The layers of the Representational Level | 9 |
| Table 4.1 | Latin declension | 91 |
| Table 4.2 | Overview of non-transparent features and their possible values | 95 |
| Table 5.1 | Distribution languages over phyla according to Diversity Value | 105 |
| Table 5.2 | Languages in sample with Ruhlen and Ethnologue classifications | 107 |
| Table 6.1 | Basic data: Distribution of non-transparent (sub)features over sample languages | 120 |
| Table 6.2 | Basic data: Distribution of non-transparent features and combinatorial features over sample languages | 124 |
| Table 6.3 | Features showing an implicational relationship | 128 |
| Table 6.4 | Features not participating in an implicational hierarchy | 139 |
| Table 6.5 | Distribution of non-transparent features in terms of category of opacity | 142 |
| Table 6.6 | Distribution of redundancy features | 147 |
| Table 6.7 | Distribution of fusion features | 149 |
| Table 6.8 | Distribution of form-based form features | 151 |
| Table 6.9 | Distribution of non-transparent features in terms of interface | 154 |
| Figure 7.1 | Implicational hierarchy of transparency in terms of category | 159 |
| Figure 7.2 | Implicational hierarchy of transparency in terms of interface | 160 |

x

# Glossing conventions and list of abbreviations

In this dissertation, all language examples are glossed as much as possible following the Leipzig Glossing Rules, available online at http://www.eva.mpg.de/lingua/resources/glossing-rules.php. However, examples copied from sources by other authors sometimes contain abbreviations that are not dealt with by those rules, or are inconsistent with them. In such cases, I have tried to stay as true as possible to the author's judgments, in order to do justice to the language under consideration. Only in some cases, I have altered the original transcription or abbreviations in the gloss in order to maintain consistency or readability.

The first line of examples provides an exact representation of (part of) a sentence in the language under consideration. This line is meant to be true to the language, rather than to orthographical rules, and therefore I use no capitals or punctuation in that line, except for commas indicating intonation breaks, question marks indicating interrogative illocution, and exclamation marks for imperative illocution. The transcription may be inconsistent across the different language analyses, because some authors provide examples in IPA, while others use an existing or newly developed orthography. Since synchronising these orthographies is unfeasible within a typological study like this one, I have adopted the orthography provided by the author. This may lead to said inconsistencies, but to my knowledge, this never influences the interpretation of the examples.

The second line of examples contains the interlinear morpheme translation, in which lexical items are given in standard roman symbols and grammatical items in small capitals. Grammatical labels are abbreviated according to the standard list of abbreviations of the Leipzig glossing rules. However, several languages require different abbreviations, that are sometimes uniquely used for that language and hence not found in any standard list. If this is the case, I have indicated the language for which the label is relevant between brackets behind the abbreviation, together with its source, so that the reader can easily track the exact denotation of the label.

Abbreviations are used only if they are relevant for an analysis, or for reasons of consistency. In other cases, a direct translation into lexical elements serves as well, e.g. pronouns are sometimes glossed as '1SG' but in other cases as 'I'. Proper names are indicated in the gloss line by the first letter of their English equivalent only, following common practice in glossing. Epenthetic elements are copied to the gloss, printed in italics, and if possible attached to the morpheme to which they belong phonotactically. If it could not be determined to which morpheme an epenthetic element should attach, it is represented as a separate entity.

Some of the sources used for Georgian and Tamil did not provide morpheme-by-morpheme glosses. For those languages, I have tried to reconstruct such glosses myself, with the help of native speaker experts. Some of their reconstructions involved changes in the first line of the examples, which I have indicated by means of square brackets. I take full responsibility for remaining errors and inconsistencies in the glosses.

The third line of each example provides a free translation in English. If relevant, a more literal translation is provided as well, which is then given between double quotation marks. Again, I have followed authors' translations, unless there was a good reason to do otherwise.

**Abbreviations in glosses**

| | |
|---|---|
| I, II, III, IV, V | semantic noun classes |
| 1 | first person |
| 2 | second person |
| 3 | third person |
| A | semantic function Actor |
| ABL | ablative |
| ABS | absolutive |
| ACC | accusative |
| ACT | active voice |
| ADJ | adjective |

| | |
|---|---|
| ADV | adverbial case |
| AFF | affirmative |
| ALL | allative |
| ANAPH | anaphoric pronoun |
| ANDT | andative (Sochiapan, Foris 2000) |
| ANIM | animate |
| ANTIP | antipassive |
| AOR | aorist |
| APPL | applicative |
| ASSOC | associative case |
| ATTR | attributive |
| AUG | augmentative |
| AUX | auxiliary |
| BEN | benefactive |
| CAUS | causative |
| CLF | classifier |
| CNTR | contrastive focus |
| COLL | collective |
| COMM | common gender |
| COMP | complementiser |
| COMPC | complementising case (Kayardild, Evans 2003) |
| COND | conditional |
| CONJ | conjunction |
| COORD | coordinator |
| COP | copula |
| COUNT | counter (Bantawa, Doornenbal 2009) |
| CVB | converb |
| DAT | dative |
| DEF | definite |
| DEM | demonstrative |
| DES | desiderative |

| | |
|---|---|
| DET | determiner |
| DIM | diminutive |
| DIR | directional |
| DS | different subject |
| DU | dual |
| EQU | equative case |
| ERG | ergative |
| EXCL | exclusive |
| EVID | evidential |
| F | feminine |
| FAC | factitive |
| FAM | familiar |
| FOC | focus |
| FUT | future |
| GEN | genitive |
| GENER | generic |
| GENR | general TAM particle (Samoan, Mosel & Hovdhaugen 1992) |
| HAB | habitual |
| HON | honorific |
| HORT | hortative |
| HUM | human |
| IMMED | immediate (Kayardild, Evans 2003) |
| IMP | imperative |
| INABIL | inability |
| INAN | inanimate |
| INCH | inchoative |
| INCL | inclusive |
| IND | indicative |
| INDEF | indefinite |
| INF | infinitive |
| INFER | inferential |

| | |
|---|---|
| INGR | ingressive |
| INSTR | instrumental |
| INTR | intransitive |
| INTS | intensifier |
| IOBJ | indirect object |
| IPFV | imperfective |
| IRR | irrealis |
| ITER | iterative |
| L | semantic function Locative |
| LAT | lative |
| LINK | linking particle |
| LOC | locative case |
| LV | locative version (Georgian, Hewitt 1995) |
| M | masculine |
| MID | middle voice |
| MODC | modal case (Kayardild, Evans 2003) |
| N | neuter |
| NEG | negative |
| NFUT | non-future |
| NHUM | non-human |
| NMLZ | nominaliser |
| NOM | nominative |
| NPST | non-past |
| NSPEC | non-specific |
| NV | neutral version (Georgian, Hewitt 1995) |
| OBJ | object |
| OBL | oblique |
| OPT | optative |
| PASS | passive |
| PFV | perfective |
| PL | plural |

| | |
|---|---|
| PLACE | place name, derivational suffix |
| POL | polite |
| POLQ | polar question |
| POSS | possessive |
| POT | potential |
| PRED | predicative |
| PRIV | privative case |
| PRN | pronominal marker |
| PRV | pre-radical vowel (Georgian, Wier 2011) |
| PROG | progressive |
| PROP | proprietive case |
| PROX | proximal |
| PRS | present tense |
| PST | past tense |
| PTCP | participle |
| PURP | purposive |
| Q | question particle |
| QUOT | quotative |
| RDP | reduplication |
| RECP | reciprocal |
| REAL | realis |
| REFL | reflexive |
| REL | relativiser |
| SBJ | subject |
| SBJV | subjunctive |
| SBJ/V | subject-verb relation (Sandawe, Steeman 2012) |
| SG | singular |
| SIM | simultaneous |
| SPEC | specific |
| SS | same subject |
| STI | indirect stance (Sheko, Hellenthal 2010) |

| | |
|---|---|
| SUB | subessive |
| THEMSUF | thematic suffix |
| TOP | topic |
| TR | transitive |
| TRNSF | transformative (Kolyma Yukaghir, Maslova 2003) |
| U | semantic function Undergoer |
| UW | unwitnessed |
| v̀, v, v̄, v́ | tone levels (lowest to highest) |
| v̋, v̏ | rising, falling tone |
| VBLZ | verbaliser |
| VEN | venitive |
| W | witnessed |

**FDG terminology**

| | |
|---|---|
| Cl | clause (ML) |
| e | State-of-Affairs (RL) |
| ep | Episode (RL) |
| $f^c$ | configurational property (RL) |
| $f^l$ | lexical property (RL) |
| IL | Interpersonal Level |
| ML | Morphosyntactic Level |
| p | Proposition (RL) |
| PL | Phonological Level |
| R | Referential Subact (IL) |
| RL | Representational Level |
| T | Ascriptive Subact (IL) |
| x | Individual (RL) |
| Xp | Phrase of type x, e.g. Np, noun phrase (ML) |

Zo wordt de onregelmatigheid, onredelijkheid, onzinnigheid [van uitzonderingen in taal] vaak goedgepraat, verborgen, weggeduwd. De taal zelf lokt dat soort reacties uit, door op bedriegelijke wijze ons eerst een regelmaat voor te spiegelen en als die regelmaat eenmaal bij ons postgevat heeft die regelmaat te doorbreken.

(Uit: Karel van 't Reve, 'Reves vermoeden'.
*Een keuze uit eigen werk* (1991), p. 129)

# Chapter 1
# Introduction

## 1.1 Topic and aims

Languages map meaning to form. A speaker expresses what she wants to say by putting her message into signed or spoken forms, which can then be translated into meaning by the hearer. Taking an information-theoretical perspective, a one-to-one mapping would be expected, that is, a consistent mapping of one meaning to one form. In natural languages, however, such transparency is continuously violated, since mismatches, redundancy and reduced forms abide. Additionally, forms and rules appear that have no semantic or pragmatic motivation. Transparency is not only preferable from an information-theoretical point of view, but also from the perspective of a language learner. There is evidence that transparent structures are easier to learn (for L1- and L2-learners) than non-transparent ones. Why, then, does opacity exist? Why is natural language not a logical and straightforwardly learnable code?

This question is highly relevant for theoretical linguistics, as it touches upon the topic of autonomous syntax: does all linguistic form follow from its function to express meaning, or is language a system on its own that can or perhaps should be studied separately from its function? This is an unresolved debate that I intend to contribute to with this dissertation. Previous studies of transparency in different languages (cf. Hengeveld 2011) have pointed out that languages differ in their degree of transparency. Languages emerging from language contact have shown to be relatively transparent (Leufkens 2013a), while older languages show more opaque features (the words non-transparent and opaque are used interchangeably in this dissertation). This suggests that there is a diachronic pattern: languages start out transparently and acquire opacity later on. This diachrony is mirrored in language acquisition, as children start their acquisition of grammar with

transparent structures, typically mastering non-transparent structures later on in the acquisition process (Slobin 1977).

The distribution of opaque features over languages is not random. Even the most transparent languages have some non-transparency in their phonology. Moreover, the opaque features attested in highly transparent languages are all cases of redundancy, meaning that a single semantic unit is expressed morphosyntactically more often than strictly necessary. On the other hand, there are some opaque features that are only attested in the most opaque languages, e.g. grammatical gender and sequence of tenses. This suggests, then, that there is an implicational order in which opaque features appear, in languages as such as well as in individuals acquiring them. Determining this ordering by means of establishing an implicational hierarchy is the main aim of this book. The implicational hierarchy of transparency that this dissertation aims to establish comprises many features otherwise thought of as incompatible. However, as this book will prove, it is in fact possible to treat these features as a coherent set. They are united by a single factor: transparency.

## 1.2 Theoretical embedding

Transparency is defined in this dissertation as a one-to-one relation between meaning and form. All structures violating transparency are called non-transparent or opaque. ‘One meaning’ and ‘one form’ are of course both highly problematic concepts. It would be an impossible task to invent a formal and semantic theory of my own to neatly delineate these concepts. Therefore, I will make use of the framework of Functional Discourse Grammar (henceforth FDG; Hengeveld & Mackenzie 2008). FDG provides a set of pragmatic, semantic, morphosyntactic and phonological primitives, as well as an architecture that relates these primitives in a cross-linguistically adequate way. The use of these well-argued concepts and architecture will be of help to define and operationalise transparency.

The notion of transparency has been discussed and studied before in theoretical linguistics, creole studies and language acquisition. This dissertation should be seen principally as part of the former field. Some evidence from creole

studies and language acquisition will be adduced, but the study does not focus on those fields.

## 1.3 Methodology

A typological study will be carried out to establish cross-linguistic patterns of opacity. 25 languages will be checked for the presence of a list of opaque features. The languages studied are all natural spoken languages. Sign languages are excluded, as well as creoles, as including both would add extra variables and complicate matters considerably.

In this dissertation, the assumption is that the implicational hierarchy is an evolutionary pathway as well, that is, that the cross-linguistic distribution of opaque features mirrors different steps on a diachronic path. However, the study will only deal with synchronic cross-linguistic data. The idea behind this is that a cross-linguistic comparison will show what the limits of variation are, since an implicational hierarchy shows which combinations of features are possible, and crucially, which combinations are impossible. Thus, a hierarchy predicts which languages will not be possible and hence which course languages can take in evolution.

## 1.4 Outline

This dissertation is organised in the following way. Chapter 2 describes the theoretical framework of Functional Discourse Grammar and places transparency in the context of that architecture. Chapter 3 further delimitates the notion of transparency, as the term is distinguished from earlier interpretations and other related concepts. The chapter will also devote attention to the simplicity debate in the field of creole studies and carefully distinguish transparency from simplicity. In chapter 4, a list of non-transparent features will be provided, together forming a metric to test the degree of transparency of languages. Research questions and hypotheses will be specified, followed by a precise description and motivation of the

methodology and language sample in chapter 5. The results will be presented and discussed in chapter 6, while the reader is referred to the appendix for the basic typological data. Chapter 7 concludes the dissertation.

# Chapter 2
# Transparency in Functional Discourse Grammar

This chapter will place the notion of transparency into the theoretical framework of Functional Discourse Grammar (Hengeveld & Mackenzie 2008). This enables an operationalisation of the term, allowing the transparency of a language to be quantified in a fine-grained fashion.

Firstly, a description of Functional Discourse Grammar will be provided in 2.1. Section 2.2 presents an FDG-based operationalisation of transparency. The current study is devoted to transparency in grammar rather than in the lexicon, a theoretical decision that will be discussed in Section 2.3. In Section 2.4, five types of non-transparency are distinguished.

## 2.1 Functional Discourse Grammar

Functional Discourse Grammar (Hengeveld & Mackenzie 2008), the successor of Functional Grammar (Dik 1978), is a structural-functional linguistic theory: structural because it aims to find explanations for linguistic structure (that is, form and its organisation), functional because it looks for such explanations in the communicative function of language. FDG is a top-down model of linguistic organisation,

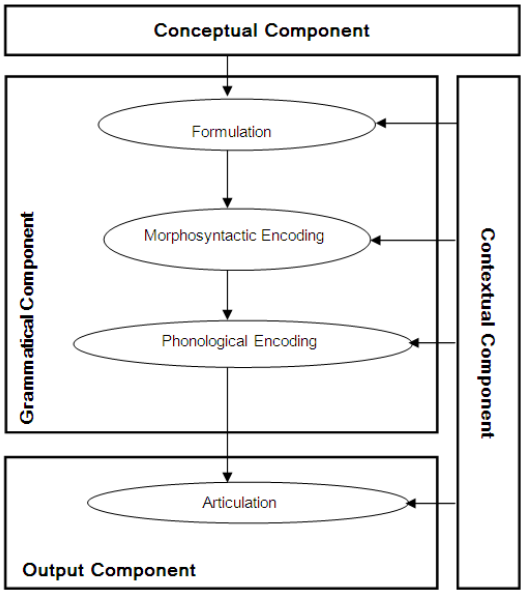



**Figure 2.1: The general architecture of FDG**

starting from the speaker's intention and working down to the spoken or signed utterance.

Figure 2.1 shows FDG's general architecture. The complete model of verbal and non-verbal interaction consists of four components, three of which are non-linguistic: the Conceptual Component, where the non-linguistic intention of the speaker is formed, the Contextual Component, containing knowledge of the context of the speech situation, and the Output Component, where the message is articulated. These non-linguistic components interact with the Grammatical Component, which is the grammar proper in FDG. A more elaborate representation of the internal architecture of the Grammatical Component is given in Figure 2.2. Processes are indicated in Figure 2.1 and Figure 2.2 by means of an oval, whereas levels are represented by rectangles.

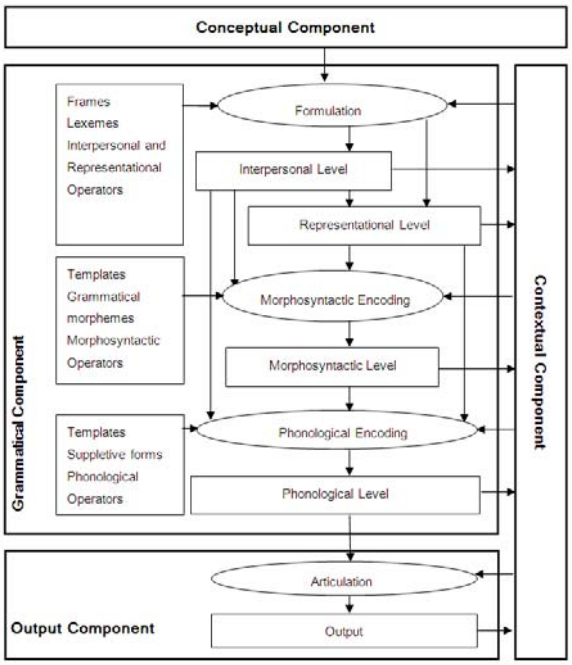



**Figure 2.2: The general architecture of FDG, including levels of representation and primitives**

A speaker's communicative intention is passed on from the Conceptual Component to the Grammatical Component. From there on, it is first subjected to the process of Formulation, meaning that the non-linguistic message is transposed into pragmatic and semantic, hence linguistic, units. The output of Formulation is represented at two hierarchically ordered levels within the Grammatical Component: the Interpersonal Level (at which pragmatic units are represented) and the Representational Level (at which semantic units are located) respectively.

Units from these levels are then passed on to the process of Morphosyntactic Encoding, which means that they are put into morphosyntactic units. The output of this first phase of Encoding lands at the Morphosyntactic Level. From there, it is passed on to Phonological Encoding, where the morphosyntactic units are converted into phonological units and end up at the Phonological Level. Finally, the Phonological Level delivers its units to the Output Component and the message is articulated by means of signs, sounds or letters. Both O'Neill (2012) and Seinhorst (2014) propose a fifth level of organisation, viz. a Phonetic Level, and argue that this fifth level is necessary to explain phonological and phonetic processes. Even though I agree with their analysis, I will not make use of a fifth level in the current study, because the architecture of such a level is as yet insufficiently known to allow for a proper study of transparency at this level.

Note that it need not be the case that a message passes all levels of linguistic organisation in FDG. All levels and processes interact with each other, as indicated by the arrows in Figure 2.1. Thus, it is possible that a unit passes from the Conceptual Component, through the Interpersonal Level, directly to the Phonological Level. This is for instance the case with interjections such as 'Ouch!', which have no semantic or morphosyntactic structure.

The four levels of the Grammatical Component all make use of specific primitives, given in Figure 2.2 in square boxes. These primitives are stored in the lexicon, which is located in the so-called Fund (cf. García Velasco 2013). Coming back to these primitives below, I will first discuss the internal structure of the four levels of linguistic organisation. These levels are each divided internally in hierarchically ordered layers. In FDG, it is common to represent these layers and other units at the Interpersonal Level by means of capitals (M, A), at the Representational Level by means of lower case letters (p, ep), at the Morphosyntactic Level by a combination of the two (Cl, Np) and at the Phonological Level by small capitals (U, IP). Equipollent (non-hierarchically ordered) units are put between square brackets. In the following exposition of FDG terms I will conform to the FDG notation, but in other chapters I will use common notations such as NP rather Np, to avoid confusion.

The upper layer in the Interpersonal Level (IL) is the Move (M), a unit that consists of one contribution to a conversation by a speaker. A Move can consist of one or more Discourse Acts (A). A Discourse Act is an action by a speaker, 'the smallest identifiable unit of communicative behaviour' (Hengeveld & Mackenzie 2008: 60). Moves and Discourse Acts are not always easy to distinguish, but the main difference is that Acts do not have a communicative goal, while a Move is always intended to provoke a reaction of or be an appropriate response to another Move. Every Discourse Act contains at least an Illocution (F), Participants (P, at least a speaker $P_1$ and often a hearer $P_2$) and a Communicated Content (C): the content of the message. Within the Communicated Content, Subacts are found. A Subact can either be an act of making reference to someone or something (a Referential Subact, R) or an act of ascribing a property to a referent (an Ascriptive Subact, T). The complete structure of the Interpersonal Level is represented in Figure 2.3.

**Figure 2.3: The layers of the Interpersonal Level**

```
M: [                                Move
        (A:     [                           Discourse Act
                (F: …)                      Illocution
                (P1: …)                     Speaker
                (P2: …)                     Addressee
                (C:     [                           Communicated Content
                        (T: …)                          Ascriptive Subact
                        (R: …)                          Referential Subact
                ])                                  Communicated Content
        ])                                  Discourse Act
])                                  Move
```

The highest layer of the Representational Level is the Propositional Content (p), a proposition that exists in a speaker's mind and is therefore subject to beliefs, doubts, etc. The next layer is that of the Episode (ep), a unit that exists in time and can therefore be real or unreal. An Episode consists of one or more thematically

coherent States-of-Affairs (e). Such States-of-Affairs have a temporal dimension too, but are located in time only with respect to each other, not with respect to the here and now. A State-of-Affairs contains a Configurational Property (f): a unit that determines the predicate-argument relations. The predicate itself is an example of a unit called Lexical Property (commonly abbreviated as f, but to make a clearer distinction with the configurational property, I will henceforth use $f^c$ and $f^l$). Arguments are entities called Individuals (x). Locations (l), Times (t), Manners (m), Quantities (q) and Reasons (r) are further semantic units that exist in some languages. The complete structure of the Representational Level is given in Figure 2.4.

**Figure 2.4: The layers of the Representational Level**

```
p:                          Propositional Content
    (ep:                        Episode
        (e:                         State-of-Affairs
            (f^c: [                     Configurational Property
                (v: …)                      any semantic category
            ])                          Configurational Property
        ])                          State-of-Affairs
    ])                          Episode
])                          Propositional Content
```

Both the Morphosyntactic and the Phonological Level are internally layered as well. Morphosyntactic and Phonological layers require less explanation than the pragmatic and semantic layers discussed above, as the concepts are much more widely known[1], but the highest layer at the Morphosyntactic Level, the Linguistic Expression (Le), is FDG-specific and therefore deserves some extra attention.

[1] I do not intend to say that the meanings of these concepts are 'widely known' in the sense that there is agreement among linguists on their definition, but since the morphosyntactic and phonological terms listed here are used cross-theoretically, I will assume that the linguist reader will have an idea of what they are. Therefore, I do not describe them further in this chapter.

Linguistic Expressions are the highest morphosyntactic units that can consist of one Word only, but also of multiple clauses. Other units at the Morphosyntactic Level are Clauses (Cl), Phrases of different types (Xp, e.g. Np), Words of different types (Xw, e.g. Nw) and Affixes (Aff). At the Phonological Level we find Utterances (U), Intonational Phrases (IP), Phonological Phrases (PP), Phonological Words (PW), Feet (F) and Syllables (S).

An important property of FDG is that it does not assume universal status for all layers. A specific layer is only assumed to exist in a language if it is relevant in that language, that is, when it is necessary to account for a structure in the language. For example, a Phonological Word is only postulated if that language shows evidence for such a unit, for instance a specific stress pattern that distinguishes Phonological Words from Phonological Phrases. This is a far-reaching difference with for instance Generative Grammar, a framework in which it is thought that units are universally present in all languages, even though they may not always show up at the surface.

The Interpersonal Level (henceforth IL) and the Representational Level (henceforth RL) both use categories of primitives called Lexemes, Frames and Operators. Lexemes are, unsurprisingly, lexical units. Lexemes should be seen as bundles of pragmatic, semantic, morphosyntactic and phonological information, in close interaction with the Contextual Component (cf. García Velasco 2013 for a discussion of the lexicon in FDG).

The second type of primitive that is found at IL and RL, is that of Frames. Frames are abstract configurations that host units with specific properties and functions and relates them to each other. For instance, predication frames are found at the layer of the Configurational Property to code semantic valency and the semantic relations between Individuals by indicating whether they are Actor, Undergoer, or Location.

Other primitives found at IL and RL are Operators: grammatical devices that operate on specific layers. Examples are absolute and relative tense operators (locating Episodes and States-of-Affairs in time) and Reportative operators at the layer of the Communicated Content. Operators are always grammatical units – their

lexical counterparts are called modifiers. Modifiers can consist of Lexemes or phrases.

The Morphosyntactic Level (henceforth ML) makes use of its own primitives, the first of which is the category of Templates. Like Frames, these are configurational units, that is, units that host other units and demonstrate their relation to each other. For instance, word order Templates are translations of pragmatic, semantic and morphosyntactic information into a linear ordering of morphosyntactic units in a sentence.

Another morphosyntactic primitive type is the Grammatical Morpheme: the smallest meaningful grammatical unit. Grammatical Morphemes are comparable to Lexemes in the sense that they are ready-made forms, however, they are grammatical rather than lexical. Grammatical Morphemes can be free morphemes and therefore, they have to be introduced at the Morphosyntactic Level, so that they participate in the syntactic configuration of the utterance.

The third morphosyntactic primitive is the Morphosyntactic Operator, which, like an operator at the Interpersonal or Representation Level, can modify a specific layer (e.g. a morphosyntactic past tense operator operates in English on a Morphosyntactic Word). Morphosyntactic Operators introduce irregular and suppletive morphosyntactic forms. For instance, an English regular past tense is modelled in FDG as a Verb (Vw, a Morphosyntactic Word) combined with a Grammatical Morpheme (the affix *-ed*). An irregular strong verb stem is a combination of a Verb and a Morphosyntactic Operator <past>, which will trigger the correct phonological output at the next level.

The Phonological Level (henceforth PL) also contains a set of primitives called Templates. These Phonological Templates are units that regulate the prosody of (combinations of) phonological units. Secondly, the PL makes use of Suppletive Forms. These are the final output forms of the Grammatical Morphemes and Morphosyntactic operators that were selected at the Morphosyntactic Level. For instance, when at ML the Grammatical Morpheme *-ed* was selected, it is replaced in Phonological Encoding by /-əd/, /ɪd/, or /-d/, depending on Phonological Word it is attached to. Note that Suppletive Forms have to be introduced *after*

Morphosyntactic Encoding (that is, in Phonological Encoding), since their form can be affected by adjacent units and hence by the syntactic configuration – the output of Morphosyntactic Encoding. The third type of phonological primitive, Phonological Operators, is responsible for e.g. intonation, stress and tone; they are phonological properties that operate on other phonological units.

## 2.2 Defining transparency in Functional Discourse Grammar

Transparency can initially be defined in general terms as a one-to-one relation between meaning and form. Non-transparent or opaque relations are all meaning-form relations that are not one-to-one. FDG can be employed to make this definition more precise, by delineating what exactly are meanings and forms. In FDG terms, one 'unit of meaning' is a unit, i.e. a primitive (function, operator or layer), at one of the upper two levels, viz. the Interpersonal Level or the Representational Level. Note that this means that not only Illocutions and Participants are meaning units, but also, for example, a Topic function and an Undergoer function. Such functions are in other frameworks not always seen as units, but rather as properties of morphosyntactic units. Note furthermore that the word 'function' has a second sense: it can also be a non-linguistic concept, referring to the effect or outcome of some process or phenomenon. For example, the function of using reduced forms is to speak more quickly. Function in that sense is not considered a meaning unit here – the term 'meaning' in this dissertation always refers to something linguistic. This entails that information present in the context but not coded linguistically will not be considered a meaning unit.

A 'unit of form' is a unit, that is, a primitive, at the lower two levels, viz. the Morphosyntactic Level or the Phonological Level. This means that layers (e.g. Morphosyntactic Words) are units of form, and so are operators and suppletive forms. Some formal units contain multiple other formal units, e.g. a Phonological Phrase can consist of multiple Phonological Words. Relations are considered transparent when all units can be related to a higher unit, e.g. when a Phonological

Phrase corresponds to a Morphosyntactic Phrase, and the Phonological Words contained by the phrase each correspond to Morphosyntactic Words.

Now that this is clear, we can create a more precise definition of transparency. The superficial definition above can be reformulated in FDG terms as in 1).

1) Transparency obtains when one unit at one of the upper two levels of linguistic organisation (IL, RL) corresponds to one unit at one of the lower two levels of linguistic organisation (ML, PL).

This definition straightforwardly captures the fact that transparency is about the relation between form and meaning. It shows that transparency is an interface property - not a property of specific levels. Note furthermore that transparency is defined as a property of the grammar, rather than of the lexicon. Lexical non-transparency, e.g. phenomena like synonymy, homonymy, and polysemy, will not be studied in this dissertation, as will be discussed in further detail in Section 2.3.

In fact, definition 1) does not make full use of FDG's potential, because it only takes into account interfaces between meaning-to-form interfaces, whereas the four levels of linguistic organisation all interact with each other so that there are meaning-to-meaning and form-to-form interfaces as well. There is not just one interface between the upper two and the lower two levels, but there are six interfaces between all four levels and their combinations (IL-RL, IL-ML, IL-PL, RL-ML, RL-PL, ML-PL). Mismatches can occur at each of these interfaces. Hence, it is possible to include all interfaces in the notion of transparency and reformulate definition 1) as definition 2).

2) Transparency obtains when one unit at one level of linguistic organisation corresponds to one unit at all other levels of organisation.

The definition now includes mismatches between, for instance, IL and RL, or between ML and PL. Surely, a mismatch between morphosyntax and phonology

would not be a violation of a one-to-one relation between meaning and form, and some readers might therefore regard definition 2) as too broad.[2] However, the inclusion of such mismatches is in my opinion an advantage, as it enables the study of an empirical question: are meaning/form mismatches (IL/RL-ML/PL mismatches) different from other mismatches (IL-RL and ML-PL mismatches), or can we take all mismatch phenomena together as one large group with a single explanation? If all mismatch phenomena can be shown to be related in a cross-linguistic distributional pattern, there is evidence that mismatch phenomena are all similar, regardless of the interface they are located at. Another possible finding is that the mismatches between the upper two and the lower two levels (IL/RL-ML/PL) do not fit into a pattern with other mismatches, that is, that mismatches at the IL-RL and ML-PL interfaces show a fundamentally different distribution over languages. In that case, we have an empirical argument to adopt definition 1) rather than 2). Until this question is answered, definition 2) is adopted. Section 5.1 will come back to this point.

An essential hypothesis in the current dissertation is that transparency is a graded notion and not a binary concept, as languages have a degree of transparency, rather than being either transparent or non-transparent. Presumably, every language violates transparency somewhere in its grammar and lexicon, so that an opposition between transparent and non-transparent languages cannot be upheld. Rather, I propose a continuum running from relatively transparent languages to relatively non-transparent languages. The degree of transparency of a language can be measured by counting non-transparent features. Counting opaque rather than transparent features may seem counter-intuitive, but is necessary as it is quite impossible to count transparent features, since in most cases, transparency involves the *absence* of certain phenomena, e.g. the absence of grammatical gender, and one cannot count what is not there. A list of non-transparent features will be provided in

[2] Henceforth, I will keep using the term 'one-to-one relation between meaning and form' as a shorthand for definition (2), as it would be too cumbersome to take over the entire definition each time it is used.

Chapter 4, but first, the next section will discuss what is not taken into account in measuring the degree of opacity of languages.

## 2.3 Non-transparency in the lexicon

As noted above, I will only study the transparency within the grammar in this dissertation, and disregard the degree of transparency within the lexicon. Non-transparency exists in the lexicon as well, since homonymy and polysemy involve relations between multiple meanings and one form, and synonymy refers to relations between one meaning and multiple forms. Presumably, languages differ in their degree to which their lexicon displays homonymy, polysemy, and synonymy, as stated for instance by Slobin (1977: 190ff.) who opposes an allegedly low degree of homonymy in Turkish to a high degree of homonymy in Serbo-Croatian. However, to my knowledge, there are as yet no ways to measure the degree of transparency of the lexicon systematically. Therefore, I will refrain from studying non-transparent properties of the lexicon in this dissertation.

Apart from homonymy, polysemy and synonymy, yet another non-transparent features will be excluded from the study because it pertains to the lexicon rather than to the grammar, viz. semantic irregularity. Semantic irregularity refers to the situation that a combination of particular forms does not result in a straightforward combination of their meanings. In derivation and compounding, this results in idiosyncratic meanings that cannot be inferred from the meanings of the separate parts (cf. Aboh & Smith 2009). For example the Dutch compound noun *bagagedrager* ‘luggage carrier’ includes not only the meanings of ‘luggage’ and ‘carrier’, but also the unpredictable aspect that this is a part of a bicycle, rather than a for example a person carrying luggage. As explained further in Section 3.1.4, this results in non-transparency, since the one-to-one relation between the meaning and form of the separate elements is obscured, e.g. *baggage* ‘luggage’ turns out to have a different meaning in different contexts. In rare cases, inflection may have a similar effect, as for example in the case of the plural noun *brethren*: this is a combination of the Lexeme *brother* and plural inflection, but denotes something else than just

'brother-PL', viz. a name for a member of a certain Christian cult. The extent to which semantic irregularity is displayed may differ from language to language, but, like homonymy, there is as yet no reliable metric that is able to quantify such cross-linguistic variation.

In FDG, cases of non-transparency in derivation, compounding and inflection are seen as pertaining to the lexicon. Returning to the *bagagedrager* example, we find in the lexicon the entries BAGAGE 'luggage', DRAGEN 'to carry', -ER 'someone who performs the action specified by the verbal stem' and BAGAGEDRAGER 'luggage carrier' – each involving mappings of a particular meaning and a particular form. Thus, the word *bagagedrager* is stored in the lexicon with its own specific meaning and as such not inherently more opaque than *bagage* or *-er* – they are all meaning-to-form mappings. The addition of idiosyncratic meaning in the combining of particular morphemes is not a systematic grammatical process, but draws into matters of world knowledge and language use, which is why it is located in the lexicon rather than in the grammar in FDG.

In linguistic tradition, especially in generative syntax and morphology, transparency is often understood as boiling down to full compositionality. According to that reasoning, a Lexeme (or any other linguistic form) is transparent when it has the exact same meaning in each different morphological or syntactic context. Taking into account that all languages exhibit unpredictable meaning in compounds and derivations, as argued convincingly by Aboh & Smith (2009), this would mean that measuring the degree of lexical transparency of a language is the same as measuring the degree of semantic regularity. As said above, this is as yet impossible, since no method is available to quantify this.

## 2.4 Categories of opacity

One-to-one relations between levels can be violated in different ways. There are four logically possible inter-level relations that are not one-to-one, viz. null-to-one, one-to-null, many-to-one and one-to-many relations. These four types of non-

transparency, along with a fifth one that involves violations of domain integrity, will be the topic of this section.

One-to-null interface relations are elements like understood arguments, phonologically empty operators and empty categories, i.e. elements that are postulated to be present in a sentence, but are not audible or visible in the actual output. Their existence is often hypothetical and solely theoretically motivated. Linguistic theories differ considerably in the extent to which they assume covert elements: generative grammarians postulate phonologically empty operators relatively easily in order to successfully explain empirical data, while functionalist theories like FDG are reluctant to use them since they claim that the existence of invisible elements cannot be falsified and is therefore theoretically undesirable. Still, even functionalists adhere to the existence of some invisible items, for example to the presence of arguments at the semantic level in languages with a low referential density, in which arguments are often not expressed overtly (cf. Bickel 2003). Even though these one-to-null relations are most certainly non-transparent, they will not be taken into consideration in the current study, as it is to my knowledge impossible to objectively determine or quantify their existence. An exception is made for zero-morphemes, also known as null affixes (cf. Bauer 2003: 37), the existence of which can be demonstrated by means of a comparison within a paradigm.

A second category of non-transparent relations is that of null-to-one relations. It includes any morphosyntactic or phonological form that is not motivated by or related to a higher-level unit. Such forms are not triggered by a pragmatic or semantic unit, but are the result of a morphosyntactic or phonological rule or process. An example is the use of dummy subjects, since those are inserted in order to satisfy a rule stating that there should always be a subject in a sentence. A dummy does not refer to anything or have any semantic content of its own; in FDG terminology one would say that there is a morphosyntactic Subject that does not relate to a Referential Subact or Individual (cf. Section 4.4.2). This type of opacity has been referred to in the literature as autonomous syntax. I choose not to use this term in this study, as it is too theoretically charged. In search of a more neutral term,

Hengeveld (2011) and Leufkens (2013a) use ‘form-based form’, which will be the term adopted here.

Form-based form features constitute prototypical cases of features with a high degree of syntacticity or phonologicity, i.e. cases of syntactic or phonological form that do not have any pragmatic or semantic counterpart or motivation. The term syntacticity is used for instance in studies into agreement (cf. Berg 1998 for a good example), in which agreement conflicts can arise when for instance a noun has two genders, one semantically assigned and one abstract. In such cases, a language that follows abstract gender is said to display a relatively high degree of syntacticity, as the gender system apparently abstracted away from semantic origins and is motivated by morphosyntactic properties. I will use the term syntacticity in a broader sense, referring to the syntacticity/phonologicity of the entire language, that is, the degree to which forms and rules are motivated by morphosyntactic considerations. In a similar vein, languages may show a high degree of phonologicity when phonological processes overrule other principles at work (cf. Hyman 2008 for an example of the use of the term phonologisation). Languages with a low degree of syntacticity and phonologicity can be said to have a high degree of semanticity and/or pragmaticity.

A third type of violation of one-to-one relations involves one-to-many relations, in which one unit of meaning is expressed by multiple forms. In such cases, one of the forms is not necessary; it is redundant, as it does not provide any additional information. This type of violation of transparency is therefore called redundancy. Redundancy can be obligatory, in which case both formal elements are always present in the output structure, or optional, if one of the redundant forms can be left implicit, for instance in pro-drop languages that show predicate-argument agreement also when the argument is implicit. Obligatory redundancy is often seen as an automatic, purely morphosyntactic operation, thus having a high degree of syntacticity, while optional redundancy is clearer pragmatic and semantic motivations. Sections 4.1.3 and 4.1.4 on different types of agreement will go into this topic and discuss the theoretical implications.

The fourth category of opacity consists of many-to-one relations, that is, relations between multiple meanings and a single linguistic form, and is called fusion. A straightforward example of such a many-to-one relation is fusional morphology, that involves the expression of different meanings by single forms, which are also referred to as portmanteau morphemes. Note that the term 'fusion' is used interchangeably with 'many-to-one relation' and should not be interpreted historically, as I by no means intend to say that all fusional morphemes have been separate morphemes in the recent or remote past.

