Towards a Theory of Communicative Efficiency in Human Languages

Habilitationsschrift

zur Erlangung des akademisches Grades

Dr. phil. habil.

der Philologischen Fakultät

der Universität Leipzig

eingereicht von

Dr. Natalia Levshina

(geboren am 29.06.1978 in Pskow, Russland)

angefertigt am Institut für Anglistik

Beschluss über die Verleihung des

akademischen Grades vom

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Habilitationsschrift selbständig und ohne unerlaubte fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die im Schriftenverzeichnis angeführten Quellen benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, 25.11.2018

Natalia Levshina

_____

Unterschrift

To Björn

Contents

Tables

Figures

Abbreviations

1	1st person
2	2nd person
3	3rd person
1+2	1st person plural inclusive
A	syntactic function of an Agent
ABS	absolutive
ACC	accusative
AGT	agentive
ART	article
ASP	aspect
B	bare (pronoun)
CAUS	causative
CONV	converb
DAT	dative
DEF	definite
DIR	directional
DO	direct object
ERG	ergative
EXP	experiential aspectual particle
F	female
FA	father
FOC	focus
GEN	genitive
IM.P	immediate past
IND	indicative
INF	infinitive
intr.	intransitive
M	male
NAR	narrative
NOM	nominative
NPST	non-past
OBJ	object
P	syntactic function of a Patient
ℙ	probability
PART	particle
PASS	pass-by (a morpheme)
PERF	perfective
PL	plural
POL	polite (pronoun)
POSS	possessive
PRO	pronoun
PST	past
REC.P	recent past
Q	question marker
SG	singular
T.LNK	topic linker
TNS	tense
tr.	transitive

Acknowledgements

The work presented here has been carried out as part of the project “Grammatical Universals”, which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 670985). I am indebted to Martin Haspelmath for enabling and inspiring this research and for giving me full creative freedom, and to all members of the Nikolai-Lab in Leipzig for their stimulating input and fruitful discussions. I’m also grateful to the participants of the Diversity Linguistics Seminar, and to the colleagues from the Institute for British Studies, the Institute for Linguistics and other departments and universities, who provided insightful feedback on some of the ideas presented here. My most loving thanks go to