A fifth type of transparency violations occurs when the principle of domain integrity is violated, which states that what belongs together at IL or RL, should be juxtaposed at ML (Hengeveld & Mackenzie, 2008: 285). Violation of domain integrity results in discontinuous units, which is why the category is named 'discontinuity', applying for instance in the French negating circumfix *ne pas*: one semantic unit (a negation operator) corresponds to two formal units at the Morphosyntactic Level (*ne* and *pas*). *Ne* and *pas* cannot occur without each other – they are not two independent units[3], which makes it impossible to speak of 'one unit of form', since the boundaries of the unit are unclear. Rather, discontinuous units, be they syntactic constituents or morphemes, involve 'one-to-fragments-relations'. Because of the similarity with one-to-many-relations, Leufkens (2013) grouped the categories of fusion and discontinuity together into one category called 'domain disintegration'. However, that term, as well as the combination, implies that fusion always results from a violation of domain disintegration. Since this is not necessarily the case, I have decided to treat the categories separately in this dissertation.

In this section, five categories of non-transparent features have been distinguished, viz. one-to-null relations or covert elements, form-based form, redundancy, fusion and discontinuity. The former category will not be taken into consideration for reasons that I have given above. Before Chapter 4 will go into the precise phenomena that fall into these categories, Chapter 3 will compare the

[3] Note that in contemporary French, *ne* is consistently dropped. This is a good example of a language change that increases the language's transparency (cf. Section 3.4.2).

definition of transparency given in FDG with definitions of the term in earlier studies from various subfields of linguistics.

# Chapter 3
# Transparency[4]

This chapter will place the notion of transparency in a wider theoretical perspective by discussing the way it was defined in earlier approaches. Some of the previous studies on transparency involve quite different interpretations of the term. Those interpretations will be discussed in section 3.1, distinguishing them from the concept of transparency as defined in the current study. This will be followed in Section 3.2 by a historical account of highly similar uses of the term transparency, notably in theoretical linguistics and creole studies. In both fields, transparency has often been associated with the concept of linguistic simplicity. I believe that those notions should be kept apart and will devote section 3.3 to argue for this position. The chapter concludes with a section on the evidence for patterns in the distribution of transparent features in the fields of typology, diachrony, and language acquisition.

## 3.1 Other interpretations of transparency: delimitation of the concept

In this section I will discuss several interpretations of the term transparency that are fundamentally different from the way I use the term in this study. In other words: this section will determine what transparency is *not*. Thus, I will further delimit the boundaries of my use of the term.

### *3.1.1 Counterbleeding and counterfeeding*

Opacity is known in phonology as an umbrella term for the notions of counterbleeding and counterfeeding (Kiparsky 1973), which are processes that arise through specific orderings of two or more phonological rules. In the case of

[4] Many of the ideas expressed in this chapter have appeared in a different form in Leufkens (2013a). I am grateful to Peter Bakker, John McWhorter and an anonymous reviewer for their feedback on earlier versions of that paper.

counterfeeding, a rule ordered second creates an element that would be a trigger for the first rule, but the first rule cannot have its effect after that second rule. For instance, Bakóvic (2011: 3) gives a hypothetical example of a vowel deletion rule V → Ø / _V that follows a palatalisation rule t → ʧ / _ [–back]. An input /tue/ does not undergo palatalisation, since the /t/ does not precede a back vowel, but it does undergo vowel deletion. This creates the output [te], to which palatalisation should apply, but it can no longer have its effect because of the rule ordering. Hence, the output is a form that is not phonologically allowed in the language, creating opacity between phonological rules and their output.

Counterfeeding works the other way around: the second rule deletes an element that served as a trigger for the first rule, so that the output appears to be a violation. For example, again following the hypothetical phonological rules by Bakóvic, the input form /tio/ undergoes palatalisation and then also vowel deletion, creating the output [ʧo]. [ʧo] superficially violates the first rule, since the [ʧ] precedes the front vowel [o] – from the output, it is not clear why palatalisation has occurred. In both counterfeeding and counterbleeding, the output obscures the working of the first rule, as an element surfaces that the first rule should have had an effect on, or does not surface while it must have been there in order for the first rule to apply. Counterbleeding and counterfeeding effects cannot be captured in traditional output-based constraints and are therefore problematic for phonological theory based on parallel OT.

This kind of opacity is entirely different from opacity as addressed in this dissertation – while phonological opacity refers to a particular effect of sequential phonological rules, opacity as studied here is always a property of interfaces between different levels of grammar.

### *3.1.2 Iconicity*

The notion of iconicity can be interpreted in a narrower sense and in a broader sense (cf. Haspelmath 2008 for an extensive analysis of the uses of the term). In the narrower sense, iconicity refers to the predictability of a meaning from its form: if a form is iconic, one can infer (part of) its meaning from its sound or form. This is for

instance the case with onomatopoeia, in which the sound of the word mimics the meaning denoted. Most linguistic form, however, is not iconic, because the relation between the form and its meaning is to a large extent arbitrary. In sign language studies, iconicity and transparency are sometimes treated as synonyms, for example in Bellugi & Klima (1975: 525):

> "… given an ASL sign, and no other information, could a nonsigner correctly ascertain its meaning? To the extent that this is possible, a sign would be considered *transparent*."

However, in this dissertation, transparency relates to the number of meanings expressed by forms and not to the predictability of meaning from form.

Iconicity in the broader sense of the word, referred to as 'diagrammatic iconicity' or 'iconicity of motivation' by Haiman (1980: 515), is discussed extensively by Haiman (1980, 1983, 1985), Givón (1985), Croft (2003a: 101ff.), Haspelmath (2008) and many others. In an iconic meaning-to-form relation "the structure of language reflects in some way the structure of experience" (Croft 2003a: 102), or as Haiman (1980: 515) puts it: "the structure of language directly reflects some aspect of the structure of reality". For example, the intuition is that if there are multiple objects, their linguistic expression (plural) will consist of a larger amount of form (a longer word) than a single object and a singular form (iconicity of quantity, Haspelmath 2008). In a similar vein, iconicity of sequence predicts that events that happened in a certain order will also be expressed linguistically in that order, so that syntax reflects semantics directly (Greenberg 1966 [1963]: 103). This dissertation will not go into this type of iconicity either, since it deals with the quantitative relation of meaning to form rather than with the arbitrariness or predictability in that relation. Crucially, in an iconic relation, the form mimics or embodies the meaning, while this is not required for a relation to be transparent (cf. Itkonen 2004 for an elaboration on the distinction between iconicity and what he calls the 1M1F principle). Note that Haiman (1980) distinguishes another subtype of iconicity: iconicity of isomorphism, which is in fact synonymous to transparency as

I have presented it in this dissertation. I will come back to this notion and Haiman's views on it in section 3.2.1.

### *3.1.3 Homomorphism and isomorphism*

The term 'one-to-one relation' used throughout this dissertation is reminiscent of the algebraic notions of homomorphism and isomorphism (cf. Homomorphism lemma 2014 for an explanation of the mathematical terms). Homomorphism applies when units in some set are structurally identical to units of another set, and isomorphism applies when homomorphism holds bi-directionally between the sets. In linguistic analysis, isomorphism holds when units at one level of organisation, e.g. semantics, always correspond to the same number of units at another level, e.g. syntax. This gives rather strong predictions, such as that two syntactic structures having similar semantics but different syntactic structures, for instance active and passive sentences, should have two different semantic structures[5]. Therefore, when arguing for homomorphism or isomorphism between levels one runs into complicated theoretical debates, which go too far to be fought out within the scope of this project. I will therefore use the terms transparency and one-to-one correspondence in a non-mathematical sense and refrain from making claims relating to homomorphism or isomorphism.

A non-mathematical linguistic interpretation of the term isomorphism is also used by Croft (2003a) and Itkonen (2004). Within isomorphism, Croft (2003a: 102ff.) distinguishes meaning-to-form mappings in the lexicon (paradigmatic isomorphism) and grammars (syntagmatic isomorphism). To Croft, isomorphism is one aspect of the larger notion of iconicity, as opposed to Itkonen (2004: 21) who sees no hierarchical relation between the two notions. For him, iconicity reflects "structural similarity […] between extralinguistic reality and language", while isomorphism refers to the principle of one meaning - one form, i.e. transparency.

[5] I am grateful to Paul Boersma for pointing this out to me.

### *3.1.4 Transparency of compounds and derivations*

In morphology, transparency is considered a property of compounds and derivations. A compound is regarded as transparent when the semantics of the separate elements are straightforwardly combined in the semantics of the compound, as for example ‘carwash’ combines the separate meanings of ‘car’ and ‘wash’. A compound like ‘hogwash’ on the other hand is considered opaque, as its meaning is not straightforwardly composed of ‘hog’ and ‘wash’. Opacity in this sense is also referred to as semantic irregularity, and discussed under that name in Section 2.3. Aboh and Smith (2009) argue that in fact all derivations and compounds are semantically irregular, since the meaning of any compound or derivation cannot be predicted from exclusively combining the semantics of the separate elements – one has to know which part of the semantics is to be used . For example, knowledge of the world is necessary to know that ‘carwash’ is the location where cars are washed, rather than the liquid that is used, as is the case with ‘mouthwash’. Taking this seriously, no compound or derivation is ever fully semantically regular. Thus, in the current study, semantic irregularity is seen as a non-transparent property that all languages exhibit at least to some degree.

McWhorter (1998) mentions a high level of semantic regularity as a characteristic of creoles and pidgins. As McWhorter is known for arguing that creoles are relatively simple, frequently taken to mean that they are transparent, the feature has gained a lot of attention in the debate on the alleged simplicity and transparency of creoles. This debate and the role of semantic regularity in it will be discussed extensively in Section 3.2.3, but it should be made clear now that transparency in this dissertation is not equated with semantic regularity. As mentioned above, opacity of compounds and derivations is seen here as one of many non-transparent feature that languages may exhibit – semantic irregularity is a non-transparent feature, but transparency does not equal semantic regularity.

Note that even though semantic irregularity is considered a non-transparent feature of languages, it will not be studied in this dissertation because following FDG, it is a property of elements in the lexicon, while this study aims to study non-transparency in the grammar (see Section 2.3). Furthermore, at the moment, no valid

method exists to establish the degree to which languages have transparent compounding and derivation. This renders it as yet impossible to say whether a language is relatively transparent or opaque in this respect.

### *3.1.5 Simplicity*

The interpretation that is perhaps most frequently given to the term transparency is that of simplicity – in my opinion, wrongly. Well-known in this respect is McWhorter (1998, 2001), who states that creole languages are fundamentally simpler than non-creole languages and judges certain characteristic creole features as grammatically simple. Moreover, McWhorter criticises the so-called semantic transparency hypothesis, proposed by Seuren & Wekker (1986; see below for a detailed account of this hypothesis). Even though McWhorter explicitly disagrees with Seuren & Wekker, the semantic transparency hypothesis and McWhorter's account of simplicity are often seen as one and the same thing in the creole literature, as will be discussed in further detail in Section 3.2.3. Not only in creole studies but also in numerous other cases, transparency and simplicity are either equated, or transparency is seen as one type of simplicity. Since the two notions are intertwined to such a high extent, a separate subsection will be devoted to detaching the notions, viz. Section 3.3.

## 3.2 Earlier studies on one-to-one correspondence

The term transparency in the sense of a one-to-one correspondence between meaning and form has been used in several subdomains of linguistics. Both Langacker (1977) and Lightfoot (1979) use transparency to explain the direction of language change, Slobin uses transparency in his work on language acquisition, and Seuren & Wekker (1986) are not the first to claim that creoles are characterised by a high degree of what they coin semantic transparency. That claim later becomes part of a vigorous debate on the supposed simplicity of creoles, a debate that has involved the challenging of long-adopted truisms and sometimes even their rejection. Despite this, confusion and inconsistency still abide with respect to terminology.

Therefore, it remains necessary to clearly define one's terms, as in Chapter 2, and show how one's definitions relate to those of others, which is what I will do in the current section. A chronological overview will be given of various transparency studies in three linguistic fields, and the views of each study will be compared to the views on transparency in this dissertation.

### *3.2.1 Theoretical linguistics*

For Langacker (1977), transparency is a category of linguistic optimality, next to various types of simplicity and perceptual optimality. Language change always results in the optimisation of one of these categories, possibly at the cost of others, so that a language may for instance gain construction simplicity while decreasing transparency. Transparency is more precisely defined by Langacker as 'a one-to-one correspondence between units of expression and units of form' (1977: 110). Several linguistic processes contribute to such correspondences: the elimination of 'morphemes with no obvious meaning or syntactic function', a tendency to 'boundary coincidence' and thirdly the minimisation of allomorphic variation – all features that are considered opaque in the current study as well, under the headers of form-based form, discontinuity and allomorphy respectively (cf. Chapter 4).

Lightfoot (1979) is less elaborate in his definition, but proposes a Transparency Principle that 'requires derivations to be minimally complex and initial, underlying structures to be 'close' to their respective surface structures' (1979: 121). According to Lightfoot, this principle guides syntactic re-analysis and is therefore a leading factor in language change, as languages that become too opaque will re-analyse non-transparent structures and make them transparent. This, too, is largely in agreement with my views on the existence of an opacity ceiling, as explained in Section 3.4.2.

While Langacker and Lightfoot both take a syntactic point of view, Bybee (1985) places the one-to-one-correspondence principle in a morphological context. She argues that such a principle may intuitively represent the most economical way to express meaning, but that its violations may have an explanation as well, as they may have positive effects for communication. To put it differently, she argues that

cases of non-transparent morphology, e.g. fusion, allomorphy or zero-morphemes, should not be seen as aberrations but rather as the result of other communicative principles overriding transparency. Thus, her work fits nicely into the so-called competing motivations paradigm (cf. Croft 2003a: 59ff.), that sees language forms as the outcome of a process of variation, competition and selection, comparable to a "survival of the fittest" scenario in biology.

Another important author in this paradigm is Haiman (1980, 1983, 1985). Haiman discusses several types of iconicity, one of which he calls iconicity of isomorphism, a somewhat confusing term in the light of sections 3.1.2 and 3.1.3 in which I have argued that neither isomorphism or iconicity are the same as transparency. However, Haiman interprets isomorphism as the one-to-one correspondence that I term transparency, which is why it has to be discussed here. For reasons of clarity, I will consistently use the word transparency where Haiman uses iconicity of isomorphism. To some extent, Haiman's work is a theoretical exposition on the existence of isomorphism, trying to answer questions such as: is full synonymy possible? When can we speak of two separate meanings? Do an active and a passive sentence have different semantic structures? As explained in Chapter 2, my study tries to answer such theoretical questions by making use of Functional Discourse Grammar.

Haiman (1985) argues for an inverse correlation between transparency and economy[6], for example by showing how morphophonological change from analysis to synthesis obscures one-to-one relations between meaning and form (1985: 160ff.). In doing so, he discusses many phenomena that will be studied in this dissertation as well, such as agreement, redundancy and fusion, and tries to explain them as an outcome of the conflict between transparency and economy. Givón (1985), on the other hand, argues that economic motivation is not necessarily in opposition to transparency. Discussions like these explicitly take the position that extra-linguistic factors such as transparency and economy compete with each other, resulting in the selection of an optimal linguistic form.

[6] Haiman (1983) argues for a similar inverse correlation between iconicity and economy, but iconicity will not be discussed in this dissertation, for reasons given in Section 3.1.2.

Such approaches are formalised for instance by Carstairs-McCarthy (e.g. 1987) and in the theory of Natural morphology by Dressler et al. (1987). Dressler (1985: 323ff.) gives 'morphotactic transparency' an explicit place in this framework, which aims to establish which linguistic phenomena are most natural. Such phenomena are expected to be most frequent in the world, even though competition with other factors, such as, again, economy and iconicity, may obscure transparency. Dressler's morphotactic transparency is somewhat broader than my interpretation of transparency, as it also claims that morphotactically transparent forms are easy to process, while the definition adopted here is more theoretical in nature and does not include easy processing. However, the notions are similar in that they are both gradual concepts and state that linguistic form should always have a 'conceptual', i.e. pragmatic or semantic motivation.

These accounts of competing motivations can all be placed on the functional side in the linguistic debate on autonomous syntax and the arbitrariness of linguistic form; a debate that has been ongoing ever since Chomsky (e.g. 1980) claimed that linguistic form is completely autonomous from meaning and communicative function (cf. Haiman 1985: 3ff.). Newmeyer (1998), a proponent of Generative Grammar and hence of Chomsky's views, articulates those views in the AUTOSYN hypothesis: "Human cognition embodies a system whose primitive terms are nonsemantic and nondiscourse-derived syntactic elements and whose principles of combination make no reference to system-external factors" (1998: 23). This hypothesis entails at least the existence of some arbitrary form, that is, "situations in which differences in form do not correlate with (relevant) differences in meaning" (1998: 28) or in my terms: non-transparency.

Adversaries of AUTOSYN refer to themselves as functionalists, as they give pride of place to the function of language, i.e. the transmission of meaning, in determining linguistic form. Radical functionalists (e.g. Cognitive Grammar, Langacker 2008) claim that all linguistic form is finally shaped by language external factors, so that there can be no linguistic form that is not ultimately caused by a pragmatic or semantic motivation. Phonology, morphology and syntax should in that scenario always be explained in pragmatic or semantic terms. More moderate

functionalists believe that autonomous linguistic form does occur, but is accidental to language, or in other words: there may be syntactic, morphological or phonological processes that cannot be the result of some pragmatic or semantic motivation. FDG takes this moderate position as it tries to explain form from its meaning, but nevertheless allows for form without a pragmatic or semantic motivation.

This dissertation hopes to contribute to this debate by measuring the amount of autonomous form per language. Systematic independent form is not treated as a given, but as a typological variable. If this proves to be a viable approach, that is, if a pattern can indeed be established in formal autonomy, linguistic theory can move beyond a simple dichotomy of formalism versus functionalism and develop a deeper insight into the nature of the relation between form and function.

### *3.2.2 Language acquisition*

Dan Slobin is a key author on the topic of transparency in the L1 acquisition domain and is like Bybee, Haiman and many others a proponent of the competing motivations paradigm (Croft 2003a: 59ff.). According to Slobin (1977), transparency[7] is one of several 'imperatives' to a language user that wants to communicate successfully. Transparent relations are said to be relatively clear, that is, they are highly intelligible structures and therefore advantageous in communication: "[…] there is a tendency for Language to strive to maintain one-to-one mapping between underlying semantic structures and surface forms, with the goal of making messages easily retrievable for listeners" (Slobin 1977: 186). Thus, Slobin expects transparent structures to be favoured in a linguistic situation where intelligibility is under pressure, viz. during acquisition (L1 and L2) and during language contact. Apart from being one of the first to claim that transparency is characteristic for creole languages (cf. Section 3.2.3), he advocates the idea that

[7] His actual term for one-to-one relations between meaning and form is 'mapping transparency'. Slobin (1980) also distinguishes 'metaphorical transparency', by which he means diagrammatic iconicity as described in Section 3.1.2.

children start out acquiring transparent structures as those are easiest to understand, and only then start acquiring the more confusing opaque relations. In Slobin (1985), this hypothesis is tested by an examination of acquisition patterns in various languages (cf. Section 3.4.3). Slobin's ideas are to a large extent formalised in the Competition Model, developed by Bates & MacWhinney (1989; MacWhinney 2005; see section 3.3.2), that also takes competition between possible forms as its starting point.

Regarding L2 acquisition, transparency is discussed most notably by Kusters (2003). The main aim in this work is to measure the complexity of verbal inflection of various languages and relate this to certain sociolinguistic properties of those languages, complexity being defined, controversially, as the difficulty for an L2-learner to acquire the language (cf. Section 3.2.2). According to Kusters, multiple factors determine such relative complexity, and one of those factors is what he calls 'the Transparency Principle, i.e. maintaining one-to-one meaning-to-form relations. Thus, according to Kusters, a language is easier to learn for an L2-acquirer the more it obeys this principle. Phenomena that violate Kusters' Transparency Principle are fusion, allomorphy, fission and homonymy (Kusters 2003: 26ff.) – of which the first two appear on my list of non-transparent features as well, viz. in Section 4.3. Fission is similar to what I call discontinuity and is hence included in this study as well, unlike homonymy, which I have left off the list as it is a feature pertaining to the lexicon rather than to the grammar (cf. Section 2.3). Kusters (2003: 357ff.) finds that languages spoken by a type II community, i.e. a community that has more L2 acquirers then L1 acquirers, tend to obey the Transparency Hypothesis more, which is in line with the idea advocated in this dissertation that language contact leads to an increase of transparency (cf. Section 3.4.2).

### *3.2.3 Creole studies*

Creoles have frequently been claimed to be relatively simple or transparent, e.g. by Kay & Sankoff (1974), Naro (1978) and Slobin (1977, 1980). Seuren & Wekker (1986) are the first to give a more precise definition of semantic transparency and consequently, they evoke the most response. They propose the semantic

transparency hypothesis: the idea that firstly, creole languages are more transparent than other languages, and secondly, that this transparency is due to the sociolinguistic circumstances in which creoles have emerged.

Semantic transparency involves the maximisation of three principles, the first being 'uniformity of treatment of semantic categories' (Seuren & Wekker 1986: 64), which strongly resembles the definition of transparency adopted here. The second principle is called universality and disprefers 'rules that are least language-specific'. This principle should account for the near-absence of morphology in creoles, as an elaborate morphological system would give room for a lot of allegedly haphazard variations. For example in the expression of a comparative, a language is less semantically transparent when it employs inflectional morphology than when makes use of lexical elements, according to this principle. Note that such universality is not part of transparency as defined in this dissertation – in order to be transparent, it is irrelevant whether some form is inflectional or unbound. The third semantic transparency principle is simplicity, i.e. a minimum "amount of processing needed to get from semantic analyses to surface structures, and vice versa" (Seuren & Wekker 1986: 66). This principle is not part of the notion of transparency in this dissertation either.

A critical response to the semantic transparency hypothesis came in 2001 when McWhorter published his second article on the alleged simplicity of creoles. In 1998, McWhorter had written that creoles are simpler languages than non-creole languages. In the (2001) article, McWhorter not only responds to severe criticism in response of his 1998 statements, but also compares his own views to the semantic transparency hypothesis, which he explicitly discards. McWhorter follows Kihm (2000) in interpreting semantic transparency as semantic atomicity: the degree to which semantic atoms are expressed as separate lexical items, rather than in 'unitary equivalents' (McWhorter 2001: 156). For example, the English lexeme *to fetch* is seen as a unitary equivalent of the semantic atoms GO, TAKE and COME, and by unifying these semantic units into one lexeme, the English word is supposedly non-transparent. Showing that Vietnamese, a non-creole language, is very 'atomistic', McWhorter argues that transparency cannot be the defining characteristic of creoles.

According to him, there might be a tendency for creoles to be relatively transparent, but they cannot be defined on the basis of transparency alone.

In fact, two issues are at stake here. Firstly, there is the question whether there is a defining characteristic at all, such that it is able to draw a synchronic linguistic boundary between creoles and non-creoles. Secondly, the question is whether transparency or simplicity constitutes that characteristic, and how those notions should be defined. McWhorter and Seuren & Wekker (1986) agree on the first question, as they believe that creoles are a separate typological class of languages. However, McWhorter argues for a distinction on the basis of simplicity, whereas Seuren & Wekker think creoles are characterised by their transparency. In my opinion, laid out in Leufkens (2013a), there is a tendency for creoles to be relatively transparent when compared to their source languages, but they cannot synchronically be defined by it, since non-creoles can be highly transparent and creoles can be highly opaque.

Without pursuing these questions any further here, what is important is that McWhorter's interpretation of semantic transparency is different from the one that Seuren and Wekker propose. The transparency hypothesis does not involve semantic atoms; in fact, Seuren & Wekker (1986: 63) argue that the notion of 'semantic element' is a highly problematic one, for instance because determining what exactly constitutes a semantic element is very much a theory-dependent decision. Even though McWhorter thus has a different idea on what transparency is, there is overlap in the definition of semantic transparency by Seuren & Wekker on the one hand and McWhorter's simplicity on the other. McWhorter (2001) proposes a simplicity metric that involves counting the number of supposedly equally significant overt distinctions and forms in different areas of grammar, i.e. the number of cross-linguistically rare phonemes, of syntactic rules, of expressed semantic distinctions and the amount of complex inflectional morphology. According to McWhorter, linguistic complexity furthermore increases with the occurrence of suppletion, allomorphy (e.g. declensional classes) and agreement. Some of these features, especially the morphological ones, are deemed non-transparent by Seuren & Wekker too. Note that McWhorter (1998) also lists tone and semantic irregularity of

compounds and derivations as complex features, but these have disappeared from that list in the 2001 article. Seuren & Wekker do not mention those features at all.

Probably due to the overlap in lists of simple and transparent features, and due to their agreement that creoles form a typological class, many creolists equate the ideas of McWhorter and of Seuren & Wekker, despite McWhorter's explicit rejection of the transparency hypothesis. For example, Braun & Plag (2003) interpret the semantic transparency hypothesis as follows:

> In Thomason's words (2001:168), "[m]orphology also tends to be extremely regular when it does exist in pidgins and creoles, without the widespread irregularities that are so very common (to the distress of students of foreign languages) in other languages' morphological systems." In what follows, we will call this 'the semantic transparency hypothesis'. […] This hypothesis is explicitly argued for by Seuren & Wekker (1986) and, in considerable detail, more recently by McWhorter (1998, 200[1]). (Braun & Plag 2003: 81)

Another example is Lefebvre (2001: 321), who "… evaluate[s] the Semantic Transparency hypothesis, as formulated in […] Seuren and Wekker (1986), and, more recently, in McWhorter (1998)." Kihm (2000: 176) remarks that "[…] studies conducted in the "simplicity" or "semantic transparency" paradigm have had a tendency to focus on the verb phrase" – and the list continues. None of these authors take McWhorter's (2001) explicit rejection of the transparency hypothesis into consideration – transparency and simplicity are treated synonymously. Even though the terminological confusion is understandable, I believe transparency and simplicity should be carefully distinguished. Section 3.3 will be devoted to this.

A number of creolists have argued against McWhorter's views on creoles as a linguistically definable class and against his simplicity measure, often pointing their arrows at the semantic transparency hypothesis at the same time. These critiques often target the supposed absence of semantic irregularity as characteristic of the simplicity or transparency of creoles. For example, Lefebvre (2001) and

Kouwenberg & LaCharité (2011) provide ample examples of irregular derivation, lexical idiosyncrasies, allomorphy, semantic ambiguities, semantically vacuous forms and other non-transparent items in a number of creoles, in order to counter the 'semantic transparency myth'. In my opinion, what they actually counter is McWhorter's (1998: 797) statement that creoles typically have semantically regular derivation to a relatively high degree, but not the broader claim that creoles tend to be more transparent than their source languages. Furthermore, a large part of the criticism can be taken away by acknowledging that transparency is a gradual notion, so that instances of opacity in a creole are not necessarily counter-examples to a claim that creoles are relatively transparent. If we weaken McWhorter's claim of simplicity being the defining characteristic of creoles to the statement that creoles are relatively transparent with respect to their source languages regarding a number of features, as Leufkens (2013a) does, the claim is able to withstand much of the arguments that have been raised against McWhorter.

## 3.3 Simplicity

The previous section has given a historic account of how transparency has been discussed in various subfields, most notably in the debate on simplicity that originated in creole studies. In this section, I will dissociate transparency from simplicity because this is essential for a proper understanding of the phenomena at hand and, consequently, for a better understanding of the properties attributed to creoles.

Following the debate in creole studies on the alleged simplicity of creoles, several definitions and metrics have been proposed. Among these, we can distinguish between definitions of absolute simplicity, that is, simplicity of a language system as such, and relative simplicity, the complexity of the acquisition of that system. Section 3.2.1 addresses different notions of absolute simplicity, while 3.2.2 goes into relative simplicity metrics. The difference between transparency and absolute and relative simplicity is explained at the end of the respective subsections.

### *3.3.1 Absolute simplicity*

Linguistic complexity can be defined in terms of the amount of form that is expressed, in combination with its hierarchical depth, i.e. the number of embedded layers. Under such definitions, a language is simpler when less linguistic material is used for a given message and when the structure of the material is more superficial. This type of simplicity metric, that does not refer to acquisition or processing but is an abstract, information-theoretic notion, is called absolute simplicity (Miestamo 2006). Simplicity in this sense largely coincides with the notion of economy as used in the literature on competing motivations (cf. Section 3.2.1).

Langacker (1977) regards simplicity as a combination of signal simplicity ('fewer and shorter units of expression', 1977: 112), perceptual optimality (saliency), constructional simplicity (syntactic depth of linguistic material) and transparency. Similarly, Dahl (2004) splits linguistic simplicity into separate notions such as structural simplicity (a low amount of material at some level of organisation) and system simplicity (simplicity of the mappings from meaning to form, comparable to transparency; 2004: 43). The overall degree of simplicity of a language is, in both Langacker's and Dahl's accounts, the total of these separate measures, of which transparency is one.

McWhorter's (2001) simplicity metric, referred to above, involves a count of the overt linguistic material as well. It adds up the number of phonemes, the number of syntactic rules, the number of obligatorily expressed semantic categories and the number of inflectional morphological distinctions. Parkvall (2008) adopts a similar but broader definition by making a list of features that add form that is considered redundant, as there are languages that can do without it. For instance, Parkvall (2008: 271) includes indefinite and definite articles as complex features, since a language like Russian does well without them. Note that both McWhorter and Parkvall consider a feature to be more complex when it is cross-linguistically rare, which is reminiscent of Seuren and Wekker's (1986) principle of universality.

Metrics of absolute complexity have repeatedly been criticised for a number of reasons. As Aboh & Smith (2009) point out, absolute simplicity deals only with overt marking, being unclear about the interaction between overtness and

covertness and about how covert markers might add to complexity. Another issue in this field is what Kusters (2003) calls 'equi-complexity': the idea that languages as a whole are equally complex, so that a high degree of complexity in one part of a grammar is always paralleled by a low degree of complexity in another domain, e.g. a complex phonology combines with a simple morphology. Few linguists still adhere to this idea since Shosted (2006) and other typological studies showed that there may be a relation, but not a significant inverse correlation between levels of complexity of different domains.

Another problem with metrics of absolute complexity is that the features that are deemed to be complex only occur in languages that have inflection. As for instance Riddle (2008), Gil (2008) and Bisang (2009) point out, isolating languages like the languages of South East Asia may not have these complex features because of the fact that their morphology and phonology do not allow them, but not because their grammars are fundamentally simple. This raises the question if isolating languages can be just as complex as agglutinative languages or in other words, whether inflectional morphemes are more complex than unbound morphemes by definition. At least Dahl (2004), Wurzel (2001) and Shosted (2006) consider inflection to be inherently more complex than juxtaposition, but they are criticised by Riddle (2008), who lists all sorts of phenomena typical for non-creole isolating languages, e.g. classifiers, and shows how these phenomena are complex, at least equally so as bound morphemes in inflectional languages. According to her, whether some meaning is expressed by means of an independent element or by means of a bound morpheme is not relevant for the degree of complexity of the language. On the other hand, Gil (2008) claims that the morphological simplicity of isolating languages does not necessarily go hand-in-hand with complexity of other domains, so that overall, isolating languages do turn out to be simpler.

In my opinion, this entire debate is still too much infected by a sense of complexity as being 'better', i.e. something that awards a language a higher status. A good candidate for a more neutral term than 'complexity' is Dahl's (2004: 103-106, 119-155) notion of maturation, which refers to the time that some features need to develop in a language, given its morphological type. Inflectional morphemes

typically take time to develop in a language, making them mature, but isolating languages may just have well exhibit mature phenomena, for example classifiers. Thus, the notion of maturity allows one to assess the complexity of languages regardless of their morphological type. Note that transparency, too, does not distinguish between agglutination and isolation from a theoretical point of view, since both inflectional and unbound morphemes can relate transparently to elements at other levels.

The various accounts described have slightly different interpretations of complexity, and consequently, different lists of complex features. However, some features keep recurring on lists of simple features, viz.:

- A relatively high amount of syntactic, morphological and phonological rules
- A relatively large phoneme inventory
- Deep structure, i.e. a relatively high degree of syntactic depth, e.g. subordination rather than coordination
- Irregularity / unpredictability
- Opacity
- Cross-linguistically rare phenomena
- Synthesis as opposed to analyticity
- Agglutination as opposed to isolation

Obviously, there is overlap between features on this list and features that I consider opaque – McWhorter's (2001: 163) list of features never found in creoles contains many features that also violate a one-to-one meaning-to-form relation, and so does Dahl's (2004: 114-115) list of 'maturation phenomena'. Both lists in turn overlap with the list of complex features of Parkvall (2008). Again, it is not surprising that accounts of simplicity have been mixed up with accounts of transparency.

However, the two notions can and should be distinguished, as can be demonstrated by means of example (1), a sentence that Turkish children play around with. It consists of two morphosyntactic words, the first of which showing a large

number of morphemes and therefore morphologically complex, according to all measures of absolute complexity described above. However, the word maintains transparency to a high degree, as nearly all morphemes correspond straightforwardly to semantic units. Thus, something that is complex in the absolute sense can be transparent at the same time, proving that the two notions measure different concepts.

(1) çekoslavakyalı-laştır-ama-dık-lar-ımız-dan mısınız?
Czechoslovakian-turn_into-INABIL-PST-PL.OBJ-1PL-among Q-2PL
'Are you among the ones whom we were not able to turn into a Czechlovakian?'

The crucial difference lies in the fact that complexity is always a property of one level, while transparency pertains to interfaces. Absolute complexity metrics measure whether there is a large amount of obligatory inflectional morphology, or a large amount of pragmatic inferences, or a large amount of semantic categories to be expressed, etc. On the other hand, transparency measures how the amount of material at one level relates to the amount of material on another level. As a result, transparency metrics do not have the problems that complexity metrics do have. Firstly, as explained above, the difference between isolation and agglutination becomes irrelevant, since both free and bound morphemes can be transparent. Secondly, the question whether complexity in one domain is inversely correlated with complexity in another (equi-complexity) can be addressed by studying the transparency, i.e. the interface relations, of the language.

In sum, studying both the simplicity and the transparency of a language allows us to measure complexity both horizontally, at specific levels of organisation, and vertically, at the interfaces between those levels. This gives us a complete picture of the organisational efficiency of entire grammars, rather than measuring overt properties of a particular domain only. Furthermore, distinguishing between simplicity and transparency allows us to understand acquisitional data better, as will be explained in the next section.

### *3.3.2 Relative simplicity*

Another way of defining simplicity is by referring to the ease of acquisition of a linguistic system. Simplicity in this sense is known as relative simplicity (Miestamo 2006). The most famous definition of simplicity in this fashion is from Kusters (2003), who defines complexity as the amount of effort needed to learn a language for an outsider, that is, a second language learner with no acquaintance whatsoever with the community speaking the target language (Kusters 2003: 6). This definition is often criticised for not taking into account the L1 of the language learner and the typological distance between the L1 and the L2, which must be of influence on the relative complexity of L2 learning.

Kusters (2003: 21ff.) compares languages on their verbal inflectional morphology, in relation to which three principles determine the degree of relative simplicity: the Economy Principle assures that as little overt form is used as strictly necessary, the Transparency Principle promotes the use of one-to-one meaning-to-form relations, and the Isomorphy Principle makes sure that pragmatic and semantic ordering is reflected in morphosyntactic ordering. All violations of these principles constitute complex features, meaning that they are difficult to learn. Kusters (2003) is unique in defining simplicity with regards to L2 learning, but the features that he lists as being complex are highly similar to the features on lists of absolute complex features that we have seen in the previous section, as Miestamo (2006) correctly signals. Furthermore, there is overlap between Kusters' complex features and the opaque features listed in Chapter 4.

Slobin (1977) goes into the question which features are easy to learn for an L1-learner. Features that Slobin deems easy to acquire are features that satisfy four competing principles, i.e. transparency, processability, economy and expressivity, the latter stating that no form should be void of meaning. Slobin's ideas are taken over in the language acquisition model of Bates & MacWhinney (e.g. Bates & MacWhinney 1989, MacWhinney 2005). Their model is initially called the Competition Model, later the Unified Competition Model, as it models language production as the outcome of a competition between linguistic forms to express a certain meaning. Factors like the ones mentioned by Slobin determine the relative

strength of the different forms, and consequently, which one is optimal in a certain situation. In acquisition, forms are optimal when they have a high so-called cue validity, which combines factors like cue availability and cue reliability. Thus, the model predicts that a feature is relative learnable when it has a high cue availability and a high cue reliability, for instance when it is salient, transparent and frequent (Bates 2005). Forms and features fulfilling such conditions are predicted by this model to be acquired first.

Again, even though accounts of relative simplicity differ on what they claim to be easy, there is large overlap in the lists of allegedly easy features. Such features are:

- In the case of L2 acquisition: similarity with feature in L1
- Highly perceptually salient features
- Highly frequent features
- Highly regular rules
- Highly transparent rules or features

Again, there is an overlap between this list of easily learnable features and the list of transparent features in Chapter 4, and between features easy to acquire and features that are simple in the absolute sense. However, it has often been shown in the literature that absolute complexity and relative simplicity do not necessarily go hand in hand. Of course, a larger number of forms takes more time to learn and in that sense, absolute complexity does result in relative complexity presumably both for L1 and L2 learners. But we know of languages that have a complex inflectional system in the sense that there are many morphemes to be learnt, while their relative complexity is strikingly low: children do seem to learn such languages easily, Turkish being a common example (Aksu-Koç & Slobin 1985: 845ff., see also Section 3.4.3). On this basis, we should not *a priori* equate absolute simplicity and relative simplicity – the two are different notions that might, but need not go together. The same is true for relative simplicity and transparency; these should not be equated *a priori*. With questions of complexity, it is always necessary to consider

not only the amount of form to be expressed or to be learnt, but also how forms relate to their meanings. Therefore, absolute complexity, relative complexity (difficulty) and transparency should be carefully distinguished and studied separately.

## 3.4 Directionality: a continuum of transparency

Previous studies into transparency (notably Slobin 1985, Hengeveld 2011, Leufkens 2013a) have measured the degree of transparency of a variety of languages. In these case studies, some types of language were found to display only few opaque phenomena, specifically contact languages, early stages of L1 acquisition and interlanguages from L2-learners. This suggests, firstly, that there is cross-linguistic variation regarding degree of transparency. Evidence for this idea is discussed in subsection 3.4.1. Secondly, the studies reveal a tendency for transparency to precede opacity, both for languages and for language learners, in line with Haiman (1985: 160), who claims: "As is well known, the development is always in the direction of opacity". Evidence for this direction in language change is discussed in Section 3.4.2, and evidence for this direction in language acquisition is discussed in Section 3.4.3.

### *3.4.1 Implicational hierarchy: the typology of transparency*

In Hengeveld (2011), four natural languages were studied in terms of their transparency. All of them turn out to have some transparent as well as some non-transparent features but the ratio is different, in agreement with the expectation that languages differ in their degree of transparency. Strikingly, Hengeveld (2011b) detects a pattern in the distribution of features. The most transparent language in the sample, viz. Sri Lanka Malay (Nordhoff 2011), exhibits the opaque features of apposition, portmanteau morphemes, phonological adaptations, and influence of phonological weight on constituent placement. Kharia (Leufkens 2011), a South Munda language, shows those features as well, but is a little less transparent than Sri Lanka Malay as it also exhibits grammatical relations and non-parallel alignment of

morphology and phonology[8]. Another language in the study is Quechua (Grández Ávila 2011), which displays the same opaque features as SLM and Kharia, but adds others. Thus, a pattern emerges of non-transparent features that consistently occur in all languages and non-transparent features that are increasingly rare. Apparently, non-transparent features are not distributed randomly over languages; the presence of one feature implies the presence of another.

For example, Hengeveld (2011b) shows that if a language has grammatical gender, it may be presumed to have fusional morphology and cross-reference as well, but not the other way around. Likewise, the presence of fusional morphology implies the presence of cross-reference, but not the other way around. Such relations are known as implicational relations, and together form implicational hierarchies, a type of linguistic universal that was introduced by Greenberg (1966 [1963]) (cf. Section 5.3). Hierarchies are especially suitable to show whether and how linguistic features from different domains are related. Specifically, the implicational hierarchy of transparency that this dissertation aims to establish is able to show the coherence of features from the pragmatic, semantic, morphosyntactic and phonological domains.

### *3.4.2 Diachrony: the transparency of creoles*

Greenberg (1978) argues that implicational hierarchies can often be explained as diachronic pathways. If this is true for a transparency hierarchy as well, we would expect that language change follows the same lines as the distributional pattern discussed in the previous section, viz. that languages start out relatively transparent, having few non-transparent features, and acquire more and more opaque phenomena over time, following the order of an implicational hierarchy. In this scenario, young languages, e.g. creoles, must be more transparent than older languages. Since this study is typological in nature, it will not investigate the truth of the hypothesis that the implicational hierarchy is a diachronic pathway. However, I will discuss the existing evidence in the literature for it.

[8] Cf. Chapter 4 for explanations of these features and why they are considered to be opaque.

Hengeveld (2011b) and Seuren & Wekker (1986) fit into a large body of literature (cf. Trudgill 2011 for an excellent overview and discussion) that states that, in the absence of language contact, language changes in this direction: from simple to more complex, from transparent to more opaque. In this scenario, complex and non-transparent features come into being when originally simple or transparent units evolve and at some point change or lose their form or function. As such, complex and opaque features are 'maturation phenomena' (Dahl 2004: 103-106, see Section 3.3.1), 'historical junk' or 'linguistic male nipples' (Lass 1997: 309). Note that Hengeveld (2011b) not only makes claims about the general direction of linguistic change, but also about the order in which particular opaque features appear. This order is in agreement with the findings of Leufkens (2013a, see below).

Of course, there are many examples of languages changing in the other direction, i.e. from complex to simple or from opaque to transparent. There are two explanations for such changes. Firstly, simplifying change can be due to language contact. It has often been noted, especially in the literature on competing motivations, that language change is the result of competition between factors like transparency and economy. In this line of reasoning, it is likely that a language will increase its transparency when the need for intelligibility is especially high, which is typically the case in language contact situations. Several studies confirm this, three of which I will briefly discuss here. For more examples and structural discussion, the reader is referred to Thomason & Kaufman (1988), Heine & Kuteva (2005), Miestamo et al. (2008), Sampson et al. (2008) and especially Trudgill (2011).

Firstly, in an extensive study, Lupyan & Dale (2010) correlate the absence of certain complex linguistic phenomena to the type of community in which the language is spoken, especially whether the community has many linguistic neighbours and hence, whether there is language contact or not. They find that esoteric communities, that is, communities with little language contact, exhibit a large number of features judged as complex by some, and as opaque by me, e.g. agreement and fusional morphology. This indicates that indeed a language acquires complex and opaque features over time, as long as it develops in relative isolation. In an open community with a lot of language contact, the tendency is the other way

around; such languages lose opaque features. This corroborates the hypothesis that transparency precedes opacity in time, but that language contact may cause a language to change in the opposite direction.

Kusters (2003: 357ff.) finds the same: languages with an increasing amount of L2 learners become simpler as they lose a number of morphological categories, and become more transparent as allomorphy is reduced, at least to some extent. This is corroborated further by Leufkens (2013a), who compares the degree of transparency of four contact languages, i.e. languages resulting from intense contact between typologically diverse languages, to the degree of transparency of their substrate and superstrate languages. Hypothetically, when speakers of typologically distant languages have a need to communicate, they will select transparent forms over opaque forms in order to be maximally intelligible. A language resulting from this kind of contact will therefore be more transparent than or as transparent as its source languages. This is indeed what Leufkens (2013a) finds: all contact languages studied have lost some opaque features with respect to their source languages, while they have not developed any new opaque features that their sub- and superstrates did not have. In sum, the results of Lupyan & Dale (2010), Kusters (2003) and Leufkens (2013a) all indicate that language contact leads to a loss of opaque features.

A second explanation for language change that makes languages more simple or transparent is the existence of a complexity ceiling or opacity ceiling, i.e. a limit on learnability that causes a feature to be lost when it crosses the threshold (cf. Slobin 1977: 192 and 1980: 230). Lightfoot (1979: 129) states: “There seems to be a tolerance level for such exceptional behaviour or ‘opacity’, and when this is reached a radical re-structuring takes place and renders the initial structures more transparent, easier to figure out and ‘closer’ to their respective surface structures.” In this vein, Audring (2009b) argues that the non-transparent and complex phenomenon of grammatical gender can only be maintained in a language when there is enough evidence for learners to acquire it, that is, when it is expressed frequently enough due to extensive gender agreement. A highly complex and non-transparent gender system like that of German is, according to this approach, only learnable because

there is so much evidence for it in the form of agreement markers – without agreement, it would not be learnable and probably lost.

### *3.4.3 Learnability: the difficulty of opacity*

A second type of language that has been shown to be relatively transparent is the interlanguage of language learners, suggesting that transparent structures are acquired before opaque structures. An argument in line with this is that children in all languages overgeneralise rules, which in essence means that they fit a difficult non-transparent structure in their language into an already acquired transparent frame. An example is the regular inflection of irregular past tense verbs in Germanic languages, e.g. children saying 'buyed' instead of 'bought'. This shows that children start out by acquiring transparent structures, and later develop opaque relations.

Furthermore, convincing evidence for the idea that transparency is easy to learn is found in the comparison of the acquisition of languages that differ in their degree of transparency. For instance, Slobin (1977: 190ff.) compares the L1-acquisition of Turkish to that of Serbo-Croatian. Turkish inflectional morphology is strongly agglutinating, very regular, and avoids homophonous forms, allowing a strong one-to-one relation between morphemes and semantic units. Slobin shows that Turkish children acquire this transparent morphology strikingly early: they correctly produce nominal inflection before they are two years old (Slobin 1977: 190). Serbo-Croatian inflectional morphology is much less transparent, as Serbo-Croatian is a synthetic language with a great deal of homonymy. According to Slobin (1977: 191), Serbo-Croatian children are able to correctly produce nominal inflection only when they are approximately 5 years old. Slobin argues that this three-year difference is a result of the high transparency of Turkish inflectional morphology versus the high opacity of Serbo-Croatian inflection.

Slobin (1977) also discusses clause embedding, a domain in which Turkish is highly opaque, while Serbo-Croatian is transparent. Again, the difference in transparency is reflected in acquisition: Turkish children need five years to master complement clauses, while Serbo-Croatian L1-learners can do this at two years of age (Slobin 1977: 191), which is early not only compared to Turkish but cross-

linguistically as well. Again, the transparent structure is learned earlier than the non-transparent one. This of course strongly suggests that, as Aksu-Koç and Slobin (1985: 855) put it, '[c]larity of semantic mapping probably facilitates acquisition'.

Of course, it is dangerous to directly compare the acquisition of Turkish with the acquisition of Serbo-Croatian – the differences could be explained by many other factors than transparency alone. It may be more correct to look for evidence for the language learner's alleged preference for transparent structure language-internally, since if it is true that transparency is acquired before opacity, acquisitional evidence should show that transparent structures are acquired at a relatively young age, while opaque structures are mastered late. Such evidence is abundant. One example of late acquisition of an opaque construction was given already: Slobin (1977) shows that the non-transparent subordination strategies of Turkish are acquired relatively late compared to the transparent verbal inflectional morphology. Another example is the acquisition of grammatical gender in Dutch. Children have acquired the concept of grammatical gender when they are three, but it takes up until seven years of age until they have correctly stored the gender of all nouns (Blom et al. 2008). Moreover, in Dutch, the diminutive suffix, characterised by a unique allomorphy, is not acquired before the age of six (Snow et al. 1980). Egyptian Arabic children master the production of the irregular plural of nouns in their language when they are six (Omar 1973).

Obviously, these examples by themselves are no definite proof either, as there might be examples of non-transparent structures acquired early to contradict these data. A problematic finding, for example, is that children also use non-target opaque structures which have to be unlearnt, such as consonant harmony in Dutch: children say [bup] rather than the target form [buk] 'book' during early phonological acquisition (Gillis 2000: 157). Such consonant harmony can presumably be explained by articulatory constraints, showing that acquisition is always an interplay between multiple factors. Thus, it is difficult to isolate the influence of transparency on language acquisition, an endeavour that falls outside the scope of this typological study. However, see Kusters (2003: 53ff.) for an overview of several studies into L1

acquisition that do support the hypothesis that transparency precedes opacity in language learners.

Finally, it is important to note that there is reason to believe that within the class of non-transparent features, there is again a particular order of appearance of features that is identical to the implicational hierarchy and diachronic pathway described in the precious sections. The opaque features of phonological assimilation and apposition turn out to be acquired relatively early, that is, earlier than other non-transparent features. For instance, Turkish vowel harmony, a non-transparent assimilation process, is acquired by Turkish children at very early ages, i.e. around 15 months (Aksu-Koç & Slobin, 1985: 845). This indicates that non-transparent features can be ordered: children start with acquiring transparent features, possibly at the same time as they acquire relatively lightly non-transparent features, ending years later with the acquisition of the most severely opaque phenomena.

The acquisition order of non-transparent features matches the typological data found in Hengeveld (2011b). This parallel direction in diachrony and in language learning is striking: both in languages and in L1-learning, transparency precedes opacity, even showing the same ordering of particular features. Slobin (2004) argues that such a parallel should not *a priori* be assumed, since it is not necessarily the case that languages develop identically in communities and language learners. However, the evidence given in Section 3.4.2 and 3.4.3 suggests that this is the case for this particular ordering of transparency and opacity.

# Chapter 4
# A list of non-transparent features

In this chapter, I will list all linguistic phenomena in which a one-to-one relation between the four levels of linguistic organisation distinguished in FDG is somehow violated. This list is created by checking possible combinations between primitives at these levels (Hengeveld 2011a). The non-transparent features are categorised according to the division introduced in Section 2.4 between redundancy, discontinuity, fusion and form-based form. After discussing the features in the respective categories in sections 4.1, 4.2, 4.3 and 4.4, Section 4.5 will provide a summary.

The list aims to be comprehensive, but some arguably non-transparent phenomena do not appear in it. Firstly, as explained in Section 2.3, features are excluded that pertain to the domain of the lexicon rather than to the grammar. Homonymy, polysemy and synonymy, as well as semantic irregularity, undoubtedly constitute cases of non-transparency: one formal element relates to multiple meanings (homonymy, polysemy, semantic irregularity), or multiple formal elements relate to one meaning (synonymy). Nonetheless, these features will be excluded from this study, since it aims to study transparency in grammar.

Another reason to leave out an opaque feature from the study is that it is near-universal and as such not interesting for the implicational hierarchy pursued. For instance, nominal apposition (Section 4.1.2) is to my knowledge allowed in all languages of the world and can therefore not be of service in discriminating the degree of transparency of a language. Such features will be discussed and exemplified in the list below, but they will not be studied.

Furthermore, there is a more practical reason to exclude some non-transparent features from the list, which is an insufficiency of previous research on the topic. For example, typological data about modal concord (Section 4.1.7) are so scarce that new data would have to be obtained for almost all languages in the sample; an endeavour that is not feasible within the time limits of this research.

Therefore, such features will be discussed in this chapter, but not taken into account in the actual study.

## 4.1 Redundancy

The category of redundancy comprises all one-to-many relations between units at different levels: one pragmatic, semantic or morphosyntactic primitive, i.e. a frame, lexeme, operator or function, relates to multiple semantic, morphosyntactic or phonological units. Thus, at least one of those units is redundant.

### *4.1.1 Multiple expressions of pragmatic information – IL-ML, IL-PL*

Pragmatic information, for instance a pragmatic function like Topic or Focus, can be expressed by special markers, e.g. discourse markers, question particles or interjections. Another frequent strategy for marking pragmatic functions or illocution is intonation: declaratives and questions are in many languages distinguished by means of intonation patterns. When two or more of such devices are used in one utterance, a one-to-many relation between pragmatic units and morphosyntactic and phonological units results. For example, when a question is expressed both by a question particle and by a specific intonation pattern, there is a relation between one pragmatic unit (Illocution) and two phonological units (a Phonological Word and an Intonational Phrase).

Reference grammars usually give information about the morphosyntax of question formation, but not about their phonological properties – only in few grammars a description of the intonation of interrogatives is given. Because of this lack of data, I cannot study this feature sufficiently, and I therefore exclude it from the study.

### *4.1.2 Nominal apposition – IL-RL*

Appositional constructions are combinations of two or more nominal elements that refer to the same entity. According to criteria by Keizer (2005), one of the elements modifies the other, but it is not an adjective and no linking element is used. A subset

of apposition is that of close appositions, in which the two nominal elements form one intonational unit, e.g. ‘my friend Manfred’, ‘the colour red’ and ‘the word transparency’, but not ‘the city of Rome’ as this includes a linking element. An example of an apposition that consists of two intonational units is ‘Manfred, my friend’.

In FDG terms, both nominal elements in an appositional construction constitute Referential Subacts: they are acts of referring to an entity, performed by the speaker. The elements are co-referential, so both Referential Subacts are used to refer to the same entity, an Individual, meaning that there is a many-to-one relation between Referential Subacts at the Interpersonal Level and an Individual at the Representational Level.

Nominal appositions are to my knowledge possible in all languages – I have never come across a language that shows any restrictions on this. Therefore, comparing languages on the presence of apposition of nouns would not make sense and even though apposition of nouns is non-transparent, I will not study this feature.

### *4.1.3 Clausal agreement or cross-reference – IL-RL*

Agreement is defined by Steele (1978: 610) as “systematic covariance between a semantic or formal property of one element and a formal property of another”. In other words, a semantic or grammatical property of one unit (the controller), whether overtly expressed or not, is expressed on some other unit (the target). Agreement can take place in various domains. Corbett (2006: 21) considers the phrase to be the most canonical domain, i.e. a standard domain for agreement, which prototypically involves agreement between nouns and their modifiers (cf. Section 4.1.4). The next most common domain for agreement is the clause, in which in many languages agreement takes place between (an) argument(s) and the predicate. Other candidates for an agreement analysis violating some of the properties that Corbett lists as canonical are for instance negative concord, modal concord and sequence-of-tenses. Many consider these not to be cases of agreement, but they do conform to Steele’s (1978) definition given above and are therefore discussed below. Some (e.g.

Audring 2009a) also consider the extra-clausal expression of features, e.g. on pronouns, as agreement, but I will not adopt that view in this study.

In FDG, agreement is seen as a copying operation at the Morphosyntactic Level. A property of an argument is copied to another unit within the clause or phrase, or an operator is copied to an operator slot in the same or another clause. Such copying pertains strictly to the Morphosyntactic Level and is therefore a non-semantic procedure. As a consequence, the copied element does not have a semantic value of its own: it is a purely morphosyntactic unit without a counterpart at a higher level. FDG strictly distinguishes between morphosyntactic copying, resulting in such empty elements, and the situation in which one semantic unit is expressed multiple times, while both expressions maintain their semantic value. This means that FDG distinguishes predicate-argument agreement (morphosyntactic copying of properties of an argument to the predicate) from so-called cross-reference (i.e. a 1-to-2 relation between a referent and multiple morphosyntactic units with semantic value; cf. Hengeveld & Mackenzie 2008: 350). A similar distinction is made in Generative Grammar. In that framework no unified account of doubling phenomena exists, but rather two groups of approaches. For instance, in the multiple expression of negation (negative concord), two kinds of explanations are proposed. One is the treatment of negative concord as syntactic agreement, advocated by Zeijlstra (2007a) among others, while in the approach by De Swart & Sag (2002), negative elements all have a semantic value. Thus, the problem of the analysis of doubling is recognised in FDG and in GG, and is approached in a similar fashion within both theories.

The distinction between agreement on the one hand and multiple expression of a semantic unit on the other is notoriously hard to draw. The crucial difference is the semantic value of the elements, but this is difficult to establish, especially if one wants to apply theory-independent criteria. A rule of thumb is that if an element can occur by itself, that is, without an overt controller, it must have semantic value of its own. Hence, the criterion is obligatoriness: if one of the two morphosyntactic elements can be left out, it cannot be a copy of the other. It follows

that we can only speak of agreement when both elements are obligatory, as for instance in the case of subject-verb agreement in French (see below).

This criterion, however, is not watertight, since the multiple expression of a semantic unit may also be obligatory. This is the case in Bantu languages that have nominal classification systems. For example, the class markers in the Swahili examples (1) and (2) are all obligatory.

Swahili – Corbett (1991: 43)

(1) ki-kapu ki-kubwa ki-moja ki-lianguka
VII-basket VII-large VII-one VII-fell
'One large basket fell.'

Swahili - Aikhenvald (2000: 395)

(2) ki-faru m-kubwa
VII-rhinoceros I-big
'A big rhinoceros.'

Following the obligatoriness criterion, we should analyse this as agreement: the class feature of the noun is copied at the Morphosyntactic Level to the other elements. All class markers except the one on the controller noun, then, must be semantically empty copies. However, in example (2), the noun 'rhinoceros' is from class VII, a class containing inanimates, but agreement is with class I, a class referring to humans, thus personifying the rhinoceros. This shows that a Swahili class marker is sufficient to alter the meaning of a noun, hence, to contribute semantics – in this case, attribution of humanity to an animal – to the sentence. This falsifies an analysis of Swahili class prefixes as empty copies of the noun prefix. Of course, one could still argue that the markers in (1) are semantically empty and the ones in (2) are not, but the main point here is that obligatoriness of markers does not guarantee that markers are semantically empty – they may still have semantic value.

Another problem with taking obligatoriness as a criterion is that agreement may apply between a target and a controller that is present contextually rather than

overtly in the sentence. This is most notably the case in pro-drop languages, in which a controller argument can be covert as long as it is clear from the context to whom or what is being referred, for instance because it is the topic of the conversation or because it is mentioned in a previous sentence. In that case, agreement may be obligatory, but this is not visible because there is no explicit overt controller. Hengeveld (2012) models this as agreement between a target and a controller present in the Contextual Component, as opposed to syntactic agreement that holds between a target and an overtly present controller.

Both these types of double expression of (properties of) arguments are non-transparent, since double expression is redundant by definition, whether occurring overtly in all sentences, or only in part of the sentences, as in pro-drop languages. However, there is of course a gradual difference between pro-drop languages and languages with syntactic agreement in the sense of Hengeveld (2012). In other words, a pro-drop language is fundamentally more transparent than a language in which arguments cannot be dropped, since redundancy is only present overtly in some clauses, rather than in all of them. To measure this difference in degree of transparency, I will indicate for each language whether the argument is obligatorily overt, in which case I speak of clausal agreement, or whether the argument can be implicit, in which case I will speak of cross-reference.

This raises another problem, namely that the distinction between pro-drop and non-pro-drop is not as clear-cut as is sometimes thought. In fact, languages show variation in the extent to which arguments are required to be expressed explicitly, a variation that Bickel (2003) quantifies by means of his notion of 'referential density'. According to the link between obligatoriness of an argument's explicitness and the extent of opacity, it would be ideal to take the referential density of a language as a measure of the degree of opacity of a language with respect to agreement. However, as will be explained in Section 5.1, I will not quantify features in a gradual but in a binary way, for reasons of time. Thus, I will draw a distinction between clausal agreement, i.e. agreement between an obligatorily overt argument and marking on the predicate, and cross-reference, in which there is double expression of (properties of) the argument as well, but with a possibly implicit

argument. Note that the definition of this distinction is slightly different from the one that is given by Hengeveld & Mackenzie (2008: 350), since that is based on the obligatoriness of agreement as such, while the distinction here is based on the 'omissability' of the controller.

A typical example of a language in which both argument expression and verbal inflection on the basis of argument properties are obligatorily overt is French: *je pens-e* '1SG think-1SG', *nous pens-ons* '1PL think-1PL'. The person and number specifications of the subject argument are obligatorily expressed by an independent NP and on the predicate. A typical pro-drop language, in which independent expression of the argument is optional, is Turkish, e.g. *(ben) gitmedim* '(I) did not go' (Lewis 1978: 68).

In sum, I will consider a language to be opaque with respect to this feature if it has clausal agreement, meaning that arguments are obligatorily present, as in French, or if it exhibits cross-reference, meaning that arguments are optionally present in the sentence, as in Tukang Besi. A language qualifies as transparent if independent expression of the argument and argument marking on the predicate are mutually exclusive. If a language has no argument marking apart from expressing the argument independently, there can be no double expression and the feature does not apply.

### *4.1.4 Phrasal agreement – RL-ML*

Another canonical type of agreement is noun-modifier agreement, i.e. the expression of a formal property of a noun on its modifier(s), whether that property is overtly expressed on the noun itself or not. This occurs for instance in Italian, as illustrated by example (3).

D. Boeke (personal communication, October 21, 2014)

(3) a. un-a bell-a ragazza
INDEF-SG.F pretty-SG.F girl(F).SG
'another woman'

b. un-Ø bel-Ø ragazzo
INDEF-SG.M pretty-SG.M boy(M).SG
'another man'

In this example, the gender and number of the noun determine the inflection on adjectives and determiners. Other properties of nouns that frequently trigger this type of agreement are definiteness and case. Other common targets include demonstratives and relativisers. I will assess for each language whether it shows copying of properties of the noun to its modifiers and if so, to what extent this is obligatory.

Again, a difference can be made between agreement between an overt controller and its target, and between an implicit controller (present contextually but not syntactically) and its target. In the phrasal domain, the controller of agreement is the noun, so that the equivalent of pro-drop in clauses would be implicitness of the head noun. Languages indeed show variation in the omissability of nouns from phrases (cf. Gil 2013), but this is not as well studied as pro-drop. Thus, most reference grammars do not make mention of this feature, which makes it practically unfeasible to study the phrasal equivalent of cross-reference. Hence, for each language in the sample, I will indicate whether there is noun-modifier agreement, as in Italian, or not, but I will not test whether the noun is obligatorily explicit or not.

### *4.1.5 Plural concord in noun phrases containing a numeral – RL-ML*

The semantic property of number can be expressed lexically, by means of a quantifier or numeral, and grammatically by marking it on the unit to which it applies, for instance by adding a plural suffix to a noun or predicate. When a numeral 'two' or higher modifies a noun, it is redundant to mark plurality grammatically as well; for instance in the phrase *five elephants*, the numeral already sufficiently expresses the plurality of the Individual denoted. Nonetheless, it is obligatory in Standard English to use the plural suffix *-s*, thus creating a non-transparent relation between the semantic number operator and its morphosyntactic

expression in an Adjective Word and an affix. I will refer to such redundant marking of plurality as 'plural concord'.

Languages deal with number marking in different ways. There are languages in which number is not expressed grammatically at all, so that plural concord is impossible. In other languages, only a subset of nouns is marked for number, for example animate nouns (Payne 1997: 96). Yet other languages only express number grammatically when the noun is individuated, that is, when it is considered a count noun or collective noun. If the noun is not marked for number, it is a non-individuated mass noun or concept noun (cf. Rijkhoff 2002: 104ff.). The distinction is marked explicitly for instance in Chukchi, where *mane-lɣ-ə-n* 'money-SG-EV-3.SG, one coin' and *mane-t* 'money-PL, several coins' are individuated and hence marked for number, while *mane-man* 'money-RDP, money in general' is the non-individuated, non-numbered form (Dunn 1999: 64). The stem *mane* as such is not inherently a mass noun or a count noun: presence or absence of the number marker determines what it is. In such languages, plural concord is possible, but since it is optional to express plurality, plural concord is optional too.

In a language in which number can be marked overtly, it is non-transparent to do so in the presence of a lexical number expression such as a numeral or a quantifier, as the English example *five elephants* demonstrated. An example of a language that is transparent with respect to this feature is Turkish, in which the plural suffix *-lVr* is mutually exclusive to numerals, e.g. *kırk harami*, literally meaning "forty thief" (Lewis 1978: 26). Some languages exhibit a mixed system, for instance Sudanese Arabic, in which nouns have to be marked for plurality when combined with a numeral from 10 up to 20, while other numerals show no plural concord. All such languages will be counted as having plural concord.

### *4.1.6 Negative concord – RL-ML*

Negation of a sentence can be expressed by various means, e.g. by inflection or by an independent negative marker (cf. Dryer 2013). In the negation of indefinites, several languages evoke negative indefinite pronouns and quantifiers, such as the English *nobody* and *no-one*. Other elements, called Negative Polarity Items (NPI's),

do not have negative semantics by themselves but are only grammatical in the presence of a negative element or in a question context, as for example English *anybody*.

A combination of regular sentential negation and a negative indefinite pronoun and/or NPI should, applying basic propositional logic, result in a positive reading of the sentence. In fact, it does so in many languages, which are called double negation languages. In other languages, negation of indefinites is expressed by means of negative existentials ('There is no person that I saw') or by means of positive indefinites ('I did not see someone'). Such means of negating indefinites are all transparent, since one negative operator always equals one morphosyntactic element.

However, there are languages in which the combination of sentential negation and a negative indefinite results in a negative reading. This phenomenon is called negative concord (e.g. Zeijlstra 2007a) and languages that exhibit it are called negative concord languages, e.g. non-standard English *I haven't seen nobody*. These come in different types – different classifications can be made according to the position of the negative elements, the context in which negative concord does or does not occur, etc. Negative concord appears to be a clear case of non-transparency, as there is one semantic negation that corresponds to multiple morphosyntactic negative elements. However, there is reason to believe that negative concord is not strictly speaking opaque, but rather a case of emphasis and concurrent semantic weakening. The problem is that it is often very difficult, if not impossible, to determine whether the lexical negative element, i.e. either the negative indefinite or NPI, has negative semantic value of its own (cf. Haspelmath 1997: 193ff.).

Firstly, there are languages in which sentential negation and the negative indefinite obligatorily co-occur. In that case, it is impossible to establish the independent negating ability of the indefinite pronoun: if *nobody* is always accompanied by *not*, there is no way to establish whether *nobody* has negative or affirmative force by itself. Secondly, languages exist in which an indefinite, in the absence of sentential negation, has negative meaning in some contexts but not in others. For example in Turkish, the indefinite *hiç kimse* has negative semantics when

it is used as an answer to a question, but it shows positive polarity if used as the subject of a question, cf. (4).

Turkish – Haspelmath (1997: 196)

(4) a. A: kim gel-di? B: hiç kimse
who come-PST.3SG INDEF anyone
'Who came?' 'Nobody.'

b. hiç kimse gel-di mi?
INDEF anyone come-PST.3SG Q
'Did anyone come?'

Cases like these make it impossible to establish whether the indefinite is inherently negative or not. This problem does not exist in all languages, but at least for some languages, the transparency of the negation of indefinites is not assessable.

Moreover, if we follow the analysis of Kiparsky & Condoravdi (2006), who study negation in Greek historically, the non-transparency of negative concord turns out to be even harder to measure. Their finding is that Jespersen's cycle in Greek is motivated by a continuous reinvention of an emphatic negation construction. A non-emphatic negation construction is available, but speakers tend to stress this negation by adding a 'minimiser' (e.g. 'not even one', 'not a tiny bit') or a 'generaliser' (e.g. 'not ever', 'no thing whatsoever'). When this emphatic negation becomes frequent, it is increasingly reduced phonologically into one morphosyntactic unit (cf. *never*, *nothing*). Furthermore, as in prototypical cases of grammaticalisation, the semantics of the emphatic negation are bleached so that the emphatic effect is lost and the construction becomes a standard negation, thus creating a new need for an emphatic negation.

In this scenario, that I adopt in this study, two negation constructions appear alongside each other: a regular one and an emphatic one. In the emphatic one, a negative indefinite may combine with sentential negation, but since there is a pragmatic effect of the multiple negation, this is not opaque. In the regular, grammaticalised negation construction, the negative indefinite is bleached and

therefore no longer an inherently negative element – as far as that can be determined. In neither case is negation opaque, so that a non-transparent construction only exists in the period that the indefinite has lost its pragmatic emphatic effect, but not yet its inherent negativity. It is impossible to establish when this situation obtains, rendering it impossible to include negative concord in this study.

#### *4.1.7 Modal concord – RL-ML*

Modal concord refers to the phenomenon of two modal expressions corresponding to one semantic modal unit only (Zeijlstra 2007b). For instance in (5), we find two modal expressions: *mandatorily* is a modifier of deontic modality on the layer of the State-of-Affairs, while *must* is a lexical element operating on the same layer.

American English - Zeijlstra (2007b: 317)

(5) Power carts must mandatorily be used on cart paths where provided.

If both modal elements were to apply as they do in isolation, the reading of the sentence would be 'it is obligatory that it is obligatory that power carts are used on cart paths where provided'. This is not the case, since in the default interpretation of this sentence, there is only one semantic deontic modality. The conclusion has to be that there is one semantic modal operator, corresponding to two morphosyntactic modal expressions – a non-transparent one-to-many relation. Note that there is no emphatic or other pragmatic effect here, so that this really is a non-transparent relation.

Modal concord as in (5) appears to be possible in many languages. However, there is virtually no typological research on modal concord, nor is it a topic covered in most reference grammars. Finding out whether the languages studied here allow it would be extremely time-consuming. This feature will therefore be left out of the current research.

*4.1.8 Temporal concord and tense copying – RL-ML*

Temporal information can be expressed lexically, by means of adverbials or adverbial phrases, and in many languages also grammatically by means of tense marking on the predicate or elsewhere. When two means of expression are combined, there is again an overlap in meaning and hence a non-transparent relation between a semantic temporal operator and its morphosyntactic expression. The combination of a lexical time marker and a grammatical one will in this dissertation be called temporal concord, as for example in the English sentence *Yesterday I was smiling*. Both *yesterday* and *was* indicate that the event occurred in the past. There is hence a one-to-two relation between meaning and form. Since temporal concord is allowed in all languages that I have encountered so far, I will assume this to be a near-universal feature and not investigate it further in this research.

A second type of multiple marking of time reference is tense copying, more generally known outside FDG as sequence of tenses or *consecutio temporum* (Comrie 1986). This involves the copying of the tense operator of a main clause to the tense slot of an embedded clause. In languages without it, an embedded clause, if tense is marked in it at all, usually contains a relative tense that takes the tense of the main clause as its deictic centre. For instance in (6), the embedded tense shows that at the time of speaking the action of dancing was in the present.

Russian - Comrie (1986: 275)

(6) tanja skaza-l-a, čto ona tancu-et

T. say-PST-F that she dance.PRS-3SG

'Tanja said that she was (litt.: is) dancing.'

However, in English and other languages with a sequence-of-tense rule, the tense of the embedded verb is 'backshifted' when the main verb has a past tense: an absolute-relative tense is used that relates the event described to the past tense of the main clause. For example in the English translation of (6), *Tanja said that she was dancing*, the embedded verb carries a past tense, rather than a present tense, because the main clause morphosyntactic operator 'past' is copied to the embedded clause

(cf. Leufkens 2013b). Hengeveld & Mackenzie (2008: 351) analyse this as an agreement-like operation called operator copying, since it is a morphosyntactic procedure that does not have a pragmatic or semantic motivation. Since the operator copy in the embedded clause has no pragmatic or semantic counterpart, a language with a tense copying rule will be considered opaque in this respect. Note that if a language has no syntactic embedding, no finite subordinate clauses, or no grammatical tense expression at all, this feature does not apply.

### *4.1.9 Spatial concord – RL-ML*

Spatial information can be expressed lexically by an adverb, adverbial phrase, adposition or by an adpositional phrase, or grammatically by means of case marking or a grammatical adposition. Using both means in one utterance results in redundancy, for instance in the Khwarshi example (7).

Khalilova (2009: 77)

(7) […] lac'alas podnos karavatiλ gił gul-un
food.GEN plate bed.SUB under put-PST.UW
'She put the plate under the bed.'

The subessive case of 'bed' expresses that something is below it, while the same (or at least an overlapping) meaning is expressed by *gił* 'under'. Hence, the combination is redundant: a one-to-two relation between a locational operator and its morphosyntactic expressions. Since spatial concord occurs in all languages with grammatical and lexical spatial marking that I have encountered so far, I will assume this is a near-universal feature of such languages and not investigate it further in this research.

### *4.1.10 Summary redundancy features*

The following features from the category of redundancy (one-to-many relations) will be studied in the languages in the sample: agreement and cross-reference in the clause, agreement and concord in the phrase, plural concord in noun phrases

containing a numeral, and tense copying. Features that are non-transparent, but nonetheless excluded from the study due to their near-universality are nominal apposition, temporal concord and spatial concord. Negative concord, multiple expression of pragmatic information and modal concord are left out because they cannot be assessed properly, due to theoretical or practical matters.

## 4.2 Discontinuity

The subcategory of discontinuity includes violations of domain integrity, which result in incomplete morphosyntactic or phonological units or fragments that cannot be used independently but require the presence of another unit. This gives rise to relations between a single meaning unit and a morphosyntactic unit that is split-up into multiple parts, involving a non-transparent relation between one meaning unit and two or more morphosyntactically incomplete units.

### *4.2.1 Extraction and/or extraposition – RL-ML*

Elements that belong together at the pragmatic or semantic level can be realised separately in morphosyntax, for example when a modifying constituent is expressed non-adjacently to its head. If the modifying phrase or clause is realised near the end of a sentence, this is called extraposition, which is commonly the result of so-called heavy shift: a morphosyntactic principle that prefers complex elements to appear near the end of sentences while simple elements appear at the beginning (cf. Section 4.4.4). Extraction, a process complementary to extraposition in which the modifying element appears to the left of its head, is usually the result of a pragmatic principle, e.g. topicalisation. An example is given in (8), in which the square brackets indicate the extraposed PP in (8b) and the extracted PP in (8c).

English - Van de Velde (2012: 433)

(8) a. We have several important books about global warming in stock.

b. We have several important books in stock [about global warming].

c. [About global warming] we have several important books in stock.

Both extraposition and extraction create a mismatch between levels, because something that is one unit of meaning is not realised as one unit of form (Van de Velde 2012: 433). For example in (8), at the semantic level, the sentence contains the Individual 'books', having the property of 'being about global warming'. This property functions as a modifier of the Individual, and is hence part of the semantic entity. However, morphosyntactically, the Individual and the modifier are not adjacent. This violates parallelism between meaning and form: one semantic unit is realised as two morphosyntactic units. Note that it is not the violation of iconicity (cf. Section 3.1.2) that is considered opaque here, but the realisation of one unit in multiple fragments.

Both extraction and extraposition are frequently analysed as movement processes, assuming that the NP is first realised as one unit at some underlying level, after which part of the NP is moved somewhere else because of a pragmatic or syntactic rule. This is the common way of thinking in the generativist tradition, but it is not the view adopted here, since FDG does not recognise movement processes. As Van de Velde (2012) argues, the morphosyntactic NP in a sentence like (8) is not one unit at an underlying level, but is generated as two separate parts from the beginning, that is, during Encoding. Note that this debate is not relevant for the current study, since the generation of the morphosyntactic unit does not influence the fact that at the Morphosyntactic Level, we find two incomplete units, corresponding to one unit at the Representational Level.

English allows for extraposition rather freely, but other languages, such as Japanese, do not. As shown in example (9), it is ungrammatical in Japanese to realise a modifying clause at a position non-adjacent to its head.

S. Iwasaki (personal communication, October 17, 2013)

(9) a. akai seetaa o kita otokonoko ni kinoo atta
red sweater ACC wear boy DAT yesterday met
'I met that boy, who was wearing a red sweater, yesterday.

b. *akai seetaa o kita kinoo otokonoko ni atta
red sweater ACC wear yesterday boy DAT met
'I met that boy yesterday, who was wearing a red sweater.'

Languages that allow extraposition, extraction or both are considered opaque, since one semantic unit corresponds to two non-adjacent morphosyntactic fragments. Languages that avoid this, are considered transparent. Of course, this is not an all-or-nothing feature – languages may differ in the degree to which they allow extraposition and extraction. A language will qualify as non-transparent regarding this feature if I find one or more example of extraposition or extraction. If I find no examples of either extraposition or extraction at all, I will assume their existence in the language is at most marginal and score the language as transparent for this feature.

### *4.2.2 Raising – RL-ML*

Argument raising occurs when an argument semantically belonging to an embedded clause behaves syntactically as an argument of a main clause. For example, sentence (10a) consists of a main clause with a pronominal dummy subject and an embedded clause, with *the horses* as its subject. In (10b), the subject of the embedded clause appears syntactically as the subject of the main clause. Thus, the argument of the predicate *ill* is morphosyntactically distanced from that predicate, even though predicate and argument form a single unit at the semantic level.

(10) a. It seems that the horses are ill.

b. The horses seem to be ill.

Argument raising creates non-parallelism between semantics and syntax and is therefore non-transparent. Hence, a language will be regarded opaque with respect to this feature if argument raising is allowed, and transparent when it is not allowed.

### *4.2.3 Circumfixes – RL-ML*

Circumfixes are discontinuous affixes consisting of two parts phonologically, but relating to one unit at the semantic and pragmatic levels. Isolating languages may exhibit circumpositions rather than circumfixes, i.e. phonologically independent elements that are placed 'around' other elements and are ungrammatical if the other element is absent. A classic example of a circumposition is the French negator *ne ... pas* 'NEG', which has disappeared in colloquial French. The semantic negation corresponds to two morphosyntactic fragments, which means that there is a non-transparent relation. A language with one or more productive circumfixes or circumpositions will be considered opaque, while a language without them is transparent in this respect.

### *4.2.4 Infixes – RL-ML*

Infixes are morphosyntactic units that are inserted inside Morphosyntactic Words. They are not discontinuous elements themselves, but create discontinuity in their hosts. It has been claimed that mesoclitics, i.e. the clitic variant of infixes, also exist (e.g. Van der Leeuw 1997), but this is highly controversial. An example of an infix in Kharia is the causative marker *<[(o)ʔ]>* or *<[(o)ʔb]>*, e.g. *botoŋ* 'fear' versus *boʔtoŋ* 'scare' (Peterson 2011: 231). If a language allows for infixation, it is considered non-transparent with respect to this feature. Note that incorporation of nouns or verbs into other elements, a common feature in for instance West Greenlandic, is not counted here, since it does not necessarily create discontinuity in the host elements.

### *4.2.5 Non-parallel alignment – ML-PL*

Another type of non-parallelism between levels pertains to the interface between the Morphosyntactic Level and the Phonological Level. Alignment at those levels, i.e. their organisation into phrases and words, should be parallel in order to be transparent, but as shown in example (11), this is not always the case.

Hengeveld & Mackenzie (2008: 18)

| (11) | [ik | wou] | [dat | hij] | kwam |
|---|---|---|---|---|---|
| | /kʊɑu | | dɑti | | kʊɑm/ |
| | I | want.PST | COMP | he | come.PST |

'I wish he would come.'

At the Phonological Level, the phonological units /dɑt/ and /i/ are combined into one Phonological Word, and so are /k/ and /ʊɑu/. However, the corresponding morphosyntactic units cannot be combined with each other. Such non-parallel alignment between morphosyntax and phonology is considered opaque in this dissertation.

While the majority of grammars contains information on phonemes, only few provide information on larger phonological units, prosody, and on the alignment of morphosyntax and phonology. Therefore, it is virtually impossible to determine whether languages have non-parallel alignment or not and this feature is excluded from the study.

### *4.2.6 Summary discontinuity features*

The non-transparent features that will be studied are extraposition and extraction, raising, circumfixation, and infixation. Since data are insufficient for a good survey, non-parallel alignment is left out of the study.

## 4.3 Fusion

In this section, all relations between more than one pragmatic or semantic unit and one formal unit are discussed. The word 'fusion' is not meant to imply that there used to be separate morphemes that have diachronically fused into one inseparable unit – this may be the origin of a fusional unit, but it need not be the case.

### *4.3.1 Cumulation of TAME and case – RL-ML*

Within the category of morphemes expressing more than one unit of meaning, Hengeveld (2007) distinguishes between cases of fusion between a lexical and a grammatical meaning, i.e. stem alternation (discussed in Section 4.3.2), and the joint expression of multiple semantic categories in a grammatical unit, which he calls cumulation. Cumulation is more commonly known as fusional morphology, and the resulting grammatical morphemes are called portmanteau morphemes. A language that predominantly makes use of fusional morphology is said to be a fusional language.

To avoid confusion, it is necessary to specify exactly what I mean by notions like 'word', 'morpheme' and 'grammatical unit', as such terms are used differently by different authors. For reasons of consistency, I will once more make use of the terminology of Functional Discourse Grammar and indicate how that relates to other more well-known terminologies. Note that names of FDG primitives are capitalised in this section, while general, theory-independent terms are not, e.g. a 'Morphosyntactic Word' is an FDG unit, but a 'word' refers to a theory-independent item. For a more elaborate description of morphological items in FDG, cf. Hengeveld (2007).

First of all, a Lexeme in FDG is a unit at the Representational Level, that is, a semantic primitive (cf. Section 2.1; Hengeveld & Mackenzie 2008: 400). While other frameworks often see a lexeme as a mapping of meaning onto form, in FDG a Lexeme is a unit of meaning only. The formal counterpart of the Lexeme at the Morphosyntactic Level is the Morphosyntactic Word. In many cases, there is a one-to-one relation between Lexemes and Morphosyntactic Words, but this need not be

the case, as a Lexeme may relate to several words (e.g. in the case of idioms) or to none, when it relates to a Stem without inflection, which is not a complete Word. A Morphosyntactic Word can also occur without relating to a Lexeme, because if it relates to a function, operator or to no higher level unit at all, it is a Grammatical Word. Grammatical Words are Morphosyntactic Words that express a pragmatic or semantic function or operator, but do not express a Lexeme, for example a copula or an article. Note that clitics are also considered Grammatical Words (cf. Section 4.4.5). 'Grammatical unit' is a cover term for Affixes and Grammatical Words.

Other morphosyntactic units distinguished in FDG are the Root, the Stem, and the Affix. A Lexeme cannot be expressed as an Affix, since an Affix is defined in FDG as a unit without lexical meaning – it can correspond to a function or operator, but not to a Lexeme. The Morphosyntactic units Stem, Root and Affix are types of Morphemes (Hengeveld & Mackenzie 2008: 404). Whether a Morpheme is a Stem, Root or Affix depends on its lexical content and its ability to occur independently. A Morpheme in FDG is the smallest morphosyntactic unit that can express a unit of meaning (a Lexeme, function or operator), and a Morphosyntactic Word consists of one or more Morphemes. In morphological theory, a Morpheme is sometimes defined as a more abstract unit, i.e. the unit or rather category that underlies a concrete realisation, also called a morph. For example, Bauer (2003: 17) distinguishes the morphs 'a' and 'an' for the indefinite article morpheme in English. This can be notated as either {a/an} or {indefinite article}, and especially the latter notation shows that a morpheme is not a purely morphosyntactic element, but something underlying the concrete morphosyntactic form. In other words, there is a difference for Bauer and other morphologists between the underlying morpheme and the concrete morph, which FDG does not makes, using the term Morpheme for all morphosyntactic units that are Morphosyntactic Words or are smaller than Morphosyntactic Words.

Now let us return to the joint expression of multiple pragmatic and semantic units in single grammatical units, i.e. portmanteau morphemes. For instance in Spanish, operators of illocution, tense, aspect, person and number are expressed together in single predicate affixes, e.g. *lleg-ó* 'arrive-IND.PST.PFV.3SG'

(Hengeveld 2004: 4). This is non-transparent, since multiple pragmatic and semantic operators correspond to a single affix. In the tradition described above, in which a distinction is made between underlying Morphemes and their realisations, the correct term is 'portmanteau morphs'; for example Bauer (2003: 19) defines these as morphs that realise more than one morpheme.

In this dissertation, I will henceforth use the term cumulation for the joint expression of multiple meanings in single grammatical units, i.e. Affixes and Grammatical Words, as do Hengeveld (2007), Bickel & Nichols (2013), and many other typologists. This is opposed to the joint expression of a lexical unit and grammatical marking in one morphosyntactic unit, which is called stem alternation. Different types of stem alternation will be dealt with separately below in Sections 4.3.2, 4.4.6 and 4.4.8. Note that, for reasons of space, I will often abbreviate the phrase 'Affixes and Grammatical Words' to simply 'affixes'.

Some semantic units are more prone to be expressed in portmanteau morphemes than others. Firstly, the joint expression of person and number is very common in the languages of the world – pronouns in particular usually include both number and person operators in one Morpheme. Pronominal portmanteaus in languages that have a gender system sometimes include this gender feature in pronominal marking as well, e.g. in Egyptian Arabic: *Muna maat-it* 'Muna die-3SG.F' (Gary & Gamal-Eldin 1982: 60). Secondly, marking of case, if that exists in a language, is frequently fused with marking of number and, if available, gender, e.g. in the Latin suffix *-orum* that marks genitive plural on nouns of the second declension, containing a majority of masculine nouns. A third set of semantic categories that is commonly expressed in portmanteau morphemes rather than in separate morphemes are those of tense, aspect, mood and evidentiality. Such operators, traditionally known as TAM markers and nowadays often as TAME markers, are in many languages expressed jointly with operators of person and number, for example in Dutch verbal inflection: *-t* 'PRS.3SG'.

For each language, I will establish whether it exhibits cumulation or not. Since person and number are cumulated in presumably nearly all languages, taking into account this category of portmanteaus fails to discriminate between a

predominantly fusional language like Spanish and an isolating language like Teiwa that has pronouns combining person and number, but hardly any other portmanteaus. Therefore, I will restrict the cumulation feature to cumulation of case marking and cumulation of TAME marking (cf. Bickel & Nichols 2013 for a similar approach). If a language displays portmanteaus expressing case and another semantic category, or portmanteaus expressing TAME plus another semantic category, or both, the language is considered non-transparent with respect to this feature. Thus, the presence of fusional morphology is measured both in the verbal and in the nominal domain. If a language displays no portmanteaus in these domains at all, it is considered transparent in this respect. If case and TAME are not expressed in the language, this feature does not apply.

### *4.3.2 Morphologically conditioned stem alternation: suppletion – RL-ML*

To mark grammatical information on a stem, a language may employ affixes. However, function marking can also be performed by changing the form of a stem, resulting in a morphosyntactic unit that expresses multiple higher level units, e.g. a Lexeme and a semantic function or a Lexeme and an operator. Such a unit is non-transparent, as it involves a many-to-one meaning-to-form relation. Several types of stem alternation exist, of which some result in opacity while others do not. One distinction that can be made is between partial modification of the base and alternation of the entire stem (cf. Hengeveld 2007). The first category will be called irregular stem formation and is treated in Section 4.3.3, the latter category is called suppletion and is the topic of this section.

Suppletion is the morphological process in which marking of grammatical information requires a stem form that is non-derivable from other stem forms of the same Lexeme (Bauer 2003: 48, Hengeveld 2007: 39). An example from French is given in (12).

Bauer (2003: 49)

(12) je vais j'irai j'allais
'I am going' 'I will go' 'I went'

These verbal stems express not only the lexical meaning of the root, but also tense and aspect information, which is why they are opaque.

For each language in the sample, I will determine whether it displays cases of suppletion. If a language does so, it is considered opaque with respect to this feature, but if there are no cases of suppletion, the languages qualifies as transparent.

### *4.3.3 Morphologically conditioned stem alternation: irregular stem formation – RL-ML*

Grammatical information can be marked, as said above, by means of affixes, by means of suppletion, but also by means of a modification to part of the stem, e.g. to its nucleus or to its segmental structure (cf. Bauer 2003: 32). If such alternations are irregular (that is, if they apply to particular stems but not to all), the lexical meaning of the stem and the grammatical meaning marked by the alternation are not formally separate – the stem expresses both in one unsegmentable form, which is why cases of irregular stem formation are non-transparent.

There is one type of irregular stem alternation that will not be studied here within the category of fusion, but rather below in the category of form-based form. This concerns morphophonologically conditioned stem alternation, which applies when a stem-final phoneme is assimilated when adjacent to a particular Morpheme (Hengeveld 2007: 39). This occurs, for example, in Hungarian: a stem-final /t/ is palatalised under the influence of the imperative suffix *-s*, as illustrated in (13). Since the alternation occurs only with the imperative suffix and not with other s-initial suffixes, this is not merely a phonological process, but a morphophonological one.

Hengeveld (2007: 39)

| (13) | köt | köš-s |
|---|---|---|
| | tie- | tie-IMP.INDEF.2SG |
| | ‘tie’ | ‘Tie!’ |

A morphophonologically based stem alternation does not result in a 2:1 meaning-to-form relation, as the altered stem form does not involve an added meaning – *köš* still only has its lexical meaning ‘tie’. Rather, the alternation is a form-based form process, as a formal, morphophonological process alters a formal unit without a pragmatic or semantic motivation. Therefore, morphophonologically motivated stem alternation will not be studied under the header of fusion, but as a form-based form feature, to be discussed further in Section 4.4.6.

Irregular stem formation that does lead to a many-to-one relation between meaning and form comes in different types, which I will now discuss one by one, basing their categorisation on Bauer (2003: 32ff.). All types involve a change of a part of a Stem as a result of the expression of an additional pragmatic or semantic unit, thus changing the stem in such a way that it cannot be separated into two distinct units. Bauer distinguishes at least four such morphological processes, viz. vowel mutation, consonant mutation, segmental modification and suprasegmental modification, of which I will now give examples.

Vowel mutation, comprising both umlaut and ablaut, is for instance found in English, in which plurality is usually expressed by means of a suffix *-s*, but in particular cases by an alternation of the stem’s vowel, e.g. *mouse* (SG) versus *mice* (PL). *Mice* is non-transparent as it combines the semantics of the Lexeme MOUSE with a plurality operator in one form. The second irregular stem formation process, consonant mutation, appears extensively in Irish Gaelic. In example (14), the initial consonant of the verb is altered to express past tense.

O'Neill (2012: 61)

(14) a. fág b. d'fhág

fɑ:g ḏɑ:g

leave leave.PST.3SG

'leave' 'left'

An example of modification of the segmental structure of a stem resulting in a many-to-one relation (Bauer 2003: 32) is again found in English, e.g. *thief* and *thieve*: voicing of the stem-final fricative reflects whether the Word expresses a State-of-Affairs or an Individual. Hence, the stem *thieve* can be said to jointly express a Lexeme and semantic information on its entity type. Finally, suprasegmental modification can also result in a Stem that expresses multiple meanings, for instance in the case of English *INsult* (noun) and *inSULT* (verb). The stress pattern marks the difference between a status as State-of-Affairs or as an Individual and thus expresses additional meaning.

For each language in the sample, I will determine whether it displays cases of irregular stem formation. If a language does so, it is considered opaque with respect to this feature. If irregular stem formation occurs as a result of morphophonological rules, it is not discussed here but categorised as a form-based form feature (see Section 4.4.6).

### *4.3.4 Summary fusion features*

The non-transparent features in the fusion category that will be studied in this dissertation are cumulation and morphologically based stem alternation. Within stem alternation, a distinction is made between suppletion and irregular stem formation. Morphophonologically motivated alternations are discussed as form-based form in Section 4.4.6.

## 4.4 Form-based form

This section lists all features that constitute a null-to-one relation between meaning and form, that is, all forms that have no higher level counterpart. Such phenomena are also known under the term 'autonomous syntax', but for reasons given in Section 2.4, I will speak of form-based form.

### *4.4.1 Grammatical gender – RL-ML*

Languages may exhibit a lexically motivated classification of nouns. An example of such so-called grammatical gender is the nominal classification of Dutch. Even though Audring (2009a), among others, shows that there are semantic motivations especially behind the use of gendered anaphoric and relative pronouns in Dutch, at least the selection of either the common article *de* or the neuter equivalent *het* is in most cases completely lexically motivated. There is for instance no semantic or morphophonological rationale behind the article use in *de kamer* 'DEF.COMM room' versus *het huis* 'DEF.N house'. This is opaque, since morphosyntactic marking occurs on the basis of a morphosyntactic feature only, and does not have a pragmatic or semantic motivation.

Nominal classification can also be semantically motivated, for instance in Kikongo, which distinguishes between at least ten noun classes, the exact number depending on how one defines class and gender. Following Dereau's (1995: 17ff.) classification, class I contains humans, class II contains humans with authority (e.g. the word denoting 'chief'), family members and natural things like 'sun' and certain higher order animal species. Class VII contains abstract nouns that have no plural, class IX contains diminutives, etc. Even though exceptions exist of nouns whose gender cannot be related to their semantics, the general gender assignment criteria are semantic in nature. Such semantic gender is transparent, since there is a one-to-one relation between the semantic class property of the noun and its marking. Of course, it is non-transparent to mark semantic class redundantly, as discussed in Sections 4.1.3 and 4.1.4, but semantic classification as such does not violate transparency.

The difference between grammatical and semantic gender may be hard to make, since classification systems are often neither the one nor the other – usually, semantic classifications have lexically motivated exceptions and even in grammatical gender systems like the Dutch one, some semantic motivations can be found. If the language has gender agreement, the distinction can nevertheless be made by observing agreement patterns of nouns of which syntactic and semantic gender do not coincide. For example, the syntacticity of the Dutch grammatical gender system is apparent through syntactic agreement, e.g. in constructions like *het meisje* 'DEF.N girl(N)', where agreement obviously occurs on the basis of lexical rather than semantic properties of the noun. If such examples are not available, I will simply follow an expert's judgement on whether a noun classification system is semantic or grammatical.

To conclude, a language is considered opaque with respect to this feature if it exhibits a grammatical gender system. It is transparent if it exhibits a semantically motivated nominal classification system, or no nominal classification at all.

### *4.4.2 Nominal expletives – RL-ML*

Another unit that does not have a pragmatic or semantic counterpart is the nominal expletive, also known as dummy subject, empty argument or pseudo-argument. For example, compare the Fongbe sentence and its English translation in (15).

Fongbe - Lefebvre & Brousseau (2002: 245)

(15) jì jà
rain fall
'It is raining.'

The argument of the predicate 'falling' is expressed explicitly in Fongbe: the rain is expressed as a participant in the event of falling. In English, however, the predicate *to rain* triggers the use of a dummy subject *it*, which has no referent. In other words, *it* does not have a pragmatic (Referential Subact) nor a semantic (Individual) counterpart and is a form-based form. By some linguists (e.g. Bennis 1986), *it* is said

to refer to a referent that is implied by the predicate, however, that option is not recognised in FDG. Apart from pronominal expletives like *it*, languages often have locational adverbs as expletives, such as *there* in *There is no-one here* in English.

Another type of pronominal expletive is given in example (16), again from Fongbe. In this case, the pronominal element in the main clause does again not refer to a pragmatic or semantic entity, but it does refer to the embedded clause – a morphosyntactic unit. Since the pronominal element does not have a higher level counterpart, it qualifies as a form-based form.

Lefebvre & Brousseau (2002: 67)

| | | | | | | |
|---|---|---|---|---|---|---|
| (16) | é | nyɔ́ | ɖɔ̀ | kɔ́kú | ní | yì |
| | it | be_good | COMP | K. | SBJV | leave |

'It is good that Koku leaves.

However, while the dummy subject of a weather predicate has no explicit referent whatsoever, the dummy in sentence like (16) is still referring to something, viz. to the complement clause that follows. Therefore, the first type of dummy is more heavily opaque than the second type.

Travis (1984) shows that there is an implicational hierarchy for different types of expletives, stating that if a language has nominal expletives at all, they will appear with weather predicates. As apparent from the examples above, Fongbe disproves this hierarchy, but still, I will assume that the hierarchy is at least a cross-linguistic tendency. Therefore, I will assess for each language whether it uses dummy subjects with weather predicates. If a language uses empty pronominal subjects with such predicates, the language is opaque, but if it uses a semantic argument like *rain*, it is transparent with respect to this feature.

Other elements that are sometimes seen as semantically vacuous are copulas. These are presumed to be morphosyntactic elements that are inserted only when a non-verbal predicate cannot bear certain verb markers (e.g. agreement affixes) and requires a verbal element to which that information can be attached. In such an analysis, copulas function as dummy verbs or verbal expletives. However,

in many and possibly all languages, copulas do have a semantic effect, as described by e.g. Peterson (2011: 375) on Kharia: "the presence or absence of [a copula] can express a semantic difference between what we may loosely term "narrative description […] and "general or self-evident truth"". Thus, it is questionable whether we should see copulas in all cases as semantically empty items – this remains a topic for future research. For now, I will not treat copulas as form-based forms.

### *4.4.3 Syntactic functions – IL-ML*

FDG reserves the term 'alignment' for so-called non-hierarchical morphosyntactic ordering in different domains, viz. the clause, the phrase or the word. In Clauses, there is non-hierarchical alignment between predicates and one or more arguments. Different types of argument functions are relevant in different languages. First of all, there are languages in which the morphosyntactic expression of arguments is dependent on pragmatic functions – FDG distinguishes the pragmatic functions Topic and Comment, Focus, Background, Contrast and Overlap. A language in which one or more of these functions are consistently marked morphosyntactically is said to exhibit interpersonal alignment.

An example of a language with interpersonal alignment is Tagalog, in which Topic arguments are obligatorily marked by means of the proclitic *ang*=, as shown in example (17). Apart from being marked by *ang*=, the semantic role (Actor, Undergoer, Location) of the Topic argument is cross-referenced on the predicate. Note that in all three sentences, the semantic roles of the arguments are the same, whereas the Topic function differs per sentence.

Bickel (2011: 8/9), italics mine

(17) a. bumilí *ang=lalake* ng=isda sa=tindahan
PFV.A.buy SPEC.TOP=man OBL=fish LOC=store
'*The man* bought fish at the/a store.'

b. binilí ng = lalake *ang = isda* sa = tindahan
PFV.U.buy OBL=man SPEC.TOP=fish LOC=store
'The/a man bought fish at the store.'

c. binilhan ng = lalake ng = isda *ang = tindahan*
PFV.L.buy OBL=man OBL=fish SPEC.TOP=store
'The/a man bought fish at the store.'

Languages that predominantly base their argument expression on semantic information are said to display representational alignment. In such languages, the semantic role of arguments determines their morphosyntactic form, e.g. their case-marker, and/or the form of the predicate.

Following Foley & Van Valin (1984), FDG (cf. Hengeveld & Mackenzie 2008: 194ff.) distinguishes between three semantic functions, viz. Actor, Undergoer and Locative. Arguments with an Actor function prototypically volitionally perform an action, and are thus similar to the argument traditionally called Agent. Undergoers passively undergo an event or state, and can as such be mapped onto the roles traditionally known as Patient, Experiencer and Theme. The semantic function Locative is a category representing oblique functions, such as Recipient and Beneficiary.

As an example of a language with an alignment system along the lines of these functions, consider the Tagalog examples above: even though Tagalog arguments themselves are not marked for their semantic function, semantic roles are relevant in the marking of the predicate. Note that this also shows that a language can exhibit interpersonal and representational alignment at the same time.

A language in which semantic role is not only marked on the predicate, but also consistently marked on arguments themselves is Acehnese. Arguments are expressed by means of pronominal clitics that, as shown in in (18), are dependent on semantic roles, viz. =*geuh* for an Undergoer, =*geu* for an Actor argument.

Durie (1985: 55-58)

(18) a. gopnyan galak = geuh that
3.HON happy=3.HON.U very
'He is very happy.'

b. keu = jih ka = geu = jôk buku = nyan lê = gopnyan
to=3.FAM INCH=3.POL.A=give book=that by=3.POL
'He gave him that book.'

Another type of representational alignment is called hierarchical alignment (Hengeveld & Mackenzie 2008: 321), in which the marking of arguments on predicates follows a hierarchy of animacy and person.

Both pragmatically and semantically based alignment are transparent, since the various morphosyntactic markings on the predicate and the argument(s) are all straightforwardly connected to pragmatic or semantic functions. A third alignment type, however, is opaque: morphosyntactic alignment is the expression of arguments and predicates that disregards pragmatic and semantic principles but is purely internal to the Morphosyntactic Level. Only in this case does FDG speak of grammatical relations, or syntactic functions, because the relations are truly grammatical and syntactic, and not in fact pragmatic or semantic. The terms Subject and Object are reserved for grammatical relations only, while words like Topic, Focus, Comment and Background are used for pragmatic functions, and Actor, Undergoer and Locative are terms used for semantic functions. In other words, a Subject is in this dissertation always a morphosyntactic entity; it can never be used to refer to an Actor argument, which is a semantic entity, or to a Topic, which is a pragmatic entity. This parallels a distinction made in the domain of case marking between semantic or concrete case, which marks semantic roles or spatial relationships, and grammatical or abstract case, which marks syntactic functions or other syntactic information (Haspelmath 2009).

A difficulty in this area of linguistic analysis is that subjects and topics are highly similar to each other, in fact so much that they cannot always be

distinguished in a principled way. Li & Thompson (1976: 459) argue that what is often analysed as a subject-predicate structure in languages is in fact a topic-comment structure, and that subjects and topics share a number of properties. In FDG, the difference can be understood by considering syntactic subjects to be internal to the Morphosyntactic Level, though triggered by information from the Contextual Component, which is external to the grammar (cf. Section 2.1). Given the high coincidence between subjects and topics, one could also argue that the availability of subjects in a language provides a transparent relation between the pragmatic and morphosyntactic units, constituting a one-to-one relation at that interface.

Li & Thompson (1976) state that languages usually have both a subject and a topic, but that one of those can be more prominent, distinguishing typologically between topic-prominent languages, subject-prominent languages, and languages in which topic and subject are prominent or non-prominent to an equal extent. They classify languages with respect to this dichotomy by using criteria such as definiteness (topics tend to be definite, while subjects need not be) and sentence-initial position (topics are often sentence-initial, while subjects can be preceded by, for instance, a verb). This is a fine-grained testing procedure that takes into account many aspects on which information is scarce in most reference grammars. Therefore, I am not able to take over Li & Thompson's metric of distinguishing between Subjects and Topics, but will use my own diagnostics, which are admittedly less fine-grained.

To determine whether a language has syntactic functions, I will use two constructions as diagnostics. There are more diagnostics for morphosyntactic alignment, but for reasons of time and space I have chosen to restrict myself to the two discussed here. The first and most common test case is that of neutralisation of pragmatic and semantic roles in intransitive predicates, which by definition have one argument commonly called the S argument. Depending on the predicate, this argument can function as an Actor, Undergoer or Locative. In some languages, the semantic function of an S argument is always expressed explicitly, as in the Acehnese example (19).

Durie (1985: 55, 56)

(19) a. lôn teungoh = lôn = jak
1 middle=1.A=go
'I am going/walking.'

b. gopnyan galak = geuh that
3.HON happy=3.HON.U very
'He is very happy.'

In other languages, semantic function is neutralised in intransitive clauses, as in the English example (20). The pronominal Actor argument in (20a) and the pronominal Undergoer argument in (20b) receive an identical morphosyntactic expression, disregarding their semantic role. Note that the semantic role can actually be expressed in English, since first and third person pronominal arguments are case-marked in transitive clauses, e.g. (20c).

(20) a. He walked.

b. He fell.

c. He was kissing him.

Since pragmatic and semantic information is ignored in (20a) and (20b), it becomes relevant to speak of a grammatical relation Subject in English. For languages like Acehnese, postulating a grammatical relation is irrelevant, since pragmatic and semantic functions can fully predict the morphosyntactic behaviour of arguments.

Note that morphosyntactic alignment can be both of the ergative and of the accusative type – these are equally non-transparent. However, if a language displays neutral alignment, not marking arguments or predicates for any function, absence of marking in intransitive clauses does not signal neutralisation but simply absence of roles, be they pragmatic or semantic. In that case, functions of any type cannot be distinguished on the basis of explicit morphosyntactic marking, but only on the basis

of word order and control. This is for instance the case in Fongbe, a language in which predicates nor arguments are explicitly marked for the roles of arguments. Therefore, an unmarked S argument is not a sign of neutralisation of semantic function marking, but simply the only option. The only clue regarding the functions of arguments in Fongbe comes from word order, since in neutral sentences, the Actor precedes the Undergoer. However, word order is relatively flexible in Fongbe, since pragmatic considerations can trigger deviations from standard word order, often additionally marker by discourse markers. Thus, word order cannot serve as a straightforward expression of semantic functions; rather, the listener has to rely on contextual information and on pragmatic cues (i.e. word order and discourse markers) to determine the roles of participants. Since in other languages too, word order is strongly affected by factors other than pragmatic or semantic functions of arguments, I will not take it into account as an expression of such functions. Furthermore, control phenomena are too complex to study within the scope of this study. Therefore, to establish whether pragmatic, semantic or syntactic functions are expressed morphosyntactically, I will only consider overt argument and predicate marking.

A second construction in which pragmatic and semantic roles are neutralised, giving relevance to the postulation of a syntactic function subject in a language, is the antipassive construction in ergative languages and the passive equivalent in accusative languages. Consider the English examples in (21).

Hengeveld & Mackenzie (2008: 326)

(21) a. The man gave the book to the boy.

b. The book was given to the boy by the man.

c. The boy was given the book by the man.

In these sentences, *the man* functions as Actor, *the book* is Undergoer and *the boy* has the semantic function Locative. These semantically different arguments are expressed identically in (21a), (21b) and (21c) respectively, i.e. as the grammatical

subject, that is unmarked for case and triggers agreement on the predicate. Since the semantic role of arguments is neutralised in this construction, there is a second reason to distinguish a syntactic function Subject in English.

Note that violation of a one-to-one relation between semantic function and its expression by a passive is only complete when there is the possibility of adding the Actor through a so-called 'by-phrase'– without the by-phrase, the valency of the predicate is reduced, so that there is no true neutralisation of semantic functions, but rather a loss of one of the functions and an entirely different predicate-argument structure. The presence of the Actor marked as an oblique argument shows that the predicate and its valency are the same, while the morphosyntactic expression of semantic functions is different, and this is what creates non-transparency.

I will count a language as non-transparent with respect to this feature if it shows neutralisation of pragmatic and semantic roles in intransitive clauses, a passive construction, or both. Even when a language predominantly shows pragmatic or semantic alignment, but has a passive construction, it is counted as opaque. If a language has neutral alignment and does not express pragmatic or semantic roles at all, neutralisation into a syntactic function is impossible and this feature does not apply.

#### *4.4.4 Influence of complexity on word order – IL-ML, RL-ML*

The FDG account of morphosyntactic placement or constituent order relies heavily on the concept of Templates, introduced briefly in Section 2.1. Templates are configurational units at the Morphosyntactic Level that are responsible for the organisation of certain units relative to each other. Thus, they organise adjacency and phonological binding, as well as the linear ordering of units (constituents, words and bound morphemes). A filled morphosyntactic Template is the result of the Encoding of pragmatic and semantic Frames, which are configurational units at the Interpersonal and Representational Levels that register the functional coherence of different units. For instance, a prototypical transitive clause may be triggered by a representational Frame consisting of a predicate, an Actor argument and an Undergoer argument. The Encoder translates this Frame into a Template in which

these elements are linearly ordered, for example in an SVO-language as $Np_A$-V-$Np_U$. In fact, Templates do not function in terms of the traditional ordering of Subject, Object and Verb, but pattern along four absolute positions ($P^I$, $P^2$, $P^M$ and $P^F$) and an in principle unlimited number of relative positions. A further description of the exact process is not of purpose here; the interested reader is referred to Hengeveld (2013) for a further account.

In sum, constituent placement is argued in FDG to be based principally on the pragmatic and semantic status of the constituents. This Formulation-based placement process can, however, be overruled by the morphosyntactic complexity of constituents. It is often observed that heavy constituents, that is, constituents that are morphosyntactically complex, are preferred to appear more to the end of the sentence while light elements appear at the beginning. This is probably due to matters of planning and processing, that allow 'easy' elements to be articulated immediately, while more complex units take more time to plan and articulate and are therefore expected to be articulated later (cf. Hawkins 1990).

A well-known phenomenon based on such influence of complexity is heavy NP-shift: the appearance of a complex constituent near the end of a sentence, that is expected to appear somewhere else according to 'default' placement rules of a language. Apart from NPs, units that can and sometimes have to undergo complexity-based placement are complement clauses, relative clauses (possibly leading to discontinuity, as shown in Section 4.2.1), adpositional phrases, and possessive phrases. An example of a dislocated complement clause comes from Sri Lanka Malay in (22), in which the complement clause, between square brackets, would normally appear pre-verbally because of its Undergoer function, but is realised post-verbally due to its weight.

Sri Lanka Malay - Nordhoff (2009: 739)

(22) se=ppe oorang thuuva pada anà-biilang
1SG=POSS man old PL PST-say
[kithang pada Malaysia=dering anà-dhaathang katha]
1PL PL Malaysia=ABL PST-come QUOT
'My elders said that we had come from Malaysia.'

The term 'heavy NP-shift' implies movement from a heavy unit, initially placed in its standard position and shifted to another location. However, the FDG account of placement does not allow for any kind of movement, so that this term is theoretically inappropriate. I will therefore speak of 'influence of complexity on word order' rather than of 'shift' or 'movement'.

The fact that the morphosyntactic weight of a constituent determines its place in a sentence is non-transparent, since a morphosyntactic property of the constituent determines morphosyntactic placement, which means that there is a null-to-one relation between the pragmatic and semantic information of the constituent, and its place in a morphosyntactic Template. Hence, complexity-based placement is a case of form-based form, so that languages that allow complexity to influence constituent ordering will be considered opaque with respect to this feature, while languages in which complexity does not play a role in placement will qualify as transparent.

Obviously, the influence of complexity as an ordering principle may coincide with the influence of other principles, for instance when a complex constituent is realised sentence-finally because of semantic motivations. In that case, it is impossible to determine whether complexity plays a role in placement. A second difficulty in assessing an influence of weight on ordering is that it is hard to actually define or quantify the weight of a constituent. For both reasons, I will not rely on isolated examples only, but count a language as exhibiting an influence of complexity on morphosyntactic ordering on the basis of a statement in the reference grammar or by an expert on the language.

*4.4.5 Function marking is predominantly head-marking – RL-ML*

Functions and operators, that is, items representing grammatical information, can be overtly marked on the unit to which the information applies, for example by means of affixes. Affixes by definition attach to a host, e.g. a Root or a Stem, and prototypically for inflectional affixes as opposed to derivational ones, this host is always of a specific morphological class. For example, the plural marker *-s* in English attaches to nouns only, whereas the third person singular marker *-t* in Dutch always attaches to verbs. Furthermore, the host of an inflectional affix is typically a Morphosyntactic Root, Stem or a Stem plus other affixes, but never a Phrase or a Clause. A further characteristic of inflectional Affixes is that they cannot freely occur in different combinations, but have to occur in a particular order (cf. Zwicky & Pullum 1983).

A different type of grammatical function marker is the clitic (cf. Anderson 2005). Clitics resemble inflectional affixes in the sense that they are bound morphemes that require attachment to hosts in order to form a phonological unit together. However, clitics differ from inflectional affixes because they are less selective as to their hosts: the morphosyntactic class of a Lexeme is in principle irrelevant for its ability to host clitics. Also, clitics can often attach to units of all degrees of complexity, meaning that their host can be a Word, a Phrase or a Clause. The ability to attach to different units renders clitics the status of Morphosyntactic Word (Hengeveld & Mackenzie 2008: 446).

A third type of function marker is the free-standing one, that is, a function marker that does not require a host but forms an independent Phonological Word on itself. Such phonologically independent function markers are in many reference grammars referred to as (functional) particles and are called Grammatical Words in FDG terms. Typically, particles scope over phrases rather than words, as do clitics. In sum, FDG distinguishes words that are Morphosyntactic Words *and* Phonological Words, viz. particles or Grammatical Words, Morphosyntactic Words that are not an independent Phonological Word, viz. clitics, and units that are both morphosyntactically and phonologically dependent on a host, viz. inflectional

affixes. Particles and clitics scope over phrases, whereas affixes scope over the head of a phrase only.

Note that this description of the differences between affixes and clitics is in terms of discrete properties, but of course these differences are in fact gradual. There is a grey area in between the prototypical affix that attaches to simple stems of one morphological class only and the prototypical clitic that attaches to units of any type. In order to do justice to the full range of variation, one would have to argue for functional morpheme in each language whether it is located more with one or other end of the clitic-affix continuum (cf. Zwicky & Pullum 1983, Leufkens 2009). Since such as a level of detail is not possible within the scope of this study, I will resort to simply following judgements of experts on the languages under consideration.

Affixal function markers are considered opaque because the morphosyntactic information of morphological class and/or complexity of the host determines the nature of the marking. For example, if grammatical information like verbal status determines the selection of a particular affix, then there is influence of a morphosyntactic property of the word on a morphosyntactic process, entailing a zero-to-one relation between the Representational Level and the Morphosyntactic Level. Therefore, a language that predominantly uses phrase-marking function markers, i.e. clitics and particles, which are indiscriminate to the type and complexity of the element over which they scope, will be considered transparent in this respect.

Of course, the notion of 'predominance' is problematic here, since it is practically impossible to quantify the amount of clitics, particles and affixes in a theory independent way and determine the ratio. In order to assess the transparency of language with respect to this feature, I will determine for each language the proportion of the amount of head-marking affixes versus the amount of phrase-marking clitics and particles. If a language shows a clear predominance of phrase-marking items, it will be considered transparent with respect to this feature. If a language displays a relatively large amount of inflectional affixes, it will be considered non-transparent with respect to this feature. If both head-markers and phrase-markers occur non-marginally and it cannot be determined which one is

predominant, the feature will remain undecided and will not be counted as either transparent or opaque.

### *4.4.6 Morphophonologically conditioned stem alternation – RL-PL*

As discussed above, Stems may undergo changes in their form under the influence of an adjacent morpheme, for example an attached affix. This is categorised as form-based form since it involves a formal process without any pragmatic or semantic motivation. Phonologically predictable alternations will be discussed in Section 4.4.8, but the current study discusses alternations that only apply to particular morphemes. For example, in the Hungarian example (13), repeated here as (23), a stem-final /t/ palatalises under the influence of the imperative suffix /-s/. This is not phonologically predictable, since only the imperative suffix triggers this particular alternation.

Hengeveld (2007: 39)

| (23) | köt | köš-s |
|---|---|---|
| | tie- | tie-IMP.INDEF.2.SG |
| | ‘tie’ | ‘Tie!’ |

I will assess the morphophonological properties of each language and determine whether they exhibit any opaque morphophonological processes applying to stems. If a language does not have such processes at all, it will be considered transparent with respect to this feature.

### *4.4.7 Morphologically and/or morphophonologically conditioned affix alternation – RL-PL*

Like stems, affixes and Grammatical Words can undergo alternations that are influenced by the particular stem they attach to. To the extent that such alternations are fully predicted by phonological rules, they are discussed in Section 4.4.10, but the current section deals with affix alternations that have a morphological or morphophonological motivation, as they are conditioned lexically by conjugation or

declension classes or apply to particular affixes only. Such alternations are non-transparent, since they do not follow from a pragmatic or semantic motivation, but are purely morpho(phono)logically determined. Note that for stems, morphologically motivated alternations and morphophonologically conditioned alternations are treated in different sections, because the first type of alternations leads to fusion while the latter is of the form-based form type. Since both morphologically conditioned and morphophonologically conditioned affix alternations belong to the form-based form category, they are treated here as one feature. Note furthermore that I will henceforth use 'affixes' where I mean 'affixes and Grammatical Words', for reasons of readability and conciseness.

Morpho(phono)logically conditioned affix alternations may occur for a number of reasons. Firstly, affixes may change their form depending on the morphological class of the particular verbal stem they attach to. In such cases, we speak of conjugation: the phenomenon that distinct verbal classes can be distinguished according to the different inflectional affixes that they take. An example of a conjugational system is found in Spanish, that has three inflection sets to mark verbs for imperfective mood (Bybee 1985: 37). Thus, Spanish verbs can be divided over three classes on the basis of which set of inflection they host. Verb conjugations are not determined semantically, nor phonologically – one cannot predict from the meaning or the phonological shape of the verb which set of inflectional affixes it gets. These conjugations should therefore be seen as a morphological property of verbs, and the selection of a particular affix allomorph is morphologically motivated.

The nominal equivalent of conjugation is declension, i.e. classification of nouns on the basis of the inflectional affixes they take, or taking another perspective, affix alternation on the basis of nominal classes. The classic example is Latin, a language in which nouns belong to one of five declension classes, assignment of which cannot be predicted on the basis of the semantics or phonology of the noun. As apparent from Table 4.1, the case and number inflection of nouns is dependent on the declension to which they belong, which strongly correlates to the gender of the noun.

**Table 4.1: Latin declension**

| | 1st declension | | 2nd declension | |
|---|---|---|---|---|
| Case | Singular | Plural | Singular | Plural |
| Nominative | stēll-**a** | stēll-**ae** | mūr-**us** | mūr-**ī** |
| Accusative | stēll-**am** | stēll-**ās** | mūr-**um** | mūr-**ōs** |
| Genitive | stēll-**ae** | stēll-**ārum** | mūr-**ī** | mūr-**ōrum** |
| Dative | stēll-**ae** | stēll-**īs** | mūr-**ō** | mūr-**īs** |
| Ablative | stēll-**ā** | stēll-**īs** | mūr-**ō** | mūr-**īs** |

In many conjugation or declension systems, some part of the stem, usually a vowel, indicates to which declension it belongs. Such elements are called thematic elements. In these systems, the selection of specific inflectional affixes does partly follow from the phonological shape of the verb or noun, for example in Dutch, in which nouns are marked for plurality either by *-en* or by *-s*. The choice for one of the allomorphs is determined morphophonologically, because roughly, *-en* is selected after stressed syllables, while *-s* occurs after unstressed syllables ending in a sonorant or coronal, and after loanwords (Kusters 2003: 28). Since this alternation only applies to this particular suffix, inflection of Dutch nouns for plural is morphophonologically motivated rather than purely phonologically.

Affix alternations may apply to particular affixes only, in which case there is reason to distinguish between different classes of affixes. This is for instance the case in West-Greenlandic, in which a distinction is made between so-called 'truncating affixes' and 'replacive affixes'. Truncating affixes, for example *-lir* 'begin', retain their initial consonant but will delete the final consonant of the stem to which they attach, so that for example the stem *siniC* 'sleep', ending in an underspecified consonant, will become *sini-lir-puq* 'sleep-begin-3SG'. When a replacive affix, e.g. *-lirtuuq* 'one who likes -ing' is attached to a stem, its initial consonant adapts to the stem-final consonant, e.g. *sin-nirtuuq* 'one who likes sleeping', from which *ni-* is deleted for other phonological reasons (M. Fortescue, personal communication, June 24, 2014). Thus, the form of the affix is not

phonologically predictable, but determined by the particular affix class to which it belongs. Since the form cannot be accounted for by a higher level motivation, this is opaque.

For each language in the sample, I will determine whether it exhibits morphologically and/or morphophonologically motivated affix alternations, due to verbal or nominal classification or due to affix classification. A language that exhibits either of these phenomena will be considered non-transparent, while a language without morphologically or morphophonologically conditioned affix alternation will be counted as transparent.

### *4.4.8 Phonologically conditioned stem alternation – RL-PL*

The phonological shape of stems may be altered under the influence of adjacent or near-adjacent phonemes. One or more phonemes of the stem may in that case adapt their place or manner of articulation, they may be voiced or devoiced, or even deleted. These phonologically and phonetically based alternations are non-transparent, as the output form is partly determined on the basis of phonological or phonetic information, without a pragmatic or semantic motivation. Phonologically or phonetically based stem alternation will be distinguished from affix alternation, because it might be the case that a language allows the one, but resists the other.

Before going into the different types of alternations, it is necessary to discuss the difference between phonologically and phonetically based alternations, which has to do with the output of (morpho)phonological rules. If the output of the rule contains segments that are all established members of the language's phoneme inventory, we are dealing with a phonological alternation. However, if an alternation rule results in a phonetic output containing segments that are not phonemes of the language, the rule involves a phonetic alternation. For example, the Dutch noun *hond* 'dog' is pronounced [hɔnt] in isolation, but when the diminutive suffix {-je} is attached, it becomes [hɔˈɲɕə]. Since the segment [ɲ] is not a phoneme in Dutch, this qualifies as a phonetic alternation rather than a phonological one. Furthermore, the combination of underlying /t+j/ becomes [ɕ], which is not a Dutch phoneme either.

In this study, it will not always be possible to determine whether an alternation is phonological or phonetic, as reference grammars do not always provide precise information on the exact phonetic shape of altered forms. Moreover, phonetic alternations are not dealt with by FDG, which sees phonetics as part of the Output Component, which is not part of the grammar proper (but cf. Seinhorst 2014 for a different view). Since drawing a strict and principled distinction between phonological and phonetic alternations is impossible within the time and scope of this study, I will simply include both types of alternations. This is not a theoretical problem, since both are equally non-transparent as they involve formal alternations without a pragmatic or semantic motivation. For reasons of readability, I will henceforth use 'phonologically conditioned alternations' when I intend 'phonologically or phonetically conditioned alternations'.

Several alternation processes exist, of which I will now discuss phoneme insertion, phoneme deletion, and alternation of voicing, place and manner of articulation. Firstly, phoneme insertion can occur to prevent two phonemes from being adjacent. For example, in Turkish, an epenthetic vowel is inserted in order to prevent the clustering of three consonants: /burn/ 'nose' + /-da/ 'LOC' > [bu.run.da] (Kornfilt 1997: 497).

A second alternation that stems may undergo as a result of a forbidden phoneme adjacency is the elision of one or more phonemes. This occurs for instance in Dutch, where it is impossible to have two adjacent identical consonants (geminates). If a geminate arises, for instance as a result of compounding (e.g. /krɑp/ 'scratch' + /pa:l/ 'pole' > 'scratching post'), one of the consonants will be deleted ([krɑpa:l]), a process which is called degemination.

Thirdly, phonemes may undergo assimilation of place of articulation, manner of articulation, and assimilation of voicing. A straightforward example of place assimilation is found in the Dutch compound [tampasta] 'tooth paste', a combination of /tɑnt/ 'tooth' and /pɑsta/ 'paste', in which the /t/ is deleted and the nasal assimilates its place to the following consonant.

For each language in the sample, I will assess if it undergoes any of the phonologically conditioned alternations described here and if so, of which types those are. A language that exhibits at least one of the various types of phonological alternations is considered non-transparent with respect to this feature.

#### *4.4.9 Phonologically conditioned affix alternation – RL-PL*

Equivalent to phonologically based alternations in stems, alternations may occur in grammatical units (i.e. affixes and Grammatical Words, again abbreviated to affixes for reasons of readability and conciseness) as a result of a disallowed adjacency of phonemes in a certain language. For example, the Dutch past tense singular suffix on is *-te* after a stem-final voiceless consonant, but *-de* after a voiced consonant or vowel. Such phonologically conditioned alternations are non-transparent, as the output form is determined on the basis of phonological information, without any pragmatic or semantic motivation. It will be determined whether the languages in the sample exhibit phonologically based alternations in Grammatical Words or affixes, or both. If so, they are considered non-transparent with respect to this feature.

#### *4.4.10 Summary form-based form features*

Non-transparent features in the form-based form category that will be studied are grammatical gender, nominal expletives, syntactic functions, influence of weight on word order, a predominance of head-marking over phrase-marking, and phonological assimilation of stems and affixes.

### 4.5 Summary of the list of non-transparent features

Table 4.2 provides an overview of all non-transparent features discussed above. For each feature, it is indicated whether it will be studied in this dissertation, together with its possible values. If a feature is excluded, the reason for exclusion is given.

**Table 4.2: Overview of non-transparent features and their possible values**

| Feature | Transparent value | Opaque value(s) | Excluded because |
|---|---|---|---|
| **Redundancy** | | | |
| Multiple expression of pragmatic information | | | Insufficient data |
| Nominal apposition | | | Near-universal |
| Clausal agreement | Absent | Present (argument obligatorily expressed overtly) | |
| Cross-reference | Absent | Present (argument optionally expressed overtly) | |
| Phrasal agreement | Absent | Present | |
| Plural concord in noun phrases containing a numeral | Absent | Optional, obligatory | |
| Negative concord | | | Impossible to measure |
| Modal concord | | | Insufficient data |
| Temporal concord | | | Near-universal |
| Tense copying | Absent | Present | |
| Spatial concord | | | Near-universal |
| **Discontinuity** | | | |
| Extraposition | Absent | Present | |

| | | | |
|---|---|---|---|
| Raising | Absent | Present | |
| Circumfixes | Absent | Present | |
| Infixes | Absent | Present | |
| Non-parallel alignment | | | Insufficient data |
| **Fusion** | | | |
| Cumulation of TAME and case | Absent | Present | |
| Morphologically based stem alternation: suppletion | Absent | Present | |
| Morphologically based stem alternation: irregular stem formation | Absent | Present | |
| **Form-based Form** | | | |
| Grammatical gender | Absent | Present | |
| Nominal expletives | Absent | Present | |
| Grammatical relations | Absent | Present | |
| Complexity determines constituent order / heavy shift | Absent | Present | |
| Predominant head-marking | Majority of function markers phrase-marking | Majority of function markers head-marking | |
| Morphophonologically conditioned stem alternation | Absent | Present | |
| Morphologically | Absent | Present | |

| | | | |
|---|---|---|---|
| conditioned affix alternation (conjugation/declension) | | | |
| Morphophonologically conditioned affix alternation | Absent | Present | |
| Phonologically conditioned stem alternation | Absent | Present | |
| Phonologically conditioned affix alternation | Absent | Present | |

# Chapter 5
# Methodology

The first section of this chapter formulates the research questions that this dissertation hopes to answer. Section 5.2 presents the sample used in answering these questions. The research methods employed, as well as the reasons for adopting this methodology, are presented in Section 5.3. Finally, Section 5.4 gives hypotheses and expectations for the outcomes of the study, based on pilot studies.

## 5.1 Research questions

The current study starts from the observation that all natural languages have a certain degree of non-transparency in their grammars. No language in the world fully maintains a strict one-to-one relation between units of different levels, but as argued throughout the previous chapters, some languages are more transparent than others. Moreover, relatively transparent languages turn out to share particular opaque properties: for example, cumulation, phonological assimilation, and nominal apposition are found in all of them. On the other hand, particular other opaque features are only found in the most non-transparent languages attested so far, such as grammatical gender, which is only present in highly opaque languages like Dutch. There is evidence that this cross-linguistic distribution of non-transparent features is reflected in diachrony and in language acquisition (cf. Section 3.4).

Thus, previous research into the transparency of languages suggests that non-transparent features are not randomly distributed over languages, but that there is a meaningful pattern in their cross-linguistic occurrence. This observation leads to the first research question in 1).

1) How are non-transparent features distributed cross-linguistically?

To test this, the presence of the non-transparent features listed in Chapter 4 will be assessed for a sample of 25 natural languages, to be further described in Section 5.2. Section 5.3 deals with the practical side of this testing method. The results of the study may show a cross-linguistic ordering of individual features or groups of features into an implicational hierarchy. If that is the case, transparency turns out to be a meaningful typological term, that is, a notion that can relate different linguistic phenomena in languages of different backgrounds to each other.

Note that in the study, features are assessed only in terms of their presence or absence, not on their importance or magnitude. Some features, such as cumulation, may in fact show gradual differences in languages: languages may show cumulation to a different degree. Measuring the extent of occurrence of the tested phenomena is, however, impossible, since this involves the invention of fine-grained metrics that are not always available in the current state of linguistics. The time available for this project did not allow for the development of such metrics, let alone for testing all languages with respect to them. Therefore, all features will be approached in a binary fashion, i.e. with respect to their presence or absence.

The non-transparent features in the list can be categorised according to the type of violation of transparency, i.e. redundancy, domain disintegration, fusion or form-based form and, secondly, according to the interface at which they violate transparency. Both of these categorisations allow us to establish whether suborderings of features can be detected. If that is the case, there is a typological validation for different categories of non-transparency as well. This issue will be addressed under the second and third research questions.

The second research question, given in 2), relates to the categorisation of non-transparent features according to their type of violation, as categorised in Section 2.4.

2) How are redundancy, fusion, domain disintegration and form-based form features distributed cross-linguistically?

In order to answer this question, each non-transparent feature from the list in Chapter 4 is categorised as exhibiting either redundancy, fusion, domain disintegration or form-based form. The results of research question 1) will then be analysed according to this division. If the results show implicational relations between the different categories, this consolidates the distinction between the subclasses of non-transparency typologically. Furthermore, it may be the case – as suggested by earlier research – that implicational relations exist between non-transparent features *within* the different categories.

In a similar vein, it is interesting to see whether any implicational relations hold between violations of transparency at different interfaces between levels of linguistic organisation. This topic is questioned by the research question, given in 3).

3) How are non-transparent features at different interfaces distributed cross-linguistically?

In order to answer this question, each non-transparent feature from the list in Chapter 4 is categorised as pertaining to a particular interface, as indicated in that chapter in the heading of the features. The results of research question 1) will be analysed according to this division.

The results of this analysis will also be able to provide an answer to a question posed in Section 2.2, in which transparency was defined as a one-to-one relation between units at all levels of linguistic structure. As explained there, transparency is usually interpreted in a narrow sense, including one-to-one relations between meaning and form only. In FDG terms, this narrow interpretation of transparency includes four interfaces: the pragmatic-morphosyntactic interface (IL-ML), the pragmatic-phonology interface (IL-PL), the semantic-morphosyntactic interface (RL-ML), and the semantic-phonological interface (RL-PL). However, FDG enables the study of two more interfaces, viz. the pragmatic-semantic interface (IL-RL) and the morphosyntactic-phonology interface (ML-PL). Relations at these interfaces are not relations between meaning and form, but between different types of meaning, and between different types of form. FDG enables the study of

violations of those relations too. Research question 3) thus enables us to establish whether there is a typologically meaningful distinction between meaning-to-form mismatches on the one hand and the other mismatches on the other.

## 5.2 The sample

This section discusses the language sample that is investigated to answer the research questions addressed in section 5.1. Firstly, it addresses the method of language sampling that was used in this study, as well as the reasons for using this particular method, and secondly, the relevant properties of the sample languages are described.

The sample used in this research was created by means of the sampling method proposed by Rijkhoff et al. (1993) and Rijkhoff & Bakker (1998). This method aims to construct a sample in which the languages included are maximally diverse in genetic terms. This type of sample is referred to as a variety sample, since it tries to find a maximal amount of linguistic diversity, as opposed to a probability sample that aims to establish a correlation between certain linguistic traits. While a probability sample is convenient for the study of a very specific phenomenon, a variety sample is more appropriate to establish broader categories of languages. Since one of the aims of the current research is to provide a broad characterisation of languages in terms of their degree of transparency, a variety sample fits the study best.

The variety sampling technique by Rijkhoff et al. (1993) involves the selection of languages from different genetic groupings and from different areas. This is necessary, since the inclusion of languages that are genetically or geographically related would bias the description of transparency towards the degree and type of transparency attested in that specific family or linguistic area, while the study aims to give a balanced overview of transparency in the world's languages. The exact number of languages selected per language family is based on the so-called Diversity Value (henceforth DV) of that family, which is a measure that indicates what the extent is of the expected typological variety within that family.

The DV combines information of the number of languages in a family and the time-depth over which these languages have developed, and computes a numerical value to reflect this. The Afro-Asiatic language family, for example, has a DV of 55.53 (Rijkhoff et al. 1993: 185), which means that Afro-Asiatic language family shows a relatively high degree of internal typological diversity as compared to the Uralic-Yukaghir family that has a DV of 4.93. When drawing a sample, this means that proportionally more Afro-Asiatic languages should be selected than Uralic-Yukaghir languages, to reach the same level of diversity. The Rijkhoff method does not specify which languages this should be, but provides guidelines for how to distribute languages over language families.

In addition to sampling techniques, an important matter in language sampling is determining the sample size. While each language sample should be sufficiently large to guarantee representativeness, practical considerations often make it impossible to study many languages. The balance between these two factors strongly depends on the type and number of properties under investigation. In the current study, the number of properties that are studied is very high and thus requires a large sample. In fact, D. Bakker (personal communication, September 29, 2010) advised the study of at least a few hundred languages to represent the possible diversity of transparency in languages of the world. This is obviously not feasible within the boundaries of a PhD project. Therefore, the sample consists of 25 languages – a number that it is feasible to study, yet covers sufficient linguistic diversity. Statistical validity cannot be obtained with a sample of this size, but this need not be problematic since the study aims to establish whether there is an interesting pattern in transparency features in the first place, rather than prove the statistical validity of such a pattern. This is a qualitative study, not a quantitative one.

A third matter in determining the sample is choosing a suitable language classification. In this study, I follow Rijkhoff et al. (1993) in using Ruhlen's (1991) classification, which is based on the work of Greenberg (1957 and later works) and is in fact controversial. Ruhlen's classification is a 'lumping' one, meaning that it groups languages together into relatively few families, while more recent classifications such as the Ethnologue (Lewis 2009) tend to take a 'splitting'

approach by distinguishing many more language families. As such, the latter approaches are often more detailed and justifiable. Furthermore, Ruhlen makes contestable decisions by assuming distant genetic relationships between languages, basing himself on scarce evidence, most notably in his postulation of an Amerind language family, which is generally rejected by contemporary Americanists (cf. Campbell 2013: 346ff.).

Despite these imperfections, I will use Ruhlen's classification because in combination with the relatively small size of the current sample, it allows for sufficient precision for my purposes, which are qualitative rather than quantitative. Firstly, it can supply a stratified sample that includes a feasible number of languages, while use of for example the Ethnologue classification would be more appropriate when using a much larger sample. Also, using Ruhlen's classification increases comparability with other typological work that also takes this classification as its point of departure. Moreover, by taking into account which of Ruhlen's classifications are nowadays considered questionable, the inclusion of languages with a doubtful genetic affiliation can be avoided, thus guaranteeing the validity of the sample. The only remaining problem that using Ruhlen's classification brings about is the possible underrepresentation of languages from the so-called Amerind family, analysed by contemporary linguists as 180 families (Campbell 2013: 367). According to such contemporary classifications, many more American languages should have been included in the sample.

Actually, to do justice to the diversity of the languages in the world with respect to communicative modality, the language sample should include sign languages as well. However, two problems arise when taking sign languages into account. Firstly, Ruhlen's classification does not include any, and secondly, the inclusion of sign languages in the sample would add an extra variable to the cross-linguistic research performed. As interesting as it would be to compare the transparency of sign languages to that of spoken languages, it is beyond the scope of the current study and will therefore be left for future research. For the same reason, the sample does not feature any creole languages.

The application of the Rijkhoff method, combined with the classification by Ruhlen (1991) and a sample size of 25, results in the sample in Table 5.1. This table shows how many languages should be selected from each language family, thereby excluding isolates, pidgins and creoles (as explained), invented languages and unclassified languages. The phyla in the left column are those distinguished by Ruhlen (1991), which is the second edition of Ruhlen's 1987 classification. The only difference between the editions that is relevant here is the fact that Korean-Japanese was classified as a branch of the Austric phylum in 1987, but seen as a separate phylum in the 1991 classification.

**Table 5.1: Distribution languages over phyla according to Diversity Value**

| Phylum (according to Ruhlen 1991) | N languages in sample (total n=25) |
|---|---|
| | |
| Afro-Asiatic | 2 |
| Altaic | 1 |
| Amerind | 2 |
| Austric | 2 |
| Australian | 2 |
| Caucasian | 1 |
| Chukchi-Kamchatkan | 1 |
| Elamo-Dravidian | 1 |
| Eskimo-Aleut | 1 |
| Indo-Hittite | 1 |
| Indo-Pacific | 2 |
| Khoisan | 1 |
| Korean-Japanese | 1 |
| Na-Dene | 1 |
| Niger-Kordofian | 2 |
| Nilo-Saharan | 2 |
| Sino-Tibetan | 1 |
| Uralic-Yukaghir | 1 |

As the sampling method guarantees the genetic and areal diversity of languages, the choice for the selection of specific languages from the families is free. However, two criteria are used for the inclusion of particular languages in the sample. Firstly, the high number and the complex nature of properties investigated in this study require an extensive and thorough description of the language involved. Therefore,

languages for which a high-quality, in-depth reference grammar is available are preferred over less well-described languages. Furthermore, an attempt is made to select languages with reference grammars that have appeared only recently, so that the author can be contacted for more information on matters of detail. This not only enables the study of more in-depth characteristics of languages, that are often not taken up in reference grammars, but also provides additional certainty concerning linguistic judgments: whether a feature is present in a language or not can of course be determined best by a speaker of the language, preferably by a native speaker linguist. Thus, by opting for those languages for which an expert can be contacted, it is possible to include more detailed information on the languages involved.

Secondly, a sample should not only cover a large genetic and areal diversity, but typological diversity too: it is important that the sample covers the full range of possible degrees of transparency and does not include too many languages with a particular degree of transparency, for instance many averagely transparent languages but few extreme cases. Therefore, a first estimate was made of the transparency of the 25 languages that were selected in a first preliminary sample. Since this estimation showed a slight bias toward the non-transparent end of the continuum, some transparent languages were included at the expense of languages estimated to be relatively non-transparent.

Eventually, it proved unfeasible to study twenty-five languages – due to time limitations, the study had to be limited to 22 languages. The originally selected languages Lango, Logba and Slave could not be studied sufficiently to include them in the final data set, which means that no languages from the Na-Dene, and one less language from the Nilo-Saharan and Niger-Kordofanian language families have been included. Obviously, this somewhat reduces the chances of a genetically, areally and typologically balanced sample, but the study still takes into account a large variety of languages with a large range of degrees of transparency, which suffices to reach the goals of this qualitative investigation into transparency.

The sampling methodology described led to the composition of the final sample given in Table 5.2. Each language is listed together with its language family name according to Ruhlen's classification and according to the more commonly

used Ethnologue classification (Lewis 2009). Since languages tend to be known under multiple names, possibly leading to confusion about which language is meant, the Ethnologue code is provided for each language.

**Table 5.2: Languages in sample with Ruhlen and Ethnologue classifications**

| **Language (Ethnologue code)** | **Phylum Ruhlen** | **Phylum Ethnologue** |
|---|---|---|
| Bantawa (bap) | Sino-Tibetan | Sino-Tibetan |
| Bininj Gun-Wok (gup) | Australian | Australian |
| Chukchi (ckt) | Chukchi-Kamchatkan | Chukotko-Kamchatkan |
| Dutch (nld) | Indo-Hittite | Indo-European |
| Egyptian Arabic (arz) | Afro-Asiatic | Afro-Asiatic |
| Fongbe (fon) | Niger-Kordofanian | Niger-Congo |
| Georgian (kat) | Caucasian | Kartvelian |
| Japanese (jpn) | Korean-Japanese | Japonic |
| Huallaga Quechua (qub) | Amerind | Quechuan |
| Kayardild (gyd) | Australian | Australian |
| Kharia (khr) | Austric | Austro-Asiatic |
| Khwarshi (khv) | Caucasian | North Caucasian |
| Kolyma Yukaghir (ykg) | Uralic-Yukaghir | Yukaghir |
| Samoan (smo) | Austric | Austronesian |
| Sandawe (sad) | Khoisan | Khoisan |
| Sheko (she) | Afro-Asiatic | Afro-Asiatic |
| Sochiapan Chinantec (cso) | Amerind | Oto-Manguean |
| Tamil (tam) | Elamo-Dravidian | Dravidian |
| Teiwa (twe) | Indo-Pacific | Trans-New Guinea |
| Tidore (tvo) | Indo-Pacific | West-Papuan |
| Turkish (tur) | Altaic | Altaic |
| West-Greenlandic (kal) | Eskimo-Aleut | Eskimo-Aleut |

## 5.3 Methodology: implicational hierarchies

In this study, the phenomenon of transparency is approached from a typological perspective by performing a cross-linguistic comparison. This type of descriptive typological research is able to determine principles and constraints that underlie natural languages by examining the range of language variation, thus showing which phenomena are possible in language and which are not. Examining the limits to variation is of use in uncovering language universals, because after all, if some phenomenon is attested in practically all languages in the world, we may conclude that this phenomenon fulfils a certain important function, whereas an unattested phenomenon presumably violates some universal principle. Croft (2003b: 341) argues that cross-linguistic comparison is the way to empirically establish generalisations, leading to these underlying principles and constraints.

The theory of Generative Grammar states that such universal principles reside in Universal Grammar: a language blueprint that is innate and unique to human cognition as opposed to the cognition of other animal species (e.g. Chomsky 1988, cf. Wasow 2003 and Haspelmath 2008 for explanations of the theoretical viewpoint). Since Universal Grammar is assumed to underlie all languages, whether its properties surface or not, it is sometimes considered sufficient to study the underlying principles of one language to identify them. Once found, the principles can be extrapolated to all other languages.

Contrariwise, functional theories of language seek universal principles in the universal function of language, i.e. communication: the transfer of ideas from a speaker to a hearer (cf. Van Valin 2003). The idea behind this is that optimal communication is dependent on certain universal factors that are irrespective of the language. For example, articulatory ease is communicatively advantageous, at least for the speaker, whether she is speaking Dutch or Sandawe. Similarly, when looked at from the hearer's perspective, a large auditory distance between phonemes is communicatively convenient, whether listening to Sochiapan Chinantec or Kharia. Such communicative principles apply to all languages of the world, and the form that languages take can be explained as the result of the interplay between

conflicting principles. They can typically be discovered by looking at the similarities between languages, and this is the reason for the traditional connection between typology and linguistic functionalism (Croft 2003a, 2003b, Van Valin 2003, Hengeveld & Mackenzie 2008: 31-37). The current research fits into this natural link between cross-linguistic comparison and an explanation of linguistic form in terms of its communicative function.

As argued above, the typological research tradition aims to detect language universals and the principles underlying them by comparing genetically unrelated languages. The most basic type of language universal is the absolute one, which states that 'all languages have trait X'. Implicational universals, a subset of linguistic universals introduced by Greenberg (1963), allow more nuances in such statements by taking the form 'all languages with trait Y, also have trait X' (Croft 2003b: 344). A series of implicational universals results in an implicational hierarchy of the following form: 'trait X implies trait Y, which itself implies trait Z', and so on. Consider for example the hierarchy of nasals: ɲ → m → n. This should be read as follows: if a language has a phoneme /ɲ/ in its inventory, then it also has the phonemes /m/ and /n/. If /m/ is in the phoneme inventory, this implies that /n/ is present as well, but it does not say anything about the presence of /ɲ/. Similarly, the presence of /n/ in a language's inventory is not informative on the presence of /m/ and /ɲ/ (Croft 2003a: 159). An arrow thus represents the statement 'implies the presence of', which should in this dissertation always be read from left to right and from top to bottom.[9]

In many domains, it has been found that different implicational hierarchies crosscut each other, when one feature is part of multiple implicational hierarchies. This is expected to be the case in the domain of transparency as well, as Leufkens (2013a) has shown that within an overall opacity hierarchy, subgroupings play a role

[9] In some typological studies, implicational relations are indicated by means of set theory symbols rather than arrows. This comes down to the same, since in the case of an implicational relation A → B, languages exhibiting feature A form a subset of languages exhibiting feature B. I have chosen to use arrows in this dissertation for reasons of readability.

as well (cf. Section 5.4). The results, then, cannot be represented in a one-dimensional implicational hierarchy, but will take the form of two- or multi-dimensional hierarchies. An abstract example of a two-dimensional hierarchy is given as 1), which combines two implication hierarchies viz. A → B → C, the second B → X → Y. Obviously, the two hierarchies cross cut in feature B, a feature of which the presence implies the presence of features A, X and Y. Note that this two-dimensional hierarchy makes no claims on implicational relation between X and A, or between Y and A.

1) A → B → C
↓
X
↓
Y

Apart from constraining the set of possible languages, implicational generalisations can provide insight into the relation between the implicationally related traits. If the presence of one feature implies the presence of another, this could of course very well be the result of a common explanation or even a causal relationship between the two. For example, the finding that the presence of an /m/ in a language's phoneme inventory implies the presence of /n/ might indicate that there is an articulatory advantage of /n/ over /m/ that makes speakers more prone to adopt the former into their language.

Some quantitative typologists warn against drawing such conclusions too fast. Most notably, Cysouw (2003) shows how alleged implicational relations can be based on mere frequency effects, instead of on a meaningful relation between traits. Unwarranted conclusions about causal relations between features can be avoided, according to Cysouw, by calculating the deviation between an expected feature distribution and the actually attested feature distribution. Even though the point raised by Cysouw should be taken seriously, it will not be taken into account here, as it is not relevant to the type of study executed. The current study is a qualitative

investigation into the nature of transparency, and aims to explore the range and distribution of degrees of non-transparency in a variety of languages.

## 5.4 Hypotheses and expected outcomes

To investigate the first research question, a typological investigation will compare the transparency of 22 natural languages, by checking whether they display the non-transparent properties listed in Chapter 4. The results will be ordered on two scales: a first scale orders the sampled languages with respect to their degree of transparency, measured as the amount of transparent features in the language, and a second scale will order the non-transparent properties from those appearing most often, to those appearing least frequently.

The main hypothesis of this dissertation is that the two scales will be related, in the sense that the degree of transparency of a language can predict not only how many, but also which non-transparent features it possesses. In other words, the hypothesis is that non-transparent features do not appear randomly over languages. Rather, correlations between certain non-transparent features and the degree of transparency of languages are expected, e.g. between grammatical gender and a low degree of transparency. This pattern of correlations can be captured in an implicational hierarchy of opaque features.

A few studies have been carried out into the transparency of specific languages, viz. Leufkens (2010), Hengeveld (2011b), and Leufkens (2013a). On the basis of these studies, some expectations exist as regards correlations between particular features and degrees of transparency. With regard to research question 1), some features are consistently found at the left side of an implicational hierarchy, i.e. they are only attested in relatively non-transparent languages. This concerns the features grammatical gender, syntactic agreement, and tense copying. On the other hand, all studies show that phonological assimilations, influence of complexity on word order, and apposition exist in all natural languages, locating them on the right side of an implicational hierarchy. Other similarities and differences between

attested hierarchies will be discussed below in the discussion of expectations for the second and third research questions.

Leufkens (2010) and Leufkens (2013a) consistently show that non-transparent properties of the redundancy category appear in all languages, even in the most transparent ones. Fusion and domain disintegration, seen as a single category in these papers, are frequently attested too, but form-based form features turn out to be rare. This leads to a first hypothesis on implications between categories of features, given in 2).

2) Form-based form → Domain disintegration → Redundancy

The hierarchy in 2) thus represents the finding that if a language exhibits one or more form-based form phenomena, there are also domain disintegration and redundancy features. If a language exhibits domain disintegration (including fusion) somewhere in its grammar, it is implied that there is redundancy as well, but no form-based form needs to be present. If a language exhibits redundancy, there need not be any domain disintegration or form-based form features in its grammar.

These findings provide a preliminary answer to research question 2), as they show that different types of mismatches (redundancy, domain disintegration, fusion and form-based form) can be ordered relative to each other.

Taking a closer look at redundancy, some expectations arise as to the ordering of features grouped in that category. Leufkens (2010) finds no clear pattern, but Leufkens (2013a) suggests that negative concord and semantic class concord imply plural concord. Furthermore, cross-reference is only exhibited by languages that have plural concord, indicating that there might be an implicational relation there too. However, counter examples exist to both patterns, so that more information is necessary to test whether these implications really hold. Hengeveld (2011b) does not go into these features.

Within the domain disintegration category, Leufkens (2010) finds some regularity as to the distribution of these features over languages, given in 3).

3) Non-parallel alignment → Discontinuity → Fusional morphology

Within fusional morphology, that includes both cumulation and fusional stem alternation in Leufkens (2010), the presence of stem alternation was found to imply the presence of cumulation, leading to the two-dimensional hierarchy in 4).

4)

| | | |
|---|---|---|
| | | Stem alternation |
| Non-parallel alignment → Discontinuity → | | ↓ |
| | | Cumulation |

Note that the presence of discontinuity in a language does not imply the presence of stem alternation – it only predicts the presence of cumulation. Only the presence of a type of fusion is implied, but not necessarily the presence of both fusion features.

Hengeveld (2011b) finds the pattern in 5), which can be seen as a collapsed version of 4).

5) Discontinuity → Stem alternation → Non-parallel alignment → Cumulation

An obvious difference between Leufkens' and Hengeveld's findings is the relative order of non-parallel alignment and discontinuity. A reason for this discrepancy might be a lack of information in Leufkens' (2010) study: the sample consisted of four languages and for one of those, no conclusive evidence on non-parallel alignment was available. The current study will not be able to solve this issue, since non-parallel alignment is excluded because of a lack of data.

Concerning form-based form, Leufkens (2010) finds an ordering of features as given in 6). Features in the hierarchy that are not studied in this dissertation, are left out.

6) Syntactic functions,
Agreement,
Grammatical gender,
Tense copying,
↓
Expletives
↓
Influence of complexity on word order,
Phonological assimilation (comparable to phonologically conditioned alternations)

Hengeveld (2011b) also finds agreement, grammatical gender and tense copying to be the most infrequent non-transparent properties, but syntactic functions appear lower in his hierarchy. Nominal expletives, on the other hand, are only attested in the most opaque languages in Hengeveld's sample. Leufkens (2013a) corroborates the rareness of agreement, grammatical gender and tense copying, as well as expletives. Syntactic functions take a middle position, resulting in a hierarchy that is altogether in line with Hengeveld (2011b).

As to the expectations for research question 3), the results from previous studies have not given concrete predictions yet. So far, there is no indication that interface has an effect on the cross-linguistic distribution of transparency, nor that subgroupings exists between meaning-to-form opacity features on the one hand, and other opaque features on the other. It does appear to be the case that phonologically based form-based form stands fundamentally lower in the hierarchy than other features, because all studies so far find phonological assimilation to be present in all sample languages. This suggests that languages are more prone to develop phonological non-transparency than opacity higher up in their grammars, in line with the fact that phonological change happens more easily than morphological change.

The three studies on which these expectations are based, i.e. Leufkens (2010), Hengeveld (2011) and Leufkens (2013a) are all based on small samples, i.e.

on 4, 5 and 11 languages respectively. The current study will test the prediction on more than twice as much languages and will thus be able to correct conflicting evidence and corroborate earlier findings. Thus, it will make the patterns attested so far more robust and give them more scientific support.

# Chapter 6
# Results and discussion

This chapter provides the answers to the research questions posed in Section 5.1, repeated here:

1) How are non-transparent features distributed cross-linguistically?
2) How are redundancy, fusion, domain disintegration and form-based form features distributed cross-linguistically?
3) How are non-transparent features at different interfaces distributed cross-linguistically?

The answers to these questions involve a large amount of typological data, which can be found in the appendix. For each sample language a sketch is provided containing information on all phenomena investigated as they manifest themselves in the language concerned. Section 6.1 addresses research question 1 by providing a schematic overview of all the data, and presents an implicational hierarchy of transparency. Furthermore, the section gives an explanation for the results in terms of the crucial notion of syntacticity, as well as an explanation for expectations that are not borne out. Section 6.2 shows the data again, now presented according to the type of violation of transparency, in order to answer the second research question. The answer to the third research question is presented in Section 6.3, in which the results are categorised according to the interface at which they apply.

## 6.1 How are non-transparent features distributed cross-linguistically?

This section will present the results of the comparative study into transparency. Section 6.1.1 will give an overview of the data and discuss the way in which they are presented. In Section 6.1.2, I will go into the pattern that is attested in the cross-linguistic distribution of transparency features, and provide an explanation for that

particular pattern in terms of syntacticity in Section 6.1.3. Finally, Section 6.1.4 explains why some features do not conform to this cross-linguistic pattern.

### *6.1.1 An overview of data*

As explained in Chapter 5, my study examines whether or not the languages in the sample are transparent with respect to the list of non-transparent features established in Chapter 4. The concrete results of this study, that is, the analyses of the non-transparent features in these languages including examples, can be found online in the form of an open access database, retrievable at transparency.humanities.uva.nl, so as to make them continuously accessible and easily searchable for an as large as possible audience, in line with current scientific research standards. A schematic overview of the data is given in Table 6.1.

Table 6.1, as well as other tables in this chapter, should be read as follows. The upmost row gives the features studied, following the overview in Section 4.5, including between brackets the number of languages that display that feature. The features, i.e. the columns, are ranked from the least attested feature at the left, to the most attested feature in the rightmost column. Each row represents a language, ordered from the most non-transparent language, i.e. the language with most non-transparent features, at the top, to the most transparent language in the lowest row.

A plus sign in a cell indicates that the feature is present in the language, in other words, that the language under consideration is non-transparent with respect to that feature. A minus shows the opposite, i.e. that the feature is absent from the language, which is therefore transparent with regard to that feature. A plus sign between brackets indicates that the feature is optionally present; in other words, that the non-transparent phenomenon is allowed to occur but does not always do so, in contrast with a minus sign which shows that the phenomenon is disallowed entirely.

The abbreviation ‘n.a.’ stands for ‘not applicable’ and means that it is impossible for the language to be classified as transparent or opaque regarding a certain feature, because an element that is necessary for the feature to apply is not available in the language. For example, a language without tense marking cannot possibly display tense copying, so that it would not make sense to count it as

transparent or opaque regarding that feature – it is neither. This is different from the value ‘n.d.’, which stands for ‘non determinable’ and means that the crucial elements are present in the language, but both to the same extent, so that it cannot be determined whether the language is transparent or opaque regarding that feature. This label applies to the feature ‘predominant head-marking’ only and is attributed to languages with a roughly equal amount of head-marking affixes and phrase-marking clitics/particles.

One further note should be made with regard to this overview: as explained in Section 5.1, features are assessed in a binary fashion, rather than as gradual phenomena. For example, stem alternation is much less widespread in Sheko than it is in Sochiapan Chinantec, but this gradual difference is not made visible in the table as they both receive a plus sign. Of course, a more fine-grained measure would be preferable, but since no graded scales are available for the features under consideration, this was outside the scope of this study.

**Table 6.1: Basic data: Distribution of non-transparent (sub)features over sample languages**

| | Nominal expletive elements (1) | Clausal agreement (1) | Tense copying (2) | Grammatical gender (2) | Circumfixes (5) | Infixes (5) | Raising (5) | Morphologically conditioned stem alternation: suppletion (8) | Morphosyntactic complexity influences word order (9) | Plural concord in noun phrases containing a numeral (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| Dutch (20) | + | + | + | + | + | - | + | + | + | (+) |
| Egyptian Arabic (18) | - | - | - | + | + | + | - | + | + | (+) |
| Georgian (15) | - | - | + | - | + | - | - | + | - | - |
| Khwarshi (14) | - | - | - | - | - | + | - | + | + | - |
| Bininj Gun-Wok (12) | - | - | - | - | - | - | - | + | - | - |
| West-Greenlandic (12) | - | - | - | - | - | - | - | - | + | + |
| Sochiapan Chinantec (12) | - | - | - | - | - | - | - | + | + | na |
| Tamil (12) | - | - | - | - | - | - | - | - | + | (+) |
| Bantawa (12) | - | - | - | - | + | + | - | - | - | (+) |
| Chukchi (11) | - | - | na | - | + | - | na | - | - | na |
| Sheko (11) | - | - | na | - | - | + | - | + | - | - |
| Sandawe (11) | - | - | na | - | - | - | - | + | - | (+) |
| Kayardild (10) | - | na | - | - | - | - | + | - | - | - |
| Huallaga Quechua (10) | - | - | na | - | - | - | + | - | + | (+) |
| Kharia (10) | - | - | - | - | - | + | - | - | + | (+) |
| Turkish (10) | - | - | - | - | - | - | + | - | - | - |
| Kolyma Yukaghir (9) | - | - | na | - | - | - | - | - | - | - |
| Samoan (8) | - | - | - | - | - | - | - | - | - | na |
| Japanese (7) | - | na | - | - | - | - | - | - | - | (+) |
| Tidore (7) | - | - | na | - | - | - | - | - | + | na |
| Fongbe (6) | - | na | - | - | - | - | + | - | - | (+) |
| Teiwa (5) | - | - | na | - | - | - | - | - | - | - |

| | Predominantly head-marking (10) | Extraction and extraposition (10) | Phrasal agreement (10) | Morphologically conditioned stem alternation: Irregular Stem Formation (12) | Cumulation of TAME and/or case (14) | Syntactic functions: (Anti-)passive (14) | Cross-reference (18) | Morphophonologically conditioned stem alternation (20) | Morpho(phono)logically conditioned affix alternation (21) | Syntactic functions: neutralisation (20) | Phonologically conditioned stem alternation (22) | Phonologically conditioned affix alternation (22) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dut | + | + | + | + | + | + | - | + | + | + | + | + |
| Egy | + | + | + | + | + | + | + | + | + | + | + | + |
| Geo | + | + | + | + | + | + | + | + | + | + | + | + |
| Khw | + | + | + | + | + | - | + | + | + | + | + | + |
| Bin | + | + | + | + | + | - | + | + | + | + | + | + |
| W-Gr | nd | + | + | - | + | + | + | + | + | + | + | + |
| Soc | nd | - | + | + | + | + | + | + | + | + | + | + |
| Tam | + | - | - | + | - | + | + | + | + | + | + | + |
| Ban | + | - | - | - | + | + | + | + | + | + | + | + |
| Chu | + | - | na | + | + | + | + | + | + | + | + | + |
| She | nd | - | + | + | - | + | + | + | + | + | + | + |
| San | nd | - | + | + | + | - | + | + | + | + | + | + |
| Kay | + | - | + | + | - | + | na | + | + | + | + | + |
| Que | - | - | - | - | - | + | + | + | + | + | + | + |
| Kha | - | - | - | - | + | + | + | + | + | - | + | + |
| Tur | nd | + | - | - | + | + | + | + | + | + | + | + |
| Kol | + | + | - | - | + | - | + | + | + | + | + | + |
| Sam | - | - | - | + | + | - | + | + | + | + | + | + |
| Jap | nd | - | - | - | - | + | na | + | + | + | + | + |
| Tid | nd | - | - | - | na | - | + | + | + | + | + | + |
| Fon | - | + | - | - | + | na | na | - | + | na | + | + |
| Tei | - | na | - | - | - | - | + | - | - | + | + | + |

In Table 6.1, all individual features are presented separately, in order to give a complete visualisation of all features studied at the lowest level. Table 6.2 presents the same data, but now particular related features are combined into four combinatorial features, and again ordered on the basis of frequency of occurrence (features) and number of non-transparent features (languages). The reasoning behind making these particular combinations of features is as follows.

First of all, 'Syntactic functions: neutralisation of semantic roles in intransitive clauses' is combined with 'Syntactic functions: neutralisation of semantic roles in (anti)passive constructions', since obviously, these features both measure the presence of a syntactic function subject.

Furthermore, the features 'Circumfixes' and 'Infixes' are combined into 'Discontinuous morphology'. These features are clearly related as well, since they involve or create discontinuous morphemes. For similar reasons, 'Extraction and extraposition', 'Raising', and 'Influence of morphosyntactic weight on word order' are combined into one combinatorial feature called 'Morphosyntactically induced displacement'. These three features are related in the sense that they create a mismatch between semantics and morphosyntax by means of a violation of the domain integrity of clauses (Raising) or phrases (Extraction and extraposition), or by means of a violation of semantically determined constituent ordering (Influence of morphosyntactic weight on word order). Furthermore, extraposition is often a direct consequence of an influence of morphosyntactic weight on ordering principles, so that it makes sense to integrate these features.

Finally, Cross-reference is combined with Clausal agreement into one combinatorial feature called 'Redundant referential marking'. The only difference between these features, as explained in Section 3.1.4, is whether the argument is obligatorily explicit (clausal agreement) or not (cross-reference), but they both measure the multiple expression of (a property of) a referent. Note that this is an improvement on Table 6.1, in which the separate cross-reference row fails to discriminate between languages without any redundant marking (e.g. Teiwa) and languages with clausal agreement (Dutch). Note furthermore that clausal agreement is still presented in a separate row as well, since it has a special status as being

analysed in FDG as a purely morphosyntactic copying procedure. Thus, it qualifies not only as a redundancy feature, but can be seen as a form-based form feature as well.

**Table 6.2: Basic data: Distribution of non-transparent features and combinatorial features over sample languages**

| | Nominal expletive elements (1) | Clausal agreement (1) | Tense copying (2) | Grammatical gender (2) | Discontinuous morphology (8) | Morphologically conditioned stem alternation: suppletion (8) | Plural concord in NPs containing a numeral (10) | Predominantly head-marking (10) |
|---|---|---|---|---|---|---|---|---|
| Dutch (18) | + | + | + | + | + | + | (+) | + |
| Egyptian Arabic (15) | - | - | - | + | + | + | (+) | + |
| Georgian (14) | - | - | + | - | + | + | - | + |
| Khwarshi (13) | - | - | - | - | + | + | - | + |
| Bininj Gun-Wok (12) | - | - | - | - | - | + | - | + |
| S. Chinantec (11) | - | - | - | - | - | + | na | nd |
| Sandawe (11) | - | - | na | - | - | + | (+) | nd |
| West-Greenlandic (10) | - | - | - | - | - | - | + | nd |
| Bantawa (10) | - | - | - | - | + | - | (+) | + |
| Kharia (10) | - | - | - | - | + | - | (+) | - |
| Chukchi (10) | - | - | na | - | + | - | na | + |
| Tamil (10) | - | - | - | - | - | - | (+) | + |
| Sheko (10) | - | - | na | - | + | + | - | nd |
| Kolyma Yukaghir (9) | - | - | na | - | - | - | - | + |
| Kayardild (9) | - | na | - | - | - | - | - | + |
| Huallaga Quechua (8) | - | - | na | - | - | - | (+) | - |
| Turkish (8) | - | - | - | - | - | - | - | nd |
| Samoan (8) | - | - | - | - | - | - | na | - |
| Tidore (7) | - | - | na | - | - | - | na | nd |
| Japanese (6) | - | na | - | - | - | - | (+) | nd |
| Fongbe (5) | - | na | - | - | - | - | (+) | - |
| Teiwa (5) | - | - | na | - | - | - | - | - |

| | Phrasal agreement (10) | Morphologically conditioned stem alternation: irregular stem formation (12) | Cumulation of TAME and/or case (14) | Morphosyntactically induced displacement (16) | Redundant referential marking (19) | Morphophonologically conditioned stem alternation (20) | Morpho(phono)logically conditioned affix alternation (21) | Syntactic function subject (21) | Phonologically conditioned affix alternation (22) | Phonologically conditioned stem alternation (22) |
|---|---|---|---|---|---|---|---|---|---|---|
| Dut | + | + | + | + | + | + | + | + | + | + |
| Egy | + | + | + | + | + | + | + | + | + | + |
| Geo | + | + | + | + | + | + | + | + | + | + |
| Khw | + | + | + | + | + | + | + | + | + | + |
| Bin | + | + | + | + | + | + | + | + | + | + |
| Soc | + | + | + | + | + | + | + | + | + | + |
| San | + | + | + | - | + | + | + | + | + | + |
| W-Gr | + | - | + | + | + | + | + | + | + | + |
| Ban | - | - | + | - | + | + | + | + | + | + |
| Kha | - | - | + | + | + | + | + | + | + | + |
| Chu | na | + | + | - | + | + | + | + | + | + |
| Tam | - | + | - | + | + | + | + | + | + | + |
| She | + | + | - | - | + | + | + | + | + | + |
| Kol | - | - | + | + | + | + | + | + | + | + |
| Kay | + | + | - | + | na | + | + | + | + | + |
| Que | - | - | - | + | + | + | + | + | + | + |
| Tur | - | - | + | + | + | + | + | + | + | + |
| Sam | - | + | + | - | + | + | + | + | + | + |
| Tid | - | - | na | + | + | + | + | + | + | + |
| Jap | - | - | - | - | na | + | + | + | + | + |
| Fon | - | - | - | + | na | - | + | na | + | + |
| Tei | - | - | - | - | + | - | - | + | + | + |

In Table 6.2, a remarkable pattern reveals itself: for quite a number of features, plus signs appear mostly in the upper part of the table, while minuses appear at the bottom. Precisely these features are, in conjunction, able to distinguish the degrees of transparency of languages in a fine-grained fashion. In other words, these features show degrees of 'seriousness' of non-transparency: while some non-transparent features are attested in all languages, others occur only rarely and are indicators of a high degree of opacity of the language in which they occur. The following section goes into these features and the pattern that they reveal.

### *6.1.2 An implicational hierarchy of transparency*

Table 6.3 provides an overview of the (combinatorial) features that show a clear boundary between pluses and minuses, thus discriminating degrees of transparency of languages. Note that the rows in this table are no longer ordered on the basis of frequency of occurrence and number of non-transparent features, but according to an underlying pattern, which is outlined by means of a bold line and which I will discuss in detail below. Such a pattern, in which binary features show a distributional ordering, is known in research design as a Guttman scale (Abdi 2010) and demonstrates the correlation between the relative transparency of languages and the particular non-transparent features that languages exhibits, thus enabling the integration of these values in an implicational hierarchy. In Table 6.3, the languages are ordered as to their degree of transparency, Dutch being the absolute winner regarding opacity as it displays all non-transparent phenomena tested, and Teiwa being the most transparent language in the sample. Counter examples to the pattern are indicated by shaded cells and will be discussed below.

Table 6.3 reveals a strikingly strong pattern, marked by the bold line. This pattern is captured by the implicational hierarchy given in 1), in which the appearance of a particular feature implies the presence of all features below it. Features separated by a comma cannot be ranked with respect to each other.

1) Nominal expletives, Clausal agreement
→ Grammatical gender, Tense copying
→ Morphologically conditioned stem alternation: suppletion
→ Phrasal agreement,
Morphologically conditioned stem alternation: irregular stem formation
→ Predominant head-marking
→ Morphophonologically conditioned stem alternation
→ Morpho(phono)logically conditioned affix alternation
→ Redundant referential marking,
Phonologically conditioned stem alternation,
Phonologically conditioned affix alternation,
Syntactic function subject

The implicational pattern in Table 6.3 is strong, but as is common in this type of research, a few counter examples are attested, which I will discuss one by one. First of all, there are two cases of conflicting results that make it impossible to rank certain pairs of features or languages with respect to each other, both indicated by a dotted line in Table 6.3. The first pair involves Tense copying and Grammatical gender, because these features are complementarily present in Egyptian Arabic and Georgian. An explanation for this could be that these features are indicators of an equal degree of non-transparency, but that Grammatical gender is an expression of opacity in the nominal domain (Egyptian Arabic), while Tense Copying is an expression of non-transparency in the verbal domain (Georgian). Thus, Georgian and Egyptian are equally non-transparent, but their opacity surfaces in different domains.

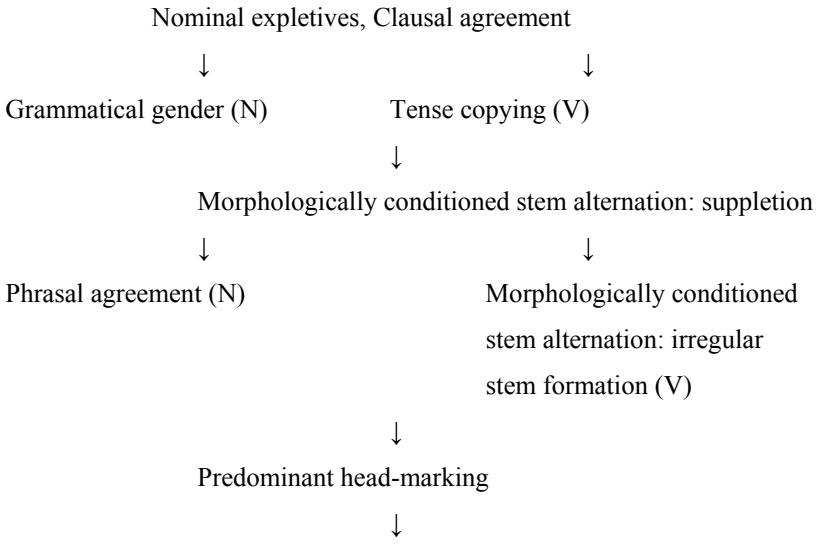
**Table 6.3: Features showing an implicational relationship**

| | Nominal expletive elements | Clausal agreement | Grammatical gender | Tense copying | Morphologically conditioned stem alternation: Suppletion | Phrasal agreement | Morphologically conditioned stem alternation: Irregular Stem Formation |
|---|---|---|---|---|---|---|---|
| Dutch | + | + | + | + | + | + | + |
| Egyptian Arabic | - | - | + | - | + | + | + |
| Georgian | - | - | - | + | + | + | + |
| Sochiapan Chinantec | - | - | - | - | + | + | + |
| Khwarshi | - | - | - | - | + | + | + |
| Bininj Gun-Wok | - | - | - | - | + | + | + |
| Sandawe | - | - | - | na | + | + | + |
| Sheko | - | - | - | na | + | + | + |
| Chukchi | - | - | - | na | - | na | + |
| Kayardild | - | na | - | - | - | + | + |
| West-Greenlandic | - | - | - | - | - | + | - |
| Tamil | - | - | - | - | - | - | + |
| Kolyma Yukaghir | - | - | - | na | - | - | - |
| Bantawa | - | - | - | - | - | - | - |
| Kharia | - | - | - | - | - | - | - |
| Turkish | - | - | - | - | - | - | - |
| Huallaga Quechua | - | - | - | na | - | - | - |
| Tidore | - | - | - | na | - | - | - |
| Japanese | - | na | - | - | - | - | - |
| Samoan | - | - | - | - | - | - | + |
| Fongbe | - | na | - | - | - | - | - |
| Teiwa | - | - | - | na | - | - | - |

| | Predominantly head-marking | Morphophonologically conditioned stem alternation | Morpho(phono)logically determined affix alternation | Redundant referential marking | Syntactic function subject | Phonologically conditioned affix alternation | Phonologically conditioned stem alternation |
|---|---|---|---|---|---|---|---|
| Dut | + | + | + | + | + | + | + |
| Egy | + | + | + | + | + | + | + |
| Geo | + | + | + | + | + | + | + |
| Soc | nd | + | + | + | + | + | + |
| Khw | + | + | + | + | + | + | + |
| Bin | + | + | + | + | + | + | + |
| San | nd | + | + | + | + | + | + |
| She | nd | + | + | + | + | + | + |
| Chu | + | + | + | + | + | + | + |
| Kay | + | + | + | na | + | + | + |
| W-Gr | nd | + | + | + | + | + | + |
| Tam | + | + | + | + | + | + | + |
| Kol | + | + | + | + | + | + | + |
| Ban | + | + | + | + | + | + | + |
| Kha | - | + | + | + | + | + | + |
| Tur | nd | + | + | + | + | + | + |
| Que | - | + | + | + | + | + | + |
| Tid | nd | + | + | + | + | + | + |
| Jap | nd | + | + | na | + | + | + |
| Sam | - | + | + | + | + | + | + |
| Fon | - | - | + | na | na | + | + |
| Tei | - | - | - | + | + | + | + |

A similar problem regarding the ranking of features in Table 6.3 concerns the presence of irregular stem formation in Tamil versus the presence of phrasal agreement in West Greenlandic. This conflict, too, can be seen as an expression of opacity in different domains: phrasal agreement pertains to NPs, while irregular stem formation in Tamil applies to the verbal domain, since infinitives and causatives are in some cases marked by means of irregular stem formations in that language. Thus, an equal degree of opacity in languages is expressed either in the nominal domain (West-Greenlandic) or in the verbal domain (Tamil). Implicational hierarchy 1) is updated in 2) to show this difference between domains.[10]

2)




[10] Since Dutch has both phrasal agreement and tense copying, it is possible to argue that an implicational relation holds between nominal expletives on the one hand and grammatical gender *and* tense copying on the other. This means that if a language has nominal expletives, it is expected to have both grammatical gender and tense copying, rather than one of those features. However, Mulder (2013) finds that Hebrew has expletives and tense copying, but no grammatical gender. Apparently then, only one of the features grammatical gender/tense copying needs to be present in order for the implication to hold.

Morphophonologically conditioned stem alternation
↓
Morpho(phono)logically conditioned affix alternation
↓
Redundant referential marking,
Phonologically conditioned stem alternation,
Phonologically conditioned affix alternation,
Syntactic function subject

One other counter example occurs that does not have consequences for the ranking of features, but is just an isolated case of a feature not in line with the overall cross-linguistic pattern. This concerns irregular stem formation in Samoan, in which plurality is expressed on a small set of stems by means of vowel lengthening, e.g. *palalū* 'flap.SG' versus *pālalū* 'flap.PL' (Mosel & Hovdhaugen 1992: 237). This is unexpected according to the distribution in Table 3, as other languages with a degree of transparency similar to that of Samoan do not display irregular stem formation at all. The phenomenon is rather marginal, since it applies to a small set of stems only (Mosel and Hovdhaugen list 24 verbal and 1 nominal stem), and many of the lengthened plurals are considered formal and archaic – Mosel & Hovdhaugen state that they are "usually not found in modern Samoan", so that Samoan appears to be moving in the direction of confirming the attested hierarchy.

### *6.1.3 Explanations for transparency and opacity*

Thus far, I have given an answer to research question 1) by demonstrating the cross-linguistic distributional pattern of transparency in the data. Of course, the question that now inevitably comes to mind is: why this particular ordering? Why is it the case that only the most non-transparent language displays nominal expletives and clausal agreement, while phonologically conditioned alternations appear in all sample languages? What is the driving force behind this striking correlation between degrees of transparency of languages and particular non-transparent features?

To answer these questions, observe that the four features on the top of the hierarchy are form-based form and obligatory redundancy features. Exactly such features were argued in Section 2.4 and 4.1.3 to have a high degree of syntacticity: they may have originated as pragmatically and semantically motivated items or constructions, but they diachronically abstracted away from their meaning and became obligatory grammatical elements and rules. In fact, these features are what Dahl (2004) considers to be maturation phenomena, i.e. features that take time to develop in languages, as they start out with a clear function but lose that function along the way (cf. Section 3.2.1). Further down the hierarchy, we find more strongly semantically motivated processes such as redundant referential marking, and lower-level processes such as phonological assimilations. The implicational hierarchy in 2), then, largely reads as a continuum of syntacticity, the most syntactic features being located at the top, fusion and optional redundancy features in the middle, and several (morpho)phonological alternation processes at the bottom.

The only counter example to an explanation of the implicational hierarchy in terms of syntacticity is the feature 'syntactic function subject', which is located at the bottom of the hierarchy, while obviously being a highly syntactic feature – the presence of a 'subject' in a language means that alignment in that language abstracts away from pragmatic and/or semantic roles, in favour of a syntactic function. The fact that it is attested in all languages can be related to the fact that there is a high coincidence between subjects and topics (cf. Li & Thompson 1976), so that alignment expression that appears to be purely syntactically motivated is in fact sometimes partly pragmatically motivated after all.

As explained in Section 4.4.3, a distinction can be drawn between purely syntactic subjects, topics that are the expression of information from the pragmatic level, and, thirdly, topics that are the expression of information from the Contextual Component. Both types of topics are transparent, since they are functions motivated by (extra-linguistic) discourse information. However, FDG is not able to discriminate between syntactic subjects and the second type of topics, that is, those that reflect information from the Contextual Component, which is extra-linguistic. Perhaps, if I had been able to make this distinction between true syntactic subjects

and contextually motivated topics, for example by using Li & Thompson's criteria to distinguish subjects from topics, relatively transparent languages would have shown to be relatively more topic-prominent, thus showing a lower degree of syntacticity and a higher degree of relevance of context and discourse activation. A first indication that this might indeed be the case can be derived from Li & Thompson's classification of Japanese as a language in which topics are as prominent as subjects – apparently, a relatively transparent language as Japanese shows a relatively high degree of contextual dependency in its argument alignment system.

I have thus far explained the ordering of non-transparent features in hierarchy 1) by means of the notion of syntacticity. Now, I still need to explain why languages differ in their degrees of (non-)transparency. The languages at the left side of Table 6.3 have developed highly syntacticised structures, even though this is disadvantageous from a learnability perspective. Why would a language develop in a non-transparent direction at all? Why does Dutch display all non-transparent features, while Fongbe and Teiwa have so few of them?

Firstly, some of the non-transparent features studied may be dysfunctional in terms of transparency of the system, but nonetheless have a communicative advantage that makes them 'profitable' for a language. Let's for example consider syntactic copying, which is the mechanism behind clausal agreement and tense copying. Both features presumably originate from pragmatic or semantic doubling rules, which typically have exceptions and conditions. For example, tense copying is in many languages conditioned by continuing applicability, a semantic notion that creates exceptions to an otherwise morphosyntactic rule (cf. Comrie 1986). Now, when such a rule abstracts away from semantics by ignoring semantic conditions, it becomes syntactic and non-transparent, but also, more regular – the conditions and exceptions will be lost. A similar argument can be made for clausal agreement, as being a regularised form of cross-reference. Both involve multiple expression of reference, but in the case of cross-reference, the explicit expression of this referent is optional. Expressing this referent explicitly in all clauses entails a decrease of transparency, but an increase in regularity, as the speaker no longer has to take into

account pragmatic considerations such as contextual activation. Hence, the introduction of non-transparent obligatory copying may constitute a communicative advantage, which explains its emergence and survival in a language like Dutch. One reason for the existence of highly opaque phenomena in languages may thus be that those phenomena have a language-internal advantage in terms of regularisation.

A second communicative advantage that several non-transparent features have is that they help in demarcating domains. This applies to agreement features and to all (morpho)phonologically conditioned alternations. Agreement involves copying of a property of one element, the controller, to another element, the target. If both the controller and target are present in the output, especially when they have the same form such as in the Italian example *un-a bell-a regazza* 'INDEF-SG.F pretty-SG.F girl(F).SG' (D. Boeke, personal communication), the cohesion between those elements is visible, thus making their constituency explicit and a sentence more easily parsable. This advantage can explain why agreement is helpful in languages, even though it is non-transparent. A similar explanation holds for (morpho)phonological alternations, as these often apply to a particular domain. For example, vowel harmony in Turkish applies to the morphosyntactic word, and thus, phonological assimilation provides information to the hearer about the boundaries of words, again involving a processing advantage. Hence, demarcation of phonological or morphosyntactic boundaries is a second function that non-transparent features may fulfil in languages, and thus another reason why they would develop despite their disadvantageous opacity.

Once a language has developed certain non-transparent features for one of the reasons given above, it may be expected that those disappear from the language – after all, they are typically hard to learn (cf. Section 3.4.3). Therefore, we still need to answer the question why certain languages remain non-transparent, i.e. why non-transparent features are retained in languages like Dutch, Egyptian Arabic and Georgian. My answer to this question is that non-transparent features need not 'bother' a language, as long they are still learnable. A feature may have lost its pragmatic or semantic content, but that does not mean it has to disappear from the language – highly syntacticised features can easily survive in a language, as

'linguistic male nipples' as Lass (1997) so aptly calls them. Thus, a highly non-transparent language like Dutch just happens to have developed non-transparent features for functional reasons, and as long as a majority of Dutch speakers acquires those features, they will be retained, as will the high degree of non-transparency of the language. The reasons that non-transparent features developed in the first place are language-internal and may differ per feature; the reason they do not disappear is that they are learnable.

I have argued how non-transparent languages come to exist, but the question why languages are as transparent as for example Fongbe and Teiwa are is still open. Of course, how a language develops from non-transparent to transparent has been discussed extensively in Section 3.4.2, in which language contact was shown to be a driving force behind such change. In short, the argument is that under pressure of language change, certain non-transparent features become less learnable or even non-learnable for a majority of the language community, so that in the end those features disappear from the language, increasing its degree of transparency.

However, this cannot be the only reason for transparency, since there are relatively transparent languages, such as Fongbe and Teiwa, that are not the result of extensive language contact, nor can be argued to have undergone a substantially higher amount of language contact than other languages. Such languages appear to retain their high degree of transparency over time. My explanation for these languages' 'preference' for transparency lies in their morphology and phonology, since in fact, they almost always turn out to be isolating. There may be a causal relation here, since isolating languages can be expected to develop fewer syntacticised phenomena. I will go into this in greater detail in Section 6.2.1 in the discussion of the place of fusion in a between-category hierarchy.

### *6.1.4 Features not fitting the hierarchy*

Now that I have extensively discussed the implicational pattern in the cross-linguistic distribution of non-transparent features, it is time to consider the features that do not conform to that pattern. Even though a large number of features do participate in the implicational hierarchy, not all features fit – some appear to be

distributed randomly over languages. Those features are listed in Table 6.4, in which the languages are ordered not with respect to the amount of opaque features that they display, which would not make sense since precisely these features are not rankable and therefore not informative on the relative transparency of the languages, but according to the order following from Table 6.3.

Interestingly, the majority of the non-patterning features involves discontinuity features, as 'Discontinuous morphology' and 'Morphosyntactically induced displacement' are both combinations of features from that category. The latter combinatorial feature also includes 'Influence of morphosyntactic weight on word order', which is a form-based form feature, but as argued above, this feature strongly relates to discontinuity. Apparently, discontinuity is a category of mismatches that does not form part of a typological transparency variable. Even though theoretically, discontinuous morphemes and constituents violate a one-to-one relation between meaning and form and are thus cases of opacity, they do not pattern typologically with other non-transparent features.

How to explain this? Firstly, 'Morphosyntactically induced dislocation' refers to a combination of syntactic features that all have to do with word order. Of course, a large range of factors, most notably pragmatic considerations and processing matters, influences word order. For example, the need to put particular constituents in focus position or to create an easily parsable sentence may overrule the transparent and consistent positioning of constituents. Perhaps, then, this large amount of factors with potential influence on word order obscures the effect of transparency as regards this feature, thus obscuring its place in an implicational hierarchy. The complex interplay of information structure and processing is apparently a stronger determining factor than transparency is.

The diffuse distribution of 'Discontinuous morphology' requires a different explanation, of a morphological rather than syntactic nature. First note that even though there are many counter examples, discontinuous morphology does show a tendency to occur more frequently in non-transparent languages than in transparent ones. This may be a side-effect of the fact that isolating languages, which resist affixation by definition, occur more on the transparent end of the language

continuum, as will be discussed below. In the meanwhile, both circumfixes and infixes show a distribution that does not neatly separate non-transparent from transparent languages – there are many counter examples. For infixes, this can be explained by the morphological preconditions that are necessary for infixation to arise. Yu (2007) distinguishes four possible origins for infixes, viz. metathesis, entrapment of an affix in between a root and a lexicalised affix, mutation of a reduplicated morpheme, and so-called morphological excrescence (affixes emerging without any direct historical antecedent, out of a morphophonological process). Thus, in order to be able to develop an infix at all, a language needs to allow for metathesis, reduplication, lexicalisation or morphological excrescence, and on top of that, an opportunity has to occur for one of those processes to create an infix. These are rather rare circumstances, which may explain why many languages do not have infixes, even though their degree of non-transparency predicts they should have.

The explanation for circumfixes is similar. I am not aware of research into the diachronic origins of circumfixation, but I suppose that a language should at least exhibit suffixes and prefixes for it to be able to develop circumfixes. Hence, languages without prefixes or suffixes cannot show circumfixes, even though they would be expected to do so with regard to their degree of non-transparency. This can explain why for instance Sochiapan Chinantec does not have circumfixes, despite its high degree of opacity: since Sochiapan lacks suffixes (Foris 2000: 6), the preconditions for circumfixation are not met.

There is also a redundancy feature that does not pattern with other non-transparent features, viz. ‘Plural concord in noun phrases containing a numeral’. I believe that concord, i.e. multiple expression of semantic properties in phrases, is a highly common, if not universal phenomenon. The only aspects in which concord shows cross-linguistic variation is which type of concord is allowed, e.g. negative concord, plural concord, modal concord, etc., and the degree to which it is grammaticalised and therefore obligatory. Taking this perspective, the distribution that is found here for plural concord is only one aspect of the entire concord distribution. I expect that if I had included more concord features in the study, I

would have found some type of concord in each language, as was the case in previous studies into transparency (cf. Hengeveld 2011b, Leufkens 2013a).

Finally, the fusion feature ‘Cumulation of TAME and/or case’ also shows a scattered distribution. This, too, may be due to the fact that I took into account cumulation of TAME and case, but left out other kinds of cumulation. If I would have looked at the cumulation of person and number as well, or at the cumulation of tense, mood, aspect and evidentiality with each other, I would most probably have found cumulation in all languages, since presumably, variation lies only in the domain in which it is found. The distribution found for TAME and case, then, is only a small piece of the picture.

**Table 6.4: Features not participating in an implicational hierarchy**

| | Discontinuous morphology | Plural concord in noun phrases containing a numeral | Cumulation of TAME and/or case | Morphosyntactically induced displacement |
|---|---|---|---|---|
| Dutch | + | (+) | + | + |
| Egyptian Arabic | + | (+) | + | + |
| Georgian | + | - | + | + |
| Sochiapan Chinantec | - | na | + | + |
| Khwarshi | + | - | + | + |
| Bininj Gun-Wok | - | - | + | + |
| Sandawe | - | (+) | + | - |
| Sheko | + | - | - | - |
| Chukchi | + | na | + | - |
| Kayardild | - | - | - | + |
| West-Greenlandic | - | + | + | + |
| Tamil | - | (+) | - | + |
| Kolyma Yukaghir | - | - | + | + |
| Bantawa | + | (+) | + | - |
| Kharia | + | (+) | + | + |
| Turkish | - | - | + | + |
| Huallaga Quechua | - | (+) | - | + |
| Tidore | - | na | na | + |
| Japanese | - | (+) | - | - |
| Samoan | - | na | + | - |
| Fongbe | - | (+) | + | + |
| Teiwa | - | - | - | - |

## 6.2 How are redundancy, fusion, domain disintegration and form-based form features distributed cross-linguistically?

In the previous section, it was shown how a number of non-transparent features participate in an implicational hierarchy. It is expected that this this also holds for larger combinations of features, i.e. for the four categories of non-transparency distinguished in Section 2.4: redundancy, discontinuity, fusion and form-based form. In Section 6.2.1 I will discuss whether the prediction that there is a between-category implicational hierarchy holds true. Section 6.2.2 will discuss the internal hierarchies in each of the categories.

### *6.2.1 An implicational hierarchy between categories of opacity*

To see whether there is an implicational hierarchy between categories of non-transparency, Table 6.3 is repeated here as Table 6.5, now expanded with an extra row that indicates the feature's category. Languages with the same degree of transparency are combined into single rows. Note that the features that cannot be ranked with respect to other features at all (i.e. the ones given in Table 6.4) are left out of this table. Unfortunately, this includes all of the discontinuity features, so that this category cannot be ranked with respect to the other categories.

In Table 6.5, form-based form features are found from the leftmost to the rightmost row. The two fusion features appear together in the left half of the table, while redundancy features are attested in the left half as well, except for one that appears in the right half. At first sight, it seems as though no between-category implication can be composed, but this changes when we take into account the notion that is crucial in the understanding of these results, viz. syntacticity. Obviously, form-based forms have a high degree of syntacticity by definition, but as has been argued above, obligatory redundancy is highly syntactic as well. If we separate the redundancy category according to syntacticity, that is, in an obligatory subset, containing Tense copying, Clausal agreement and Phrasal agreement, and a non-obligatory subset, consisting of redundant referential marking, Table 6.5 turns out to show a between-category relation after all, given in 3). In this hierarchy, the

category of form-based form has been split up into multiple categories as well, into a syntactic, a morphophonological and a phonological form-based form group.

3) Form-based form (syntactic)
→ Redundancy (obligatory)
→ Fusion
→ Form-based form (morphophonological alternations)
→ Redundancy (optional)
→ Form-based form (phonological alternations)

The only clear counter-example to this hierarchy is the feature 'Syntactic function subject', which is attested in all languages while it is expected to be rare according to 3). The reason for this has been discussed already in Section 6.1.3.

Hierarchy 3) complies with what has been found by Hengeveld (2011b) and Leufkens (2013a), since those studies ranked form-based form above redundancy too. Fusion, in both studies part of a larger category called 'domain disintegration', was located in between, also in agreement with the current findings. However, the previous studies did not distinguish different types of redundancy, nor did they distinguish between syntactic form-based form and (morpho)phonologically

**Table 6.5: Distribution of non-transparent features in terms of category of opacity**

| | Nominal expletive elements | Clausal agreement | Tense copying | Grammatical gender | Morphologically conditioned stem alternation: suppletion | Morphologically conditioned stem alternation: irregular stem formation |
|---|---|---|---|---|---|---|
| | Form-based form | Redun-dancy (obl) | Redun-dancy (obl) | Form-based form | Fusion | Fusion |
| Dutch | + | + | + | + | + | + |
| Egyptian Arabic | - | - | - | + | + | + |
| Georgian | - | - | + | - | + | + |
| Sochiapan Chinantec, Khwarshi, Bininj Gun-Wok, Sandawe, Sheko | - | - | - / na | - | + | + |
| Chukchi, Kayardild | - | - / na | - / na | - | - | + |
| West-Greenlandic | - | - | - | - | - | - |
| Tamil | - | - | - | - | - | + |
| Kolyma Yukaghir, Bantawa | - | - | - / na | - | - | - |
| Kharia, Turkish, Huallaga Quechua, Tidore, Japanese | - | - / na | - / na | - | - | - |
| Samoan | - | - | - / na | - | - | + |
| Fongbe | - | na | - | - | - | - |
| Teiwa | - | - | na | - | - | - |

| | Phrasal agreement | Predominantly head-marking | Morphophonologically conditioned stem alternation | Morphophonologically conditioned affix alternation | Redundant referential marking | Syntactic function subject | Phonologically conditioned affix alternation | Phonologically conditioned stem alternation |
|---|---|---|---|---|---|---|---|---|
| | Redun-dancy (obl) | Form based form | Form based form | Form based form | Redun-dancy | Form based form | Form based form | Form based form |
| Dut | + | + | + | + | + | + | + | + |
| Egy | + | + | + | + | + | + | + | + |
| Geo | + | + | + | + | + | + | + | + |
| Soc, Khw, Bin, San, She | + | + / nd | + | + | + | + | + | + |
| Chu, Kay | + / na | + | + | + | + / na | + | + | + |
| W-Gr | + | nd | + | + | + | + | + | + |
| Tam | - | + | + | + | + | + | + | + |
| Kol, Ban | - | + | + | + | + | + | + | + |
| Kha, Tur, Que, Tid, Jap | - | - / nd | + | + | + / na | + | + | + |
| Sam | - | - | + | + | + | + | + | + |
| Fon | - | - | - | + | na | na | + | + |
| Tei | - | - | - | - | + | + | + | + |

conditioned form-based form. Therefore, it cannot be determined whether the hierarchies are completely parallel.

Before I turn to discussing the category-internal implicational relations, let me point at another interesting finding relating to one of the opacity categories, viz. fusion. Since fusion is considered to be non-transparent in this dissertation, it is not surprising that languages generally considered to be fusional (Dutch, Egyptian Arabic, Sochiapan Chinantec) are among the most non-transparent languages in the sample. However, both other traditionally distinguished morphological types (agglutinative and isolating) are theoretically equally transparent, so that it is not to be expected that an implicational hierarchy of transparency reflects groupings in terms of agglutination or isolation. Still, if we look at the degree of transparency of languages of different morphological types, we observe that all the isolating languages in the sample (Fongbe, Teiwa, Japanese in the nominal domain, Samoan) are located at the transparent end of the table, while agglutinative languages are found in the middle.

There is no theoretical reason for isolating languages to be more transparent than agglutinative languages – the transparency metric does not take account of that. Note that the feature 'predominantly head marking' cannot account for this finding either, since it does not measure whether a language has bound morphemes or not, i.e. whether the language is agglutinative or isolating, but whether there are proportionally more affix-like or more clitic-like bound morphemes. Apparently then, there is an effect of word-status on transparency – one-to-one relations between meanings and words are more transparent than one-to-one relations between meanings and morphemes. Perhaps, this is due to a different mental status of words versus morphemes. Another explanation can be that phonological change, usually the starting point for language change, is less severe in isolating languages, for example because phonological rules crossing word boundaries are less frequent than phonological rules crossing morpheme boundaries. With less phonological change, isolating languages can be expected to develop fewer maturation phenomena, or in other words, they syntacticise less often.

### *6.2.2 Implicational hierarchies within categories of opacity*

The categories of non-transparency cannot only be ordered with respect to each other in a between-category hierarchy, but show internal orderings as well. These are captured in within-category hierarchies that will now be discussed one by one, each time compared to the findings of previous studies into transparency (cf. Section 5.3).

Firstly, Table 6.6 presents the features in the redundancy category, following the ordering of both features and languages of Table 6.3. Table 6.6 shows once more that optional redundancy, i.e. cross-reference, is in fact a very common process – almost all languages show this form of redundancy in the clausal domain. Thus, an implicational relation holds between optional and obligatory redundancy, which is manifested in the clausal domain as an implicational relation between clausal agreement and redundant referential marking. A second implicational relation exists between agreement in the clause and agreement in the phrase. Then, thirdly, clausal agreement implies the presence of tense copying, which is interesting because it means that argument copying implies operator copying (cf. Section 4.1.8 and Hengeveld & Mackenzie 2008: 351). These three category-internal hierarchies are combined in hierarchy 4).

4) Clausal agreement → Tense copying[11] → Phrasal agreement → Redundant referential marking

The current results are not easily compared to results of previous transparency studies, since those studied different sets of features that were also categorised differently. For example, Leufkens (2013a) studied cross-reference and various types of concord under the header of redundancy, but agreement and tense copying as form-based form features. Still, her results match the results of the current study

[11] Note that Hebrew does not display tense copying, but does exhibit phrasal agreement and redundant referential marking copying (cf. Mulder 2013). In fact, as argued in footnote 10, the second part of the hierarchy should say 'Tense copying or grammatical gender', but since this hierarchy includes redundancy phenomena only, grammatical gender is left out.

to a large extent: agreement and tense copying are only attested in the most opaque languages, whereas some type of concord is found in all languages. One difference is that none of the four languages in Leufkens (2013a) exhibits cross-reference, but this may very well be due to the fact that they are contact languages and therefore have a higher degree of transparency. Implicational hierarchy 4) is also largely in agreement with Hengeveld (2011b), who finds the same ordering for agreement, tense copying and apposition, of which cross-reference is a subtype. Hengeveld (2011b) nor Leufkens (2013a) distinguish between clausal and phrasal agreement.

**Table 6.6: Distribution of redundancy features**

| | Clausal agreement | Tense copying | Phrasal agreement | Redundant referential marking |
|---|---|---|---|---|
| Dutch | + | + | + | + |
| Egyptian Arabic | - | - | + | + |
| Georgian | - | + | + | + |
| Sochiapan Chinantec | - | - | + | + |
| Khwarshi | - | - | + | + |
| Bininj Gun-Wok | - | - | + | + |
| Sandawe | - | na | + | + |
| Sheko | - | na | + | + |
| Chukchi | - | na | na | + |
| Kayardild | na | - | + | na |
| West-Greenlandic | - | - | + | + |
| Tamil | - | - | - | + |
| Kolyma Yukaghir | - | na | - | + |
| Bantawa | - | - | - | + |
| Kharia | - | - | - | + |
| Turkish | - | - | - | + |
| Huallaga Quechua | - | na | - | + |
| Tidore | - | na | - | + |
| Japanese | na | - | - | na |
| Samoan | - | - | - | + |
| Fongbe | na | - | - | na |
| Teiwa | - | na | - | + |

Table 6.7 presents the distribution of the fusion features that were studied, in an ordering that reflects the underlying implicational pattern attested in Table 6.3. The table shows that suppletion implies irregular stem formation, which boils down to the conclusion that if grammatical information is expressed by means of stem alternation, the stem will be partly modified rather than replaced completely. This implicational relation is captured in hierarchy 5).

5) Morphologically conditioned stem alternation: suppletion
→ Morphologically conditioned stem alternation: irregular stem formation

Since none of the previous studies into transparency made a distinction between different types of stem alternation, it cannot be determined whether hierarchy 5) complies with previous findings.

**Table 6.7: Distribution of fusion features**

| | Morphologically conditioned stem alternation: suppletion | Morphologically conditioned stem alternation: irregular stem formation |
|---|---|---|
| Dutch | + | + |
| Egyptian Arabic | + | + |
| Georgian | + | + |
| Sochiapan Chinantec | + | + |
| Khwarshi | + | + |
| Bininj Gun-Wok | + | + |
| Sandawe | + | + |
| Sheko | + | + |
| Chukchi | - | + |
| Kayardild | - | + |
| West-Greenlandic | - | - |
| Tamil | - | + |
| Kolyma Yukaghir | - | - |
| Bantawa | - | - |
| Kharia | - | - |
| Turkish | - | - |
| Huallaga Quechua | - | - |
| Tidore | - | - |
| Japanese | - | - |
| Samoan | - | + |
| Fongbe | - | - |
| Teiwa | - | - |

An overview of the distribution of non-transparent features in the form-based form category is given in Table 6.8. These features can be ranked in the implicational hierarchy given as 6).

6) Nominal expletives
→ Grammatical gender
→ Predominantly head-marking
→ Morphophonologically conditioned stem alternation
→ Morphophonologically conditioned affix alternation
→ Phonologically conditioned stem alternation,
Phonologically conditioned affix alternation,
Syntactic function subject

Note that four implicational relations between individual features build up this hierarchy. Firstly, there is an implicational relation between morphosyntactic features (Nominal expletives, Grammatical gender, Predominantly head-marking) and (morpho)phonological alternations. This is a first indication of an effect of the level at which opacity appears, a point that will be pursued further in the following section. A second indication for this level effect is the fact that implicational relations hold between morphophonological alternations in stems and phonological alternations in stems, and between morphophonological alternations in affixes and phonological alternations in affixes. The latter implications are not very telling, since phonologically conditioned alternations in stems and affixes appear in all sample languages, but they are nonetheless important in establishing the aforementioned level effect. Finally, an implicational relation holds between the presence of morphophonologically conditioned stem alternation and the presence of morphophonologically conditioned affix alternation, which shows that stems have a stronger domain integrity than affixes cross-linguistically.

The ordering of form-based form features established in this study is in agreement with earlier findings of Hengeveld (2011b) and Leufkens (2013a). The only difference concerns the ranking of the feature ‘Syntactic function subject’, but

**Table 6.8: Distribution of form-based form features**

| | Nominal expletives | Grammatical gender | Predominantly head-marking | Morphophonologically conditioned stem alternation | Morphophonologically conditioned affix alternation | Phonologically conditioned stem alternation | Phonologically conditioned affix alternation | Syntactic function subject |
|---|---|---|---|---|---|---|---|---|
| Dutch | + | + | + | + | + | + | + | + |
| Egyptian Arabic | - | + | + | + | + | + | + | + |
| Georgian | - | - | + | + | + | + | + | + |
| S. Chinantec | - | - | nd | + | + | + | + | + |
| Khwarshi | - | - | + | + | + | + | + | + |
| Bininj Gun-Wok | - | - | + | + | + | + | + | + |
| Sandawe | - | - | nd | + | + | + | + | + |
| Sheko | - | - | nd | + | + | + | + | + |
| Chukchi | - | - | + | + | + | + | + | + |
| Kayardild | - | - | + | + | + | + | + | + |
| West-Greenlandic | - | - | nd | + | + | + | + | + |
| Tamil | - | - | + | + | + | + | + | + |
| Kolyma Yukaghir | - | - | + | + | + | + | + | + |
| Bantawa | - | - | + | + | + | + | + | + |
| Kharia | - | - | - | + | + | + | + | + |
| Turkish | - | - | nd | + | + | + | + | + |
| Huallaga Quechua | - | - | - | + | + | + | + | + |
| Tidore | - | - | nd | + | + | + | + | + |
| Japanese | - | - | nd | + | + | + | + | + |
| Samoan | - | - | - | + | + | + | + | + |
| Fongbe | - | - | - | - | + | + | + | na |
| Teiwa | - | - | - | - | - | + | + | + |

as explained in Section 6.1.3, the method that these studies used for testing this feature should probably be interpreted differently as there is a strong correlation between syntactic functions and pragmatic ones.

## 6.3 How are non-transparent features at different interfaces distributed cross-linguistically?

Table 6.3 is repeated as Table 6.9, now expanded with a row that gives the interface at which the transparency violation takes place. Table 6.9 clearly shows that two implicational relations hold between interfaces at which transparency is violated. Firstly, there is an implicational relation between the Interpersonal Level and the Representational Level, in the sense that features at interfaces between the pragmatic level and other levels (IL-RL, IL-ML and IL-PL) are more easily violated than features at interfaces between the semantic levels and other levels (RL-ML, RL-PL). In fact, there is one feature that violates transparency at the IL-RL interface, viz. cross-reference: a mismatch between multiple Referential Subacts and one semantic Individual, and one non-transparent feature at the IL-ML interface, viz. the neutralisation of pragmatic functions in a syntactic function subject. Both these features turn out to be highly common, in fact, all languages display these violations of pragmaticity. Violations of transparency at the RL-ML interface are less common, thus giving rise to the implicational relation in 7). This implication can be understood as stating that the obscuring of a one-to-one relation between a pragmatic entity and any other entity is less severe than the obscuring of a transparent relation between a semantic entity and another entity. In yet other words: pragmaticity can be violated more easily than semanticity.

7) Violation at RL interface (RL-ML, RL-PL)

↓

Violation at IL interface (IL-RL, IL-ML, IL-PL)

Secondly, violations of transparency at the interfaces with the Morphosyntactic Level are higher up in the hierarchy than those at the interfaces with the Phonological Level. The fact that phonologically conditioned alternations of both stems and affixes exist in all languages, while morphophonological alternations are less common, shows that the integrity of phonological elements is violated more easily than the integrity of morphological units. Morphologically conditioned alternations, in which an entire stem is affected, are even less common. Thus, non-transparency at an interface between the Phonological Level and another level is less 'severe' than opacity at an interface between the Morphosyntactic Level and any other level. This implicational relation is captured by hierarchy 8).

8) Violation at ML interface (IL-ML, RL-ML)

↓

Violation at PL interface (IL-PL, RL-PL)

Section 2.4 and Section 5.1 both posed the question whether there is a distributional difference between non-transparency at meaning-to form interfaces (i.e. opaque features at the IL-ML, RL-ML, IL-PL and RL-PL interfaces) and non-transparency at the other interfaces (IL-RL and ML-PL). While the study included one violation at IL-RL, no feature was included pertaining to the ML-PL interface. In fact, the only opaque feature at the ML-PL interface is non-parallel alignment (cf. Section 4.2.5), which was excluded from the study because of a lack of data on the topic. Therefore, a comparison between meaning-to-form features and other opaque features is a comparison between cross-reference (Redundant referential marking) and all other features. Of course, this is not very informative, since a single feature is being used to represent an entire category. Therefore, more research into non-transparent features on the IL-RL and the ML-PL interfaces is necessary to make better-substantiated claims on this topic.

**Table 6.9: Distribution of non-transparent features in terms of interface**

| | Nominal expletive elements | Clausal agreement | Grammatical gender | Tense copying | Morphologically conditioned stem alternation: Suppletion | Phrasal agreement | Morphologically conditioned stem alternation: Irregular Stem Formation |
|---|---|---|---|---|---|---|---|
| | RL ML | RL ML | RL ML | RL ML | RL ML | RL ML | RL ML |
| Dutch | + | + | + | + | + | + | + |
| Egyptian Arabic | - | - | + | - | + | + | + |
| Georgian | - | - | - | + | + | + | + |
| Sochiapan Chinantec, Khwarshi, Bininj Gun-Wok, Sandawe, Sheko | - | - | - | - / na | + | + | + |
| Chukchi, Kayardild | - | - / na | - | - / na | - | + / na | + |
| West-Greenlandic | - | - | - | - | - | + | - |
| Tamil | - | - | - | - | - | - | + |
| Kolyma Yukaghir, Bantawa | - | - | - | - / na | - | - | - |
| Kharia, Turkish, Huallaga Quechua, Tidore, Japanese | - | - / na | - | - / na | - | - | - |
| Samoan | - | - | - | - | - | - | - |
| Fongbe | - | na | - | - | - | - | - |
| Teiwa | - | - | - | na | - | - | - |

| | Predominantly head-marking | Morphophonologically conditioned stem alternation | Morphophonologically determined affix alternation | Redundant referential marking | Syntactic function subject | Phonologically conditioned affix alternation | Phonologically conditioned stem alternation |
|---|---|---|---|---|---|---|---|
| | RL ML | RL ML | RL ML | IL RL | IL ML | RL PL | RL PL |
| Dut | + | + | + | + | + | + | + |
| Egy | + | + | + | + | + | + | + |
| Geo | + | + | + | + | + | + | + |
| Soc, Khw, Bin, San, She | + / nd | + | + | + | + | + | + |
| Chu, Kay | + | + | + | + / na | + | + | + |
| W-Gr | nd | + | + | + | + | + | + |
| Tam | + | + | + | + | + | + | + |
| Kol, Ban | + | + | + | + | + | + | + |
| Kha, Tur, Que, Tid, Jap | - / nd | + | + | + / na | + | + | + |
| Sam | - | + | + | + | + | + | + |
| Fon | - | - | + | na | na | + | + |
| Tei | - | - | - | + | + | + | + |

# Chapter 7
# Conclusions

In this chapter, I will review the outcomes of the study performed and discuss the conclusions that can be drawn from it. Furthermore, I will examine the implications for the theoretical debate on the complexity of languages, language change and language learnability. Finally, I will give suggestions for the direction that research into these topics might take.

The study that was reported on in this dissertation has provided answers to the research questions given here:

1) How are non-transparent features distributed cross-linguistically?
2) How are redundancy, fusion, domain disintegration and form-based form features distributed cross-linguistically?
3) How are non-transparent features at different interfaces distributed cross-linguistically?

Answers were formulated in terms of implicational hierarchies that are presented in combination in Figure 7.1 and Figure 7.2. These figures should be read as follows. Each individual non-transparent feature forms a small rectangle. Each feature rectangle is captured in larger boxes, representing the category of opacity in which the feature belongs in Figure 7.1, and the interface at which the feature plays a role in Figure 7.2. Category and interfaces boxes are printed bold or with a dotted line. Each feature implies the presence of the features below it on the vertical axis, regardless of the horizontal axis. For example, the feature 'Nominal expletives' implies 'Grammatical gender', but also 'Tense copying' and 'Phonologically conditioned stem affix alternations'. Boxes that have an equal height are not ranked with respect to each other. For example, 'Nominal expletives' and 'Clausal agreement' cannot be ranked with respect to each other.

Implications between categories of opacity hold as marked by arrows, so that, for instance, the presence of any feature from the category ‘Obligatory redundancy’ implies the presence of a feature from the category ‘Fusion’. In the same vein, violations of transparency at interfaces of the Representational Level imply the presence of violations of transparency at interfaces of the Interpersonal Level.

**Figure 7.1: Implicational hierarchy of transparency in terms of category**

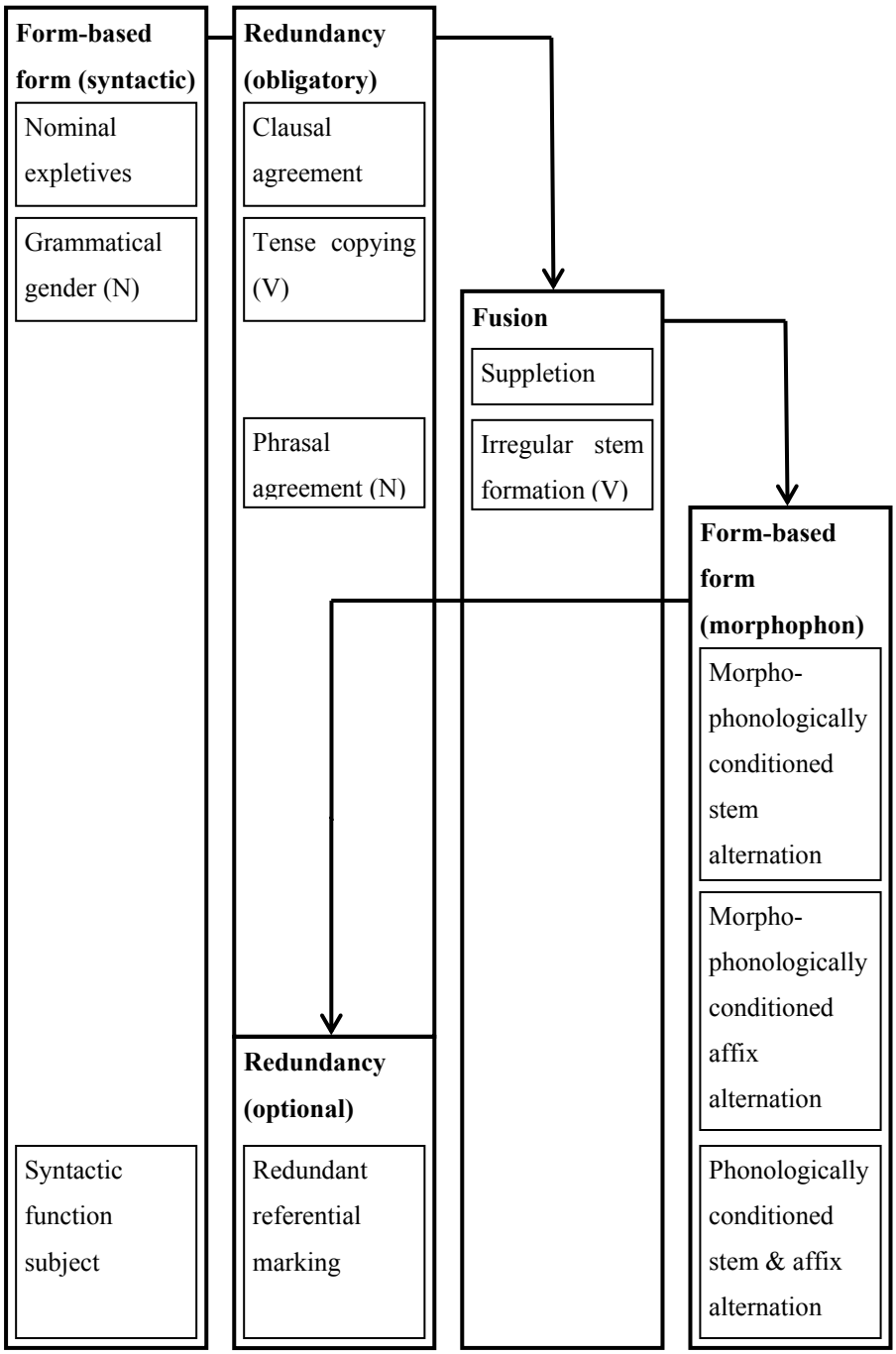
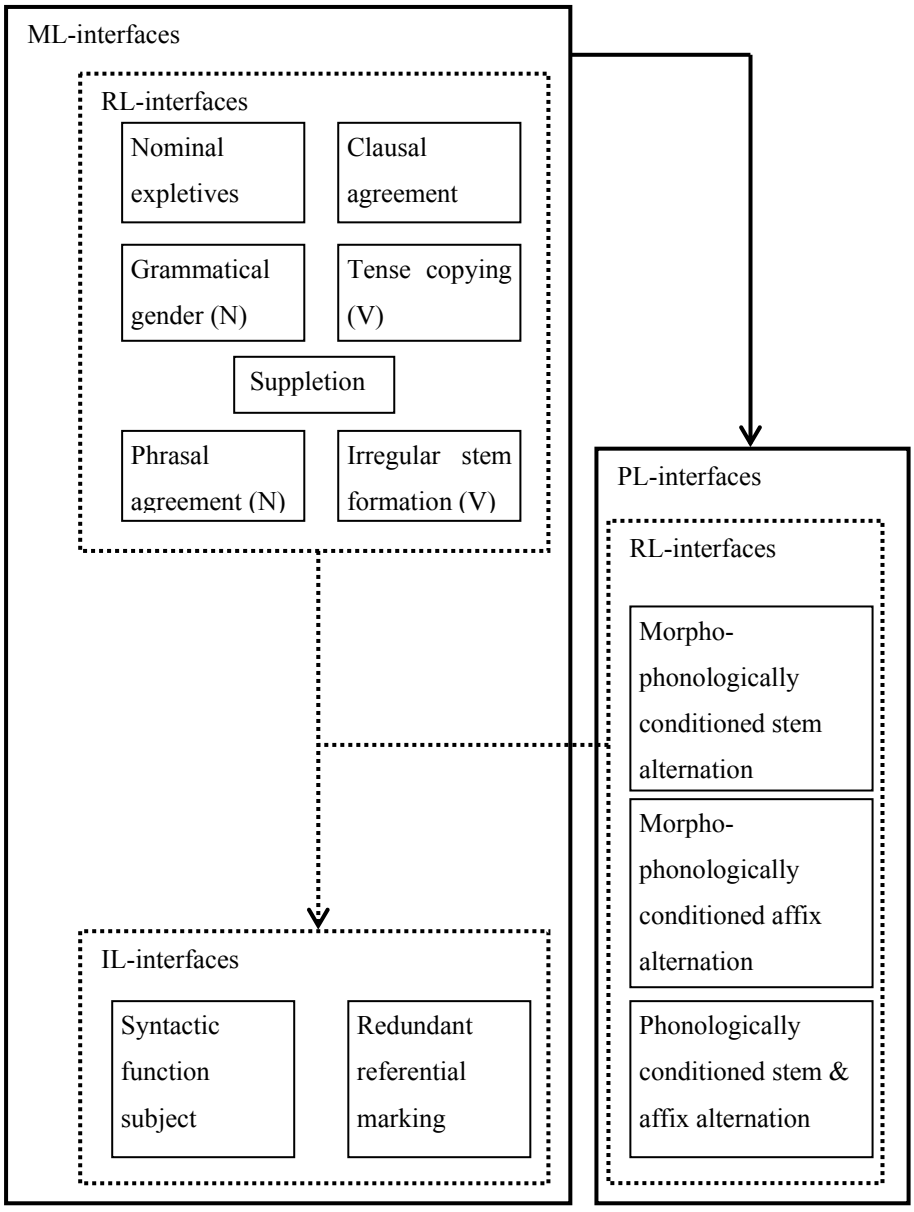


**Figure 7.2: Implicational hierarchy of transparency in terms of interface**



The two-dimensional hierarchies in Figure 7.1 and Figure 7.2 visualise how the hierarchies can be explained by syntacticity, both as an effect of opacity category and as an effect of level. Firstly, the ordering of non-transparent features is an ordering of the most syntacticised categories, viz. form-based form and obligatory redundancy, to the most pragmatic/semantic category, i.e. optional redundancy. Fusion is located in between. Secondly, the ordering of non-transparent features reflects an ordering of interfaces, since non-transparency occurs more easily at pragmatic interfaces than at semantic interfaces, and more easily at the Phonological Level than at the Morphosyntactic Level. Thus, a non-transparent language is a language in which purely syntactic rules and elements have taken over from pragmatically and semantically motivated rules and elements. The presence of form-based form and obligatory copying of (properties of) elements is the crucial characteristic of such languages.

A small number of non-transparent features turned out not to participate in this hierarchy. For two of these features, viz. plural concord in noun phrases containing a numeral, and cumulation of TAME and case morphemes, this is argued to be the result of too small a scope of the metric. Presumably, concord and cumulation are near-universal features, and languages differ only in the domain in which concord or cumulation manifests itself. Testing plural concord in noun phrases containing a numeral, and cumulation of TAME and case only, isolated from other types of concord and cumulation, resulted in a scattered distribution.

Moreover, features in the discontinuity category did not show a distributional pattern. The diffused occurrence of discontinuous morphosyntactic elements can be explained by the fact that such elements are the outcome of an interplay between several word order rules, which are typically affected by multiple factors, such as pragmatic considerations and processing matters. These factors may obscure the working of transparency in forming continuous elements that are positioned consistently. Thus, a distributional pattern based on transparency is overruled by other ordering principles. The fact that languages that are on the basis of their degree of non-transparency expected to show discontinuous morphology do not always do so is explained by considering the origins of circumfixes and infixes.

In fact, such affixes only appear in languages as a result of a complex interplay between processes such as reduplication, metathesis, and morphophonological change. If such processes do not exist in a language, the preconditions for having discontinuous morphology are simply not met, thus obscuring the degree of non-transparency of that language.

An interesting finding in this study is that isolating languages turn out to be generally more transparent than agglutinative languages, even though the features studied do not measure morphological type in a direct sense. The correlation between the isolating type and a high degree of transparency shows that a non-compound, non-inflected word is more transparent than a single morphosyntactic word consisting of multiple stems or inflected stems, even when those stems and inflections transparently relate to other levels of organisation. If one accepts the hypothesis that transparency is easier to acquire, then these findings support the idea that inflectional morphemes are fundamentally more difficult to acquire than isolated morphemes, as argued by Dahl (2004), among others.

Furthermore, the findings in this dissertation mostly corroborate the findings of earlier studies into transparency, viz. Hengeveld (2011b) and Leufkens (2013a). Moreover, the findings are in line with findings of many acquisitional studies that demonstrate an increased relative complexity of languages features that are highly syntactic. For example, grammatical gender has often been said to be exceptionally difficult to acquire for both L1 and L2 learners. In agreement with this, it is located at the top of the opacity hierarchy established in this dissertation. Note furthermore that the features at the top of the hierarchy are all on Dahl's (2004: 114-115) list of maturation phenomena, showing that they take time to develop not only in a speaker, but diachronically too.

Previous studies of transparency in language used samples of at most 11 languages. The current dissertation studies a sample of 22 languages and therefore gives more stability and scientific validity to the results. Of course, future research can straightforwardly add to this by including more and more languages. Ideally, such research would include more non-transparent features, such as non-parallel alignment and various other types of concord that this study could not take into

account because of a lack of data. Furthermore, research into transparency would benefit from the development of a more fine-grained metric that does not rely on measuring features in a binary fashion, but on gradual metrics for each feature.

The findings of this dissertation could also be relevant for further studies in other subfields of linguistics. Especially interesting would be research into learnability, testing the hypothesis that non-transparency is fundamentally more difficult for L1 and L2 learners than transparency. Obviously, studies of language contact could further corroborate the claim by Leufkens (2013a) that particular types of language contact correlate with an increase of transparency in a language, in line with argumentation and evidence for instance by Trudgill (2011), showing that short-term adult language contact leads to language simplification.

# References

**In-text references**


Abdi, Hervé. 2010. Guttman Scaling. In Neil Salkind (ed.), *Encyclopedia of Research Design*. Thousand Oaks, CA: Sage. Available online at https://www.utdallas.edu/~herve/abdi-GuttmanScaling2010-pretty.pdf (retrieved on July 17, 2014).

Aboh, Enoch. 2004. *The morphosyntax of complement-head sequences. Clause structure and word order patterns in Kwa*. New York : Oxford University Press.

Aboh, Enoch & Norval Smith. 2009. Simplicity, simplification, complexity and complexification. Where have the interfaces gone? In Enoch Aboh & Norval Smith (eds.), *Complex processes in new languages*, 1-25. Amsterdam: John Benjamins.

Aikhenvald, Alexandra Y. 2000. *Classifiers: A typology of noun categorization devices*. Oxford: Oxford University Press.

Aksu-Koç, Ayhan & Dan Slobin. 1985. The acquisition of Turkish. In Dan Slobin (ed.), *The crosslinguistic study of language acquisition*, 839-878. Hillsdale: Lawrence Erlbaum Associates.

Anderson, Stephen R. 2005. *Aspects of the theory of clitics*. Oxford: Oxford University Press.

Audring, Jenny. 2009a. *Reinventing pronoun gender* (PhD dissertation, Free University Amsterdam). Utrecht: LOT.

Audring, Jenny. 2009b. Gender assignment and gender agreement. *Morphology* 18, 93-116.

Baković, Eric. 2011. Opacity and ordering. In John Goldsmith, Jason Riggle & Alan Yu (eds.), *The handbook of phonological theory* (2nd edition), 40-67. Malden, MA: Wiley-Blackwell.

Bates, Elizabeth & Brian MacWhinney (eds.). 1989. *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press.

Bauer, Laurie. 2003. *Introducing linguistic morphology* (2nd edition). Edinburgh: Edinburgh University Press.

Bellugi, Ursula & Edward S. Klima. 1975. Two faces of sign: Iconic and abstract. *Annals of the New York Academy of Sciences* 280, 514-538.

Bennis, Hans 1986. *Gaps and Dummies*. Dordrecht: Foris Publications.

Berg, Thomas. 1998. The resolution of number conflicts in English and German agreement patterns. *Linguistics* 36 (1), 41-70.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79, 708-736.

Bickel, Balthasar. 2011. Grammatical relations typology. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 399-444. Oxford: Oxford University Press.

Bickel, Balthasar & Johanna Nichols. 2013. Exponence of Selected Inflectional Formatives. In Matthew S. Dryer & Martin Haspelmath (eds.), The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/21 (accessed on July 27, 2014).

Bisang, Walter. 2009. On the evolution of complexity – Sometimes less is more in East and mainland Southeast Asia. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as a variable concept*, 34-49. Oxford: Oxford University Press.

Blom, Elma, Daniela Polišenská & Fred Weerman. 2008. Articles, adjectives and age of onset: The acquisition of Dutch grammatical gender. *Second Language Research* 24, 297-331.

Braun, Maria & Ingo Plag. 2003. How transparent is creole morphology? A study of early Sranan word-formation. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 2002*, 81-104. Dordrecht: Kluwer.

Bybee, Joan L. 1985. *Morphology. A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Campbell, Lyle. 2013. *Historical linguistics. An introduction* (3rd edition). Edinburgh: Edinburgh University press.

Carstairs-McCarthy, Andrew. 1987. *Allomorphy in inflexion*. London: Croom Helm.

Chomsky, Noam. 1980. *Rules and representations*. New York: Columbia University Press.

Chomsky, Noam. 1988. *Language and problems of knowledge: The Managua lectures*. Cambridge, MA: MIT Press.

Comrie, Bernard. 1986. Tense in indirect speech. *Folia Linguistica* 20, 265-296.

Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville. 2006. *Agreement*. Cambridge: Cambridge University Press.

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

Croft, William. 2003b. Typology. In Mark Aronoff and Janie Rees-Miller (eds.), *The Blackwell Handbook of Linguistics* (2nd edition). Oxford: Basil Blackwell, 337-369.

Cysouw, Michael. 2003. Against implicational universals. *Linguistic Typology* 7, 89-101.

Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

De Swart, Henriëtte & Ivan A. Sag. 2002. Negation and negative concord in Romance. *Linguistics and Philosophy* 25 (4), 373-417.

Dereau, Leon. 1955. *Cours de Kikongo*. Namur: Wesmael-Charlier.

Dik, Simon. 1978. *Functional Grammar*. Amsterdam: North-Holland.

Dressler, Wolfgang U. 1985. On the Predictiveness of Natural Morphology. *Journal of Linguistics* 21 (2), 321-337.

Dressler, Wolfgang U., Willi Mayerthaler, Oswald Panagl and Wolfgang U. Wurzel. 1987. *Leitmotifs in Natural Morphology* (Studies in language companion series 10). Amsterdam: John Benjamins.

Dryer, Matthew S.. 2013. Negative Morphemes. In Matthew S. Dryer & Martin Haspelmath (eds.), The World Atlas of Language Structures Online.

Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/112 (accessed on July 23, 2014).

Dunn, Michael John. 1999. *A grammar of Chukchi* (unpublished doctoral dissertation). Canberra: Australian National University.

Durie, Mark. 1985. *A grammar of Acehnese: On the basis of a dialect of North Aceh* (Verhandelingen van het Koninklijk Insitituut voor Taal-, Land- en Volkenkunde 112). Dordrecht/Cinnaminson, NJ: Foris.

Foley, William A. & Robert D. Van Valin Jr. 1984. *Functional syntax and Universal Grammar*. Cambridge: Cambridge University Press.

Foris, David Paul. 2000. *A grammar of Sochiapan Chinantec* (Studies in Chinantec Languages 6). Dallas: SIL International and The University of Texas at Arlington.

García Velasco, Daniel. 2013. *Where is word meaning in Functional Discourse Grammar?* Paper presented at the International Workshop on The Lexicon in Functional Discourse Grammar, Vienna, September 5-6.

Gary, Judith Olmsted and Saad Gamal-Eldin. 1981. *Cairene Egyptian Colloquial Arabic*.
London: Croom Helm.

Gil, David. 2008. How complex are isolating languages? In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity. Typology, contact, change*, 109-132. Amsterdam: John Benjamins.

Gil, David. 2013. Adjectives without Nouns. In Matthew S. Dryer & Martin Haspelmath (eds.), The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/61 (accessed on July 23, 2014).

Gillis, Steven. 2000. Fonologische ontwikkeling. In Gillis, Steven & Annemarie Scharlaekens (eds.), *Kindertaalverwerving. Een handboek voor het Nederlands*, 131-184. Groningen: Martinus Nijhoff.

Givón, Talmy. 1985. Iconicity, isomorphism and non-arbitrary coding in syntax. In John Haiman (ed.), *Iconicity in syntax*, 187-219. Amsterdam: John Benjamins.

Grández Ávila, Magaly. Language transparency in Functional Discourse Grammar: The case of Quechua. *Linguistics in Amsterdam* 4 (2), 22-56.

Greenberg, Joseph H. 1957. *Essays in Linguistics*. Chicago: University of Chicago Press.

Greenberg, Joseph. 1966 [1963]. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg (ed.), *Universals of language* (2nd edition), 73-113. Cambridge: MIT Press.

Greenberg, Joseph. 1978. Diachrony, synchrony and language universals. In Joseph Greenberg (ed.), *Universals of human language. Vol. I: Method & theory*, 61-91. Stanford: Stanford University Press.

Haiman, John. 1980. The iconicity of grammar: Isomorphism and motivation. *Language* 56 (3), 515-540.

Haiman, John. 1983. Iconic and economic motivation. *Language* 59 (4), 781-819.

Haiman, John (ed.). 1985. *Iconicity in syntax*. Amsterdam: John Benjamins.

Haiman, John. 1985. *Natural syntax. Iconicity and erosion*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Clarendon Press.

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19, 1-33.

Haspelmath, Martin. 2009. Terminology of case. In Andrej Malchukov & Andrew Spencer (eds.), *Handbook of case*, 505–517. Oxford: Oxford University Press.

Hawkins, John A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21 (2), 223-261.

Heine, Bernd & Kuteva, Tania. 2005. *Language contact and grammatical change*. Cambridge: Cambridge University Press.

Hengeveld, Kees. 2004. Morphological types in Functional Discourse Grammar. In Kees Hengeveld (ed.), *Morphology in Functional Discourse Grammar. Working Papers in Functional Grammar* 79, 1-10.

Hengeveld, Kees. 2007. Parts-of-speech systems and morphological types. *ACLC Working Papers* 2 (1), 31-48.

Hengeveld, Kees (ed.). 2011. Transparency in functional discourse grammar [Special issue]. *Linguistics in Amsterdam* 4 (2).

Hengeveld, Kees. 2011a. Introduction: Transparency in Functional Discourse Grammar. *Linguistics in Amsterdam* 4 (2), 1-22.

Hengeveld, Kees. 2011b. Epilogue: Degrees of transparency. *Linguistics in Amsterdam* 4 (2), 110-114.

Hengeveld, Kees. 2013. A new approach to constituent order typology. In Lachlan J. Mackenzie & Hella G. Olbertz (eds.), *Casebook in Functional Discourse Grammar*, 15-38. Amsterdam: John Benjamins.

Hengeveld, Kees & Lachlan J. Mackenzie. 2008. *Functional Discourse Grammar. A Typologically-Based Theory of Language Structure*. Oxford: Oxford University Press.

Homomorphism lemma. 2014. *Encyclopædia Britannica Online*. Available online at http://www.britannica.com/EBchecked/topic/270579/homomorphism (accessed on July 30, 2014).

Hyman, Larry M. 2008. Enlarging the scope of phonologization. *University of California Berkeley Phonology Lab Annual Report*, 382–409. Available online at http://linguistics.berkeley.edu/phonlab/annual_report/documents/2008/Hyman_Phonologization_PLAR.pdf (accessed on July 18, 2014).

Itkonen, Esa. 2004. Typological explanation and iconicity. *Logos and Language* V (1), 21-33.

Kay, Paul & Gillian Sankoff. 1974. A language-universals approach to pidgins and creoles. In David DeCamp & Ian Hancock (eds.), *Pidgins and creoles: Current trends and prospects*, 61-72. Washington: Georgetown University Press.

Keizer, Evelien. 2005. The discourse function of close appositions. *Neophilologus* 89, 447-467.

Khalilova, Zaira, 2009. *A grammar of Khwarshi* (PhD dissertation, Leiden University). Utrecht: LOT.

Kihm, Alain. 2000. Are creole languages "perfect" languages? In John McWhorter (ed.), *Language change and language contact*, 163-199. Amsterdam: John Benjamins.

Kiparsky, Paul. 1973. Abstractness, opacity and global rules. In Osamu Fujimura (ed.), *Three Dimensions of Linguistic Theory*, 57-86. Tokyo: Institute for Advanced Studies of Language.

Kiparsky, Paul & Cleo Condoravdi. 2006. Tracking Jespersen's cycle (manuscript). Patras: University of Patras. Available online at http://web.stanford.edu/~kiparsky/Papers/lesvosnegation.pdf (retrieved on November 28, 2013).

Kornfilt, Jaklin. 1997. *Turkish*. New York: Routledge.

Kouwenberg, Sylvia & Darlene LaCharité. 2011. The typology of Caribbean Creole reduplication. *Journal of Pidgin and Creole Languages* 26 (1), 194-218.

Krueger, John R. 1962. *Yakut manual* (Indiana University Publications, Uralic and Altaic Series 21). The Hague: Mouton.

Kusters, Wouter. 2003. *Linguistic complexity: The influence of social change on verbal inflection* (PhD dissertation, Leiden University). Utrecht: LOT.

Langacker, Ronald W. 1977. Syntactic reanalysis. In Charles Li (ed.), *Mechanisms of syntactic change*, 56-139. Austin: University of Texas Press.

Langacker, Ronald W. 2008. *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Lass, Roger. 1997. *Historical linguistics and language change*. Cambridge: Cambridge University Press.

Leeuw, Frank van der. 1997. *Clitics. Prosodic studies* (PhD dissertation, University of Amsterdam). Amsterdam: Holland Academic Graphics.

Lefebvre, Claire. 2001. On the semantic opacity of creole languages. *Journal of Pidgin and Creole Languages* 16, 321-353.

Leufkens, Sterre. 2009. *Words vs. Clitics* (unpublished Research MA paper). Amsterdam: University of Amsterdam.

Leufkens, Sterre. 2010. *The transparency of creoles* (unpublished Research MA thesis). Amsterdam: University of Amsterdam.

Leufkens, Sterre. 2011. Kharia: A transparent language. *Linguistics in Amsterdam* 4 (2), 75-95.

Leufkens, Sterre. 2013a. The transparency of creoles. *Journal of Pidgin and Creole Languages* 28 (2), 323-362.

Leufkens, Sterre. 2013b. Time reference in English indirect speech. In Lachlan J. Mackenzie, & Hella Olbertz G. (eds.), *Casebook in Functional Discourse Grammar*, 189-212. Amsterdam: John Benjamins.

Lewis, G.L. 1978. *Turkish grammar*. Oxford: Clarendon Press.

Lewis, Paul M. (ed.). 2009. *Ethnologue: Languages of the world* (16th edition). Dallas: SIL International. Available online at http://archive.ethnologue.com/16/ (accessed repeatedly in 2010-2014).

Li, Charles & Sandra Thompson. 1976. Subject and topic: A new typology of language. In Charles Li (ed.), *Subject and topic*, 457-490. New York: Academic Press.

Lightfoot, David W. 1979. *Principles of diachronic syntax*. Cambridge: Cambridge University Press.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5 (1), 1-10.

MacWhinney, Brian. 2005. A unified model of language acquisition. In J. F. Kroll & A. M. B. de Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, 49-67. New York: Oxford University Press.

McWhorter, John. 1998. Identifying the creole prototype: Vindicating a typological class. *Language* 74, 788-818.

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5, 125-166.

Miestamo, Matti. 2006. On the feasibility of complexity metrics. In Krista Kerge & Maria-Maren Sepper (eds.), *Conference proceedings of The Annual Finnish and Estonian Conference of Linguistics, May 6-7, 2004*, 11-26. Tallin: TLÜ.

Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan reference grammar*. Oslo: Scandinavian University Press.

Mulder, Mijke. 2013. Transparency in Modern Hebrew: A Functional Discourse Grammar analysis. *Linguistics in Amsterdam* 6 (1), 1-27.

Naro, Anthony J. 1978. A study on the origins of pidginization. *Language* 54 (2), 314-347.

Newmeyer, Frederick J. 1998. *Language form and language function*. Cambridge: MIT Press.

Nordhoff, Sebastian. 2009. *A grammar of upcountry Sri Lanka Malay* (PhD dissertation, University of Amsterdam). Utrecht: LOT.

Nordhoff, Sebastian. 2011. Transparency in Sri Lanka Malay. *Linguistics in Amsterdam* 4 (2), 96-110.

O'Neill, Gareth. 2012. *Initial consonant mutation in Irish Gaelic: a Functional Discourse Grammar analysis* (unpublished MA thesis). Amsterdam: University of Amsterdam.

Omar, Margaret. 1973. *The acquisition of Egyptian Arabic as a native language*. The Hague: Mouton.

Parkvall, Mikael. 2008. The simplicity of creoles in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity. Typology, contact, change*, 265-285. Amsterdam: John Benjamins.

Payne, Thomas. 1997. *Describing morphosyntax*. Cambridge: Cambridge University Press.

Peterson, John. 2011. *A grammar of Kharia. A South Munda language*. Leiden: Brill.

Riddle, Elisabeth. 2008. Complexity in isolating languages: Lexical elaboration versus grammatical economy. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity. Typology, contact, change*, 133-152. Amsterdam: John Benjamins.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & P. Kahrel. 1993. A method of language sampling. *Studies in Language* 17 (1), 169-203.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2, 263-314.

Rijkhoff, J. 2002. *The noun phrase*. Oxford: Oxford University Press

Ruhlen, Merritt. 1991. *A guide to the world's languages. Vol. 1: Classification* (2nd edition). Stanford, CA: Stanford University Press.

Sampson, Geoffrey, David Gil & Peter Trudgill (eds.). 2009. *Language complexity as a variable concept*. Oxford: Oxford University Press.

Seinhorst, Klaas. 2014. Phonetics in Functional Discourse Grammar. *Web papers in Functional Discourse Grammar* 87. Available online at http://home.hum.uva.nl/fdg/working_papers/WP-FDG-87.pdf (retrieved on February 21, 2014).

Seuren, Pieter & Herman Wekker. 1986. Semantic transparency as a factor in creole genesis. In Pieter Muysken (ed.), *Substrata versus universals in creole genesis*, 57-70. Amsterdam: John Benjamins.

Shosted, Ryan K. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10, 1-40.

Slobin, Dan I. 1977. Language change in childhood and history. In John Macnamara (ed.), *Language learning and language thought*, 185-214. New York: Academic Press.

Slobin, Dan I. 1980. The repeated path between transparency and opacity in language. In Ursula Bellugi & M. Studdert-Kennedy (eds.), *Signed and Spoken Language: Biological Constraints on Linguistic Form*, 229-243. Weinheim: Verlag Chemie GmbH.

Slobin, Dan I. (ed.). 1985. *The crosslinguistic study of language acquisition*. Hillsdale: Lawrence Erlbaum Associates.

Slobin, Dan I. 2004. From ontogenesis to phylogenesis: What can child language tell us about language evolution? In J. Langer, S.T. Parker & C. Milbrath (eds.), *Biology and Knowledge Revisited: From Neurogenesis to Psychogenesis*, 255-286. Mahwah, NJ: Lawrence Erlbaum Associates.

Snow, Catherine, Norval Smith & Marian Hoefnagel-Höhle. 1980. The acquisition of some Dutch morphological rules. *Journal of Child Language* 7, 539-553.

Steele, Susan. 1978. Word order variation: a typological study. In J. H. Greenberg, C. A. Ferguson and E. A. Moravcsik (eds.), *Universals of Human Language. Vol. IV: Syntax*, 585-624. Stanford: Stanford University Press.

Thomason, Sarah. 2001. *Language contact: An introduction*. Washington DC: Georgetown
University Press.

Thomason, Sarah & Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Travis, Lisa. 1984. *Parameters and effects of word order variation* (doctoral dissertation). Cambridge, MA: MIT.

Trudgill, Peter. 2011. *Sociolinguistic typology. Social determinants of linguistic complexity*. Oxford: Oxford University Press.

Valin, Robert D. Jr. van. 2003. Functional linguistics. In Mark Aronoff and Janie Rees-Miller (eds.), *The Blackwell Handbook of Linguistics* (2nd edition) , 319-336. Oxford: Basil Blackwell.

Velde, Freek van de. 2012. PP extraction and extraposition in Functional Discourse Grammar. *Language Sciences* 34, 433-454.

Wasow, Thomas. 2003. Generative Grammar. In Mark Aronoff and Janie Rees-Miller (eds.), *The Blackwell Handbook of Linguistics* (2nd edition) , 295-318. Oxford: Basil Blackwell.

Wurzel, Wolfgang U. 2001. Creoles, complexity, and linguistic change. *Linguistic Typology* 5 (2/3), 377-387.

Zeijlstra, Hedde. 2007a. Negation in natural language: On the form and meaning of negative elements. *Language and Linguistics Compass* 1 (5), 498-518.

Zeijlstra, Hedde. 2007b. Modal concord. In T. Friedman & M. Gibson (eds.), *SALT XVII*, 317-332. Ithaca, NY: Cornell University.

Zwicky, Arnold M. & Geoffrey K. Pullum. 1983. Cliticization vs. inflection: English n't.
*Language* 59 (3), 502-513.

**Descriptions of sample languages**

BANTAWA

Bickel, Balthasar & Johanna Nichols. 2007. Inflectional morphology. In Timothy Shopen (ed.), *Language typology and syntactic description, Vol. 3: Grammatical categories and the lexicon* (2nd edition), 169–240. Cambridge: Cambridge University Press.

Doornenbal, Marius. 2009. *A grammar of Bantawa* (PhD dissertation, Leiden University). Utrecht: LOT.

BININJ GUN-WOK

Evans, Nicholas. 2003. *Bininj Gun-Wok. A pan-dialectal grammar of Mayali, Kunwinjku and Kune*. Canberra: Pacific Linguistics.

CHUKCHI

Dunn, Michael John. 1999. *A grammar of Chukchi* (unpublished doctoral dissertation). Canberra: Australian National University.

DUTCH

Audring, Jenny. 2009a. Reinventing pronoun gender (PhD dissertation, Free University Amsterdam). Utrecht: LOT.

Snow, Catherine, Norval Smith, and Marian Hoefnagel-Höhle. 1980. The acquisition of some Dutch morphological rules. Journal of Child language 7, 539-553.

EGYPTIAN ARABIC

Gary, Judith Olmsted and Saad Gamal-Eldin. 1981. *Cairene Egyptian Colloquial Arabic*.
London: Croom Helm.

FONGBE

Bobyleva, Ekaterina. 2013. *The development of the nominal domain in creole languages*. A comparative-typological approach (PhD dissertation, University of Amsterdam). Utrecht: LOT.

Lefebvre, Claire & Anne-Marie Brousseau. 2002. *A grammar of Fongbe* (Mouton grammar library 25). Berlin: Mouton de Gruyter.

GEORGIAN

Harris, Alice C. 1981. *Georgian syntax. A study in relational grammar*. Cambridge: Cambridge University Press.

Hewitt, Brian George. 1987. *The typology of subordination in Georgian and Abkhaz* (Empirical approaches to language typology 5). Berlin: Mouton de Gruyter.

Hewitt, Brian George. 1995. *Georgian. A structural reference grammar*. Amsterdam: John Benjamins Publishing.

Valin, Robert D. van. 1990. Semantic parameters of split intransitivity. *Language* 66 (2), 221-260.

Vamling, Karina. 1989. *Complementation in Georgian*. Lund: Lund University Press.

Wier, Thomas. 2011. *Georgian morphosyntax and feature hierarchies in natural language* (PhD dissertation). Chicago: University of Chicago.

HUALLAGA QUECHUA

Grández Ávila, Magaly. 2011. Language transparency in Functional Discourse Grammar: The case of Quechua. *Linguistics in Amsterdam* 4 (2), 22-56.

Weber, David J. 1989. *A grammar of Huallaga (Huánuco) Quechua*. Berkeley and Los Angeles: University of California Press.

Weber, David J. 1998. *Rimaycuna. Quechua de Huánuco. Diccionario del Quechua del Huallaga* (Serie Linguística Peruana N. 48). Lima: Instituto Linguistico de Verano.

JAPANESE

Hinds, John. 1986. *Japanese*. London: Croom Helm.

Iwasaki, Shoichi. 2002. *Japanese*. Amsterdam: John Benjamins.

Iwasaki, Shoichi. 2013. *Japanese* (revised edition). Amsterdam: John Benjamins.

KAYARDILD

Evans, Nicholas. 1995. *A grammar of Kayardild* (Mouton grammar library 15). Berlin: Mouton de Gruyter.

KHARIA

Peterson, John. 2011. *A grammar of Kharia. A South Munda language*. Leiden: Brill.

KHWARSHI

Khalilova, Zaira, 2009. *A grammar of Khwarshi* (PhD dissertation, Leiden University). Utrecht: LOT.

Forker, Diana. 2012. The bi-absolutive construction in Nakh-Daghestanian. *Folia Linguistica* 46, 75-108.

KOLYMA YUKAGHIR

Maslova, Elena. 2003. *A grammar of Kolyma Yukaghir* (Mouton Grammar Library 27). Berlin: Mouton de Gruyter.

SAMOAN

Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan reference grammar*. Oslo: Scandinavian University Press.

SANDAWE

Eaton, Helen. 2010. *A Sandawe grammar* (SIL e-Books 20). Available online at http://www-01.sil.org/silepubs/Pubs/52718/52718_EatonH_Sandawe_Grammar.pdf (accessed repeatedly in 2010-2014).

Hunziker, Daniel, Elisabeth Hunziker & Helen Eaton. 2008. A description of the phonology of the Sandawe language. SIL Electronic working papers. Available online at http://www.sil.org/silewp/2008/silewp2008-004.pdf (retrieved on November 28, 2013)

Steeman, Sander. 2012. *A grammar of Sandawe* (PhD dissertation, Leiden University). Utrecht: LOT.

SHEKO

Aklilu, Yilma. 1988. *The phonology and grammar of Sheko* (MA thesis). Addis Ababa: Addis Ababa University, Institute of Ethiopian Studies.

Hellenthal, Anneke Christine. 2010. *A grammar of Sheko* (PhD dissertation, Leiden University). Utrecht: LOT.

SOCHIAPAN CHINANTEC

Foris, David Paul. 1973. Chinantec syllable structure. International Journal of American *Linguistics* 39 (4), 232-235.

Foris, David Paul. 2000. *A grammar of Sochiapan Chinantec* (Studies in Chinantec Languages 6). Dallas: SIL International and The University of Texas at Arlington.

TAMIL

Andronov, Mikhail. 2004. *A reference grammar of the Tamil language*. München: Lincom.

Asher, Ron. 1982. *Tamil* (Lingua Descriptive Studies 7). Amsterdam: North-Holland.

Lehmann, Thomas. 1989. *A grammar of modern Tamil*. Pondicherry: Pondicherry Institute of Linguistics and Culture.

Mallinson, Graham. 1986. Languages with and without extraposition. *Folia Linguistica*, 20 (2), 147-163.

Schiffman, Harold F. 1999. *A reference grammar of spoken Tamil*. Cambridge: Cambridge University Press.

Steever, Stanford. 2005. *The Tamil auxiliary verb system*. New York: Routledge.

TEIWA

Klamer, Marian A. F. 2003. *A grammar of Teiwa* (Mouton grammar library 49). Berlin: Mouton de Gruyter.

Klamer, Marian A. F. To appear. Plural words in Alor-Pantar. In Klamer, Marian (ed.), *Alor Pantar Languages: History and Typology*. Leiden: Brill.

TIDORE

Author unknown. *Valence in Tidore* (unpublished questionnaire). Available online at http://chl.anu.edu.au/linguistics/projects/Conferences/EastNusantara/ValenceTidore.rtf (retrieved on April 10, 2014).

Staden, Miriam van. 2000. *Tidore: A linguistic description of a language of the North Moluccas* (PhD dissertation). Leiden: Leiden University.

TURKISH

Lewis, Geoffrey L. 1978. *Turkish grammar*. Oxford: Clarendon Press.

Kornfilt, Jaklin. 1997. *Turkish*. New York: Routledge.

WEST GREENLANDIC

Bittner, Maria and Ken Hale. 1996. Ergativity: Toward a Theory of a Heterogeneous Class. *Linguistic Inquiry* 27, 531-604.

Fortescue, Michael. 1984. *West Greenlandic*. Sydney: Croom Helm Australia.

Sadock, Jerrold M. 2003. *A grammar of Kalaallisut (West Greenlandic Inuttut)*. München: Lincom.

# Summary in English

Languages are often seen as mappings between meanings and forms. In order to transfer a message, a speaker maps the intended meaning onto forms, such that a hearer can decode the forms and comprehend the meaning. From such a perspective, it appears intuitively most efficient for a language to map each meaning onto a single form. In fact, research into language acquisition suggests that such so-called transparent relations are easiest to acquire for L1 and L2 learners. Nonetheless, it is commonly observed that in all languages of the world, non-transparent (or: opaque) structure is manifested to some extent in the grammar and in the lexicon. The point of departure of this dissertation is that even though all languages exhibit non-transparency, they do so to a variable degree, meaning that all languages are opaque, but some languages are more opaque than others.

A few small studies tested the hypothesis that languages show variation in the extent to which they employ opaque relations in their grammars. These studies, viz. Leufkens (2010), Hengeveld (2011b) and Leufkens (2013) all corroborated this hypothesis, and, interestingly, they showed that non-transparent features are not randomly distributed over languages, but that they show implicational relations. Certain non-transparent features are attested in virtually all languages of the world, while others are exhibited only by languages with the highest degree of non-transparency. This interesting finding sparked the idea that there is a typology of transparency, such that not only languages can be ranked in terms of their degree of non-transparency, but that non-transparent features too can be ranked in terms of their correlation with the degree of non-transparency of languages.

In order to further corroborate, but also extend the findings of these earlier studies, this study compares 22 languages as regards their degree of transparency. Of course, such a comparative study necessitates a fine-grained metric that can distinguish between transparent and non-transparent relations, and thus, between highly transparent and less transparent languages. Such a metric is developed in this study on the basis of the theoretical framework of Functional Discourse Grammar

(henceforth FDG; Hengeveld & Mackenzie 2008), which allowed for a precise delimitation of the concept 'one unit of meaning' and 'one unit of form'.

FDG distinguishes between four levels of linguistic organisation: a pragmatic level (the Interpersonal Level, IL), a semantic level (the Representational Level, RL), the Morphosyntactic Level (ML) and the Phonological Level (PL). Each of these levels is itself layered and makes use of specialised sets of primitives that are stored in the lexicon. Now, transparency holds when one unit at one of the levels of organisation corresponds to exactly one unit at all other levels of organisation, for example when a pragmatic act of reference relates to a single semantic entity, which in turn evokes a morphosyntactic Noun Phrase, corresponding to a Phonological Phrase.

Transparency can be violated in five ways. Firstly, if one (pragmatic or semantic) unit is expressed by multiple formal (i.e. morphosyntactic or phonological) units, at least one of those formal units is redundant. Thus, a first category of non-transparent features is that of *redundancy*. A second category, called *form-based form* in this study, is constituted by formal elements that have no higher level counterpart whatsoever. *Fusion* is a third type of opacity, referring to cases where multiple meanings are expressed in a single form. Finally, *discontinuity* forms a fourth category, including all violations of domain integrity, resulting in fragmentary units that obscure boundaries of units. A logically possible fifth category consisting of one-to-null elements, i.e. meanings that can be argued to be present but that are not visible at the linguistic surface, is excluded from the study, because its existence is theoretically too controversial.

The term transparency has frequently been used in the linguistic literature, most notably in the subfields of theoretical linguistics, language acquisition studies and creole studies. Theoretical linguists such as Lightfoot (1979), Bybee (1985), Kusters (2003) and Dahl (2004) tend to see transparency as an optimal property of languages, in competition with other factors such as economy and expressivity. Within this so-called competing motivations paradigm, this competition is seen as the driving force behind language change. Similar views are adopted by language acquisition experts such as Slobin (1977) and Bates and MacWhinney (1989), who

argue that at least L1 learners have a preference for transparent structures, which are therefore acquired first.

In creole studies, transparency is usually treated on a par with simplicity, a controversial notion in that field since McWhorter (1998, 2001) stated that creoles were a typological class distinct from non-creoles, characterised by their high degree of simplicity. This statement received fierce criticism, directed firstly at the idea that creoles would in some sense be deficient or primitive, which McWhorter denied, and secondly at the particular metric that McWhorter proposed to measure complexity. This metric mainly involves the counting of overt morphological markings, and was therefore said to forego important complexities in creoles and isolating languages that are not morphologically marked, but present nonetheless.

The claim that creoles are relatively simple compared to other languages is often equated with the transparency hypothesis as originally developed by Seuren & Wekker (1986), who claim that a relatively high degree of transparency is characteristic of creoles. However, in my opinion, transparency and simplicity should by no means be seen as the same. The most crucial difference is that simplicity measures the amount of overt distinctions and rules in a particular domain of grammar or at a particular level of organisation, while transparency is a property of interfaces between such levels. Especially when examining relative complexity, i.e. difficulty of L1 or L2 acquisition, one needs to take both simplicity and transparency into account to be able to explain for instance the relatively early acquisition of the complex but transparent verbal morphology of Turkish.

To measure the degree of transparency of languages, a list of 20 non-transparent features was composed. These features can be categorised as regards the category of opacity to which they belong, i.e. redundancy, fusion, discontinuity and form-based form, and as regards the interface between levels at which they take place. By checking for each language in the sample which of the 20 features it displays, languages can be ranked in terms of the amount of non-transparency in their grammar, hence in terms of their degree of transparency. The features were tested in a binary way, since a finer grained metric that weighs features with respect to another or defines multiple values per feature is as yet not available.

In the category of redundancy, languages are tested on the presence of various agreement features. Of course, in a language with argument-predicate agreement, properties of the argument are expressed multiple times in a clause. In this study, such double referential expression is called agreement if both the argument and the predicate marking are obligatorily explicit, and referred to as cross-reference if the independent argument can be left implicit, e.g. in pro-drop languages. Further non-transparent features in the redundancy category are noun-modifier agreement, tense copying, i.e. the copying of a tense operator from the main clause to a complement clause, and plural concord, which refers to the expression of plurality on a noun modified by a numeral 'two' or higher.

A first non-transparent feature in the category of discontinuity that is tested in this study is extraposition, that is, the realisation of a head of a relative clause separately from its antecedent. Such a discontinuous constituent violates domain integrity, as what belongs together semantically is not realised as one unit morphosyntactically. The same is true in the case of argument raising, since in that case, an argument of a complement clause is realised as part of the main clause, where it does not belong semantically. Furthermore, circumfixes violate domain integrity as they are relations between one semantic operator and multiple phonological strings, and infixes create discontinuity in their hosts.

In the category of fusion, three non-transparent features are distinguished. Firstly, it is examined whether languages exhibit cumulation, that is, joint expression of multiple semantic categories into one morphological element, called a portmanteau morpheme. Since cumulation of person and number in pronouns is attested in almost all languages, it is more fruitful to consider cumulation of TAME (for the verbal domain) and of case (in the nominal domain) with other semantic categories only. Secondly, the study includes morphologically conditioned stem alternations, i.e. alternations to stems such that the morphological output form expresses lexical and grammatical information in one unsegmentable element. Two types of such stem alternations exist: suppletion, involving a complete replacement of the stem, and irregular stem formation, in which only part of the stem is altered.

Finally, languages are assessed as regards the presence of form-based form features. Firstly, it is shown whether the language exhibits nominal expletives with weather predicates, as those are pronouns without a referent. Secondly, this category contains the non-transparent feature of grammatical gender, a nominal classification system that has no pragmatic or semantic motivation. Similarly, languages may show syntactic functions like subject, which are not motivated by pragmatic roles such as topic and focus, nor by semantic roles like Actor and Undergoer. A fourth feature studied in this category is influence of complexity on word order, which determines whether the syntactic weight of constituents determines their morphosyntactic placement, rather than pragmatic or semantic ordering principles. It is also examined whether functions are marked in languages by means of phrase markers, i.e. clitics and particles, or by head-markers, i.e. affixes. The latter case is non-transparent, because the function marker is sensitive to the morphosyntactic complexity of its host. Finally, languages are tested on the presence of morphophonologically and phonologically conditioned alternations of morphemes, since these are all processes that have no pragmatic or semantic trigger, but purely a formal motivation.

The presence of these features is assessed in a sample of 22 languages. The sample was composed by means of the variety sampling method by Rijkhoff et al. (1993), which ensures the genetic diversity of the sample. By including more languages from families that show a greater internal diversity, the sample also aims to represent a large range of typological variation. The original plan was to test 25 languages, but due to time limitations, three languages had to be dropped. Due to this, and due to the use of Ruhlen's language classification that presupposes distant genetic relationships, some families may be underrepresented in the sample, most notably the American language families.

The examination of the features in these languages show that Dutch is unequivocally the most opaque language in the sample, while Teiwa and Fongbe are the most transparent ones. An explanation for non-transparency in languages is given in terms of communicative advantages that particular non-transparent features may have, such as regularisation and the demarcation of constituent boundaries.

Once non-transparent features have developed in a language for such reasons, they will only disappear if they become too difficult to acquire, but otherwise, they can be retained as 'linguistic male nipples' (Lass 1997). A high degree of transparency of a language can be explained as the result of language contact, as is argued in Chapter 3 but also for example by Trudgill (2011). The transparency of languages like Fongbe and Teiwa, which cannot be said to have undergone extensive language contact, may be accounted for by their morphology and phonology, as they are isolating languages. Perhaps, the expression of pragmatic and semantic units in distinct words rather than morphemes strengthens the transparent relation between meaning and form. It may also be the case that phonological change, which is often the starting point of language change, plays an insignificant role in isolating languages, which therefore do not develop non-transparent features.

The cross-linguistic distribution of non-transparent phenomena enables a ranking of non-transparent features into an implicational hierarchy, which matches earlier findings in typology, diachrony and language acquisition. Categorising the features with respect to their category of opacity, it becomes clear that this ordering reflects an ordering in terms of syntacticity, i.e. the degree to which linguistic forms are motivated by morphosyntactic information only. It turns out that form-based form features are rarely attested in languages, making them the best predictors of a high degree of non-transparency in a language. They are followed by obligatory redundancy features such as argument-predicate agreement, then by fusion features, which in turn implicate the presence of optional redundancy such as cross-reference. Finally, (morpho)phonologically conditioned alternations are attested in almost all of the languages studied. This latter finding shows that apart from an effect of opacity category, there is also an effect of interface. It turns out to be the case that a transparent relation between pragmatic units and other elements is violated more easily than a relation between a semantic unit and another element. In the same way, phonological elements are more easily opaque than morphological units are.

A few features could not be ranked with respect to other features in an implicational hierarchy. Firstly, discontinuous constituents turn out to appear in many languages even though they are not expected according to their degree of

transparency. This may be due to the fact that pragmatic information and processing matters also influence word order, thus obscuring the workings of transparency. Discontinuous morphology (infixes and circumfixes) has such strong morphological preconditions that they do not appear in languages in which they are expected. Finally, concord and cumulation are probably near-universal, but the subtypes tested here (plural concord and cumulation of TAME and case) did not show a pattern.

In conclusion, the study successfully ranks a multitude of non-transparent features with respect to each other, showing that transparency is a highly relevant variable in typology and multiple other domains of linguistics.

# Samenvatting in het Nederlands

Talen worden vaak gezien als een verzameling relaties tussen betekenissen en vormen. Om een bepaalde boodschap over te brengen, moet de spreker die boodschap omzetten van betekenissen naar vormen, zodat de hoorder die vormen weer kan omzetten naar de bedoelde betekenis. Op die manier beschouwd lijkt het maximaal efficiënt wanneer iedere betekenis aan precies één vorm gerelateerd is. Onderzoek naar eerste- en tweedetaalverwerving wijst bovendien uit dat zulke relaties het makkelijkst te leren zijn. Desondanks blijkt dat alle natuurlijke talen van de wereld op zijn minst enige vorm van non-transparantie laten zien in hun grammatica en lexicon. Deze dissertatie gaat uit van het idee dat alle talen non-transparantie kennen, maar dat sommige talen 'ondoorzichtiger' zijn dan andere.

Enkele kleine studies, nl. Leufkens (2010), Hengeveld (2011b) en Leufkens (2013), hebben laten zien dat er inderdaad variatie is in de mate waarin talen non-transparant zijn. Bovendien bewijzen deze studies dat verschillende non-transparante verschijnselen niet willekeurig verspreid zijn in talen, maar dat ze implicationele verbanden laten zien. Bepaalde non-transparante verschijnselen komen voor in alle talen van de wereld, terwijl andere alleen voorkomen in sterk non-transparante talen. Deze interessante constatering heeft geleid tot het idee dat er een typologie van transparantie bestaat, oftewel dat talen niet alleen geordend kunnen worden op basis van hun mate van transparantie, maar dat ook non-transparante eigenschappen geordend kunnen worden op basis van hun voorkomen in al dan niet transparante talen.

In deze studie worden 22 talen met elkaar vergeleken op hun graad van transparantie, om aldus de eerder gevonden resultaten te repliceren, te preciseren en uit te breiden. Natuurlijk kan een dergelijke vergelijking alleen plaatsvinden op basis van een meetinstrument dat kan vaststellen wat transparant is en wat non-transparant is, en op die manier welke taal relatief transparant is en welke relatief non-transparant. Een dergelijk meetinstrument is ontwikkeld in deze studie met behulp van een taalkundige theorie genaamd *Functional Discourse Grammar* (voortaan

FDG; Hengeveld & Mackenzie 2008). Deze theorie stelde me in staat om concrete definities te geven van ‘één betekenis’ en ‘één vorm’ en op die manier het begrip transparantie te operationaliseren.

FDG onderscheidt vier niveaus van linguïstische organisatie, namelijk een pragmatisch niveau (*Interpersonal Level*, IL), een semantisch niveau (*Representational Level*, RL), een morfosyntactisch niveau (*Morphosyntactic Level*, ML), en een fonologisch niveau (*Phonological Level*, PL). Deze niveaus zijn intern gelaagd en maken gebruik van bouwstenen zoals Lexemen en Woorden, die zijn opgeslagen in het lexicon. Een fenomeen is transparant wanneer een bouwsteen op een van deze niveaus correspondeert met precies één bouwsteen op de andere niveaus, bijvoorbeeld wanneer een pragmatische *Act of Reference* correspondeert met een semantische *Individual*, die op zijn beurt een *Noun phrase* oproept, en vervolgens uitgedrukt wordt in een *Phonological Phrase*.

Transparantie kan op vijf manieren geschonden worden. Ten eerste kan een enkel pragmatisch of semantisch element uitgedrukt worden door meer dan een morfosyntactisch of fonologisch element. In dat geval is een van die formele elementen redundant. Deze eerste categorie van non-transparantie heet daarom *redundantie*. Een tweede categorie, in deze dissertatie *vorm-gebaseerde vorm* geheten, omvat alle formele (morfosyntactische of fonologische) elementen die niet corresponderen met een element op het pragmatisch of semantisch niveau en dus als het ware betekenisloos zijn. *Fusie* is de derde categorie van non-transparantie, die alle gevallen omvat van enkele vormen die corresponderen met meer dan een betekeniselement. Ten vierde is er de categorie *discontinuïteit*, waarin schendingen van domeinintegriteit zijn opgenomen, die tot gevolg hebben dat de grenzen van elementen (bijv. woorden) niet meer herkenbaar zijn, zodat we niet meer kunnen spreken van ‘één vorm’. Een vijfde categorie bestaat uit pragmatische of semantische elementen die aangenomen worden onderliggend aanwezig te zijn maar niet zichtbaar zijn in de concrete uiting. Deze vijfde categorie is uitgesloten van dit onderzoek, omdat het bestaan van dergelijke elementen theoretisch te controversieel is.

De term transparantie is vaak gebruikt in de linguïstiek, vooral in de theoretische taalkunde, in het gebied van taalverwerving, en in de creolistiek. Theoretisch taalkundigen als Lightfoot (1979), Bybee (1985), Kusters (2003) en Dahl (2004) zien transparantie als een optimale eigenschap van talen, die in competitie is met andere factoren zoals economie en expressiviteit. Binnen dit paradigma van 'concurrerende factoren' wordt die competitie gezien als de drijvende kracht achter taalverandering. Experts op het gebied van taalverwerving, zoals Slobin (1977) en Bates & MacWhinney (1989), hanteren een soortgelijke visie aangezien ze stellen dat eerstetaalverwervers een voorkeur hebben voor transparante structuren, die dan ook als eerste verworven worden.

In de creolistiek wordt transparantie meestal gezien als synoniem aan simpliciteit, een notie die controversieel is sinds McWhorter (1998, 2001) stelde dat creooltalen fundamenteel simpeler zijn dan niet-creolen en daarom een aparte typologische klasse vormen. Dit idee is heftig bekritiseerd, ten eerste omdat McWhorter ermee zou bedoelen dat creooltalen op enige manier primitief of gebrekkig zouden zijn, wat hij ontkent. Een tweede punt van kritiek richtte zich op het meetinstrument dat McWhorter voorstelde om complexiteit te meten, dat bestaat uit het tellen van het aantal in de zin zichtbare morfosyntactische elementen. Volgens zijn critici zou McWhorter hiermee voorbijgaan aan belangrijke complexe eigenschappen van creooltalen en andere isolerende talen die zich niet in morfosyntactische zin openbaren, maar wel degelijk complex zijn.

Het idee dat creolen relatief simpel zijn in vergelijking met andere talen wordt vaak gelijkgesteld met de transparantiehypothese van Seuren & Wekker (1986), die behelst dat een relatief hoge graad van transparantie karakteristiek is voor creooltalen. Mijns inziens moeten transparantie en simpliciteit echter nooit als synoniem gezien worden. Het meest cruciale verschil is dat simpliciteit het aantal aan de oppervlakte zichtbare elementen meet in een specifiek domein of op een specifiek niveau van een taal, terwijl transparantie een eigenschap is van de interfaces tussen niveaus. Vooral bij het bestuderen van eerste- en tweedetaalverwerving moeten zowel simpliciteit als transparantie in ogenschouw worden genomen, aangezien men anders niet kan verklaren dat bijvoorbeeld de

complexe, maar relatief transparante werkwoordsmorfologie van het Turks relatief vroeg verworven wordt door kinderen.

Om de mate van transparantie van diverse talen te meten, heb ik een lijst opgesteld bestaande uit 20 non-transparante eigenschappen. Deze eigenschappen kunnen worden gecategoriseerd aan de hand van het type non-transparantie dat ze vertonen, nl. redundantie, fusie, discontinuïteit of vorm-gebaseerde vorm, en aan de hand van de interface waarop ze plaatsvinden. Door voor iedere taal in het voor dit onderzoek samengestelde sample na te gaan welke van deze 20 eigenschappen hij bezit, konden de talen geordend worden op hun mate van transparantie. Aan iedere eigenschap werd een binaire waarde toegekend, aangezien er op dit moment geen meetinstrument bestaat dat voor deze eigenschappen een preciezere, graduele waarde kan toekennen.

In de categorie redundantie zijn talen getest op de aanwezigheid van verschillende types congruentie. In een taal waarin eigenschappen van argumenten ook op het predicaat worden uitgedrukt, is sprake van een verdubbeling van informatie. In deze dissertatie wordt dat congruentie genoemd als zowel het argument als de predicaatsmarkering verplicht expliciet zijn, maar spreek ik van cross-referentie als het argument ook impliciet kan zijn, zoals in pro-droptalen. Andere non-transparante eigenschappen in deze categorie zijn congruentie tussen een zelfstandig naamwoord en de modificeerder daarvan, en tijdscongruentie, oftewel het kopiëren van de werkwoordstijd van de hoofdzin naar een ondergeschikte bijzin. Verder wordt gekeken naar meervoudsconcordantie, dat wil zeggen het verschijnsel dat meervoud gemarkeerd wordt op een naamwoord dat gemodificeerd wordt door een telwoord ‘twee’ of hoger.

Een non-transparante eigenschap in de categorie discontinuïteit is extrapositie: het uitdrukken van een relatieve bijzin op een locatie niet grenzende aan het hoofd van die bijzin. Een dergelijke discontinue constituent schendt domeinintegriteit, aangezien elementen die op semantisch niveau bij elkaar horen niet bij elkaar staan op morfosyntactisch niveau. Hetzelfde geldt in het geval van *raising*, waarbij een argument van een ondergeschikte bijzin wordt gerealiseerd als een syntactisch element van de hoofdzin, zodat semantiek en morfosyntaxis niet

parallel lopen. Circumfixen schenden domeinintegriteit, aangezien ze relaties betreffen tussen een semantische operator en meerdere fonologische strings, en ook infixen schenden domeinintegriteit aangezien ze discontinuïteit creëren van de stam waaraan ze zich hechten.

In de categorie fusie worden drie non-transparante verschijnselen onderscheiden. Ten eerste wordt onderzocht of er sprake is van cumulatie, oftewel een gezamenlijke expressie van meerdere semantische categorieën in één morfologisch element, dat een *portmanteau* morfeem genoemd wordt. Aangezien vrijwel alle talen van de wereld cumulatie van persoon en getal in pronomina laten zien, is de bestudering daarvan niet zinvol. Nuttiger is om de cumulatie van markeerders van tijd, aspect, modaliteit en evidentialiteit (TAME) met andere semantische categorieën in werkwoordsmarkering te bekijken, en de cumulatie van naamvalsmarkering met andere categorieën op zelfstandig naamwoorden te bekijken. Ten tweede onderzoek ik het vóórkomen van morfologisch geconditioneerde stamalternantie, zodanig dat de resulterende stam zowel een lexicale als een grammaticale betekenis uitdrukt middels één ondeelbare vorm. Stamalternantie komt voor in twee typen, namelijk suppletie, waarbij de stam een volledig andere vorm krijgt onder invloed van grammaticale markering, en onregelmatige stemvorming, waarbij een gedeelte van de stem (bijvoorbeeld de klinker) wordt aangepast.

Tot slot worden talen onderzocht op de aanwezigheid van non-transparantie van het type vorm-gebaseerde vorm. Zo wordt bekeken of de taal dummypronomina gebruikt bij weerpredicaten, aangezien dergelijke pronomina geen referent hebben. Ten tweede wordt bekeken of de taal grammaticaal geslacht heeft, d.w.z. een nominaal classificatiesysteem dat niet gemotiveerd wordt vanuit de pragmatiek of semantiek. Op dezelfde manier kunnen talen een syntactische functie zoals subject hebben die niet teruggaat op een pragmatische of semantische functie. Een vierde eigenschap in deze categorie is de invloed van complexiteit op woordvolgorde, oftewel het verschijnsel dat het morfosyntactisch gewicht van een zinsdeel de locatie van dat zinsdeel in de zin bepaalt. Dat is non-transparant, aangezien woordvolgorde dan niet bepaald wordt door pragmatische of semantische overwegingen. Verder

wordt bekeken of grammaticale functies in de talen uitgedrukt worden door markeerders die de hele frase markeren, nl. door clitics of partikels, of door markeerders die alleen het hoofd van de frase markeren, nl. affixen. Het laatste geval is non-transparant, omdat de markeerder dan gevoelig is voor de morfosyntactische complexiteit van zijn gastheer. Tot slot worden talen getest op de aanwezigheid van morfofonologisch en fonologisch geconditioneerde alternanties in zowel woordstammen als affixen. Ook dergelijke alternanties hebben een puur formele, en geen pragmatische of semantische motivatie.

De aanwezigheid van bovengenoemde non-transparante eigenschappen is getest in een sample van 22 talen. Dit sample is samengesteld volgens de *variety sampling* methode van Rijkhoff et al. (1993), die een zo groot mogelijke genetische diversiteit van het sample nastreeft. Doordat deze methode meer talen uit een familie selecteert naarmate de interne variatie binnen een familie groter is, is ook gepoogd een zo groot mogelijke typologische variatie te creëren. Aanvankelijk was het de bedoeling 25 talen op te nemen in het sample, maar vanwege tijdsbeperkingen zijn drie talen komen te vervallen. Hierdoor, en door het gebruik van de classificatie van Ruhlen die talen samenneemt die hooguit een verre verwantschap vertonen, is het mogelijk dat met name Zuid-Amerikaanse talen enigszins ondervertegenwoordigd zijn in dit onderzoek.

De uitkomsten van het hier beschreven onderzoek naar de aanwezigheid van non-transparante eigenschappen in 22 talen laten zien dat het Nederlands de minst transparante taal in het sample is, terwijl het Teiwa en het Fongbe het meest transparant zijn. Een verklaring voor een hoge graad van non-transparantie is dat bepaalde non-transparante eigenschappen communicatief voordelig kunnen zijn, bijvoorbeeld omdat ze een regulariserend effect hebben of omdat ze de grenzen van constituenten markeren. Als non-transparante eigenschappen zich in een taal eenmaal ontwikkeld hebben, zullen ze slechts verdwijnen wanneer ze niet meer leerbaar zijn, maar anders zullen ze aanwezig blijven als ‘linguïstische mannentepels’ (Lass 1997). Een hoge mate van transparantie in een taal is te verklaren als het resultaat van intensief taalcontact, zoals in Hoofdstuk 3 wordt beschreven, en ook door bijvoorbeeld Trudgill (2011) wordt verdedigd. De

transparantie van Fongbe en Teiwa, talen die geen intensief taalcontact hebben ondergaan, kan verklaard worden door hun morfologie en fonologie, aangezien dergelijke transparante talen isolerend zijn. Wellicht wordt een transparante relatie tussen betekenis en vorm versterkt wanneer die betekenis uitgedrukt wordt middels aparte woorden, in plaats van morfemen. Een andere mogelijkheid is dat fonologische verandering, vaak het beginpunt van taalverandering, in isolerende talen een relatief kleine rol speelt, waardoor die talen geen non-transparante eigenschappen ontwikkelen.

De cross-linguïstische distributie van non-transparante verschijnselen maakt het mogelijk om non-transparante eigenschappen te ordenen in een implicationele hiërarchie, die overeenstemt met eerdere resultaten uit de typologie, diachronie en taalverwerving. Door de eigenschappen naar categorie in te delen wordt duidelijk dat de gevonden ordening te verklaren is aan de hand van het begrip syntacticiteit, oftewel de mate waarin vormen te verklaren zijn vanuit morfosyntactische informatie. Uit de resultaten blijkt namelijk dat syntactische eigenschappen uit de categorie vorm-gebaseerde vorm alleen in de minst transparante talen voorkomen, en dus de beste voorspellers zijn van een hoge graad van non-transparantie van talen. Deze eigenschappen woorden gevolgd door redundantie-eigenschappen zoals congruentie, en vervolgens door fusie-eigenschappen. Daaronder in de hiërarchie staan optionele redundantie-eigenschappen zoals cross-referentie, en als laatste vinden we vorm-gebaseerde vormeigenschappen op (morfo-)fonologisch niveau. Met name dit laatste gegeven laat zien dat er behalve een effect van het type non-transparantie, ook een effect is van de interface waar de non-transparantie zich afspeelt: een transparante relatie tussen een pragmatische eenheid en een ander element wordt makkelijker geschonden dan een relatie tussen een semantische eenheid en een ander element, en op dezelfde manier zijn relaties met fonologische elementen vaker non-transparant dan relaties met morfosyntactische elementen.

Enkele non-transparante verschijnselen konden niet geordend worden ten opzichte van andere eigenschappen. Ten eerste blijken discontinue zinsdelen voor te komen in talen waar ze op basis van de mate van transparantie van die talen niet

verwacht worden. Dit zou kunnen komen doordat pragmatische informatie en verwerkingsfactoren ook woordvolgorde beïnvloeden, en op die manier de werking van transparantie tenietdoen. Discontinue morfologie (circumfixen en infixen) hebben dermate sterke morfologische precondities dat ze niet verschijnen in talen waar ze, op basis van hun graad van transparantie, wel verwacht worden. Verder zijn concordantie en cumulatie waarschijnlijk vrijwel universeel, maar laten de subtypes die hier getest zijn (respectievelijk meervoudsconcordantie en cumulatie van TAME-markering en naamvalsmarkering met andere categorieën) op zichzelf geen patroon zien.

Deze uitzonderingen daargelaten heeft dit onderzoek een groot aantal non-transparante eigenschappen ten opzichte van elkaar geordend. Het laat daarmee zien dat transparantie een zeer relevante variabele is in de typologie en andere domeinen van de linguïstiek.

# Curriculum Vitae

Sterre Leufkens was born in Delft on February 4th, 1986. After completing the gymnasium with a cum laude degree in 2004, she started her bachelor in linguistics at the University of Amsterdam. This bachelor included the acquisition of Egyptian Arabic and Modern Standard Arabic. Sterre graduated cum laude in 2007, and took a gap year in which she worked as a research assistant. In 2008, she began a research master in Linguistics, following courses in theoretical linguistics, typology, and (formal) semantics. A tutorial with Kees Hengeveld on the transparency of Kharia led to an MA thesis on the transparency of creole languages (later published as Leufkens 2013), which in turn led to a PhD project that Sterre started in September 2010 after graduating cum laude from the RMA. This PhD was funded by means of a grant *Promoveren in de geesteswetenschappen* from NWO.

During her studies and her PhD, Sterre has participated in many organisational activities. She organised conferences and events as board member of student's association VOS, as board member of the alumni circle for linguists at the UvA, and as the coordinator of research group Learnable and Unlearnable Languages. She presented her work at local events, but also at international conferences, such as the FDG conference in Lisbon (2010), *Societas Linguistica Europea* in Stockholm (2012), and at the Creole Grammars & Linguistic Theories' conference in Paris (2011). Furthermore, Sterre presented and discussed her work during a three month research stay at the Max Planck Institut für Evolutionäre Anthropologie in Leipzig, in the fall of 2013.

Outside the university, Sterre engaged in writing a linguistic popular-scientific blog together with friend and linguistics student Marten van der Meulen, in which they bring linguistic knowledge to a larger audience. Furthermore, as a volunteer for LGBT-organisation COC, Sterre discussed homosexuality at secondary schools in Amsterdam. Being an enthusiastic amateur violin player, she was part of the Ricciotti Ensemble, an orchestra bringing music to people who for some reason are unable to experience live music on their own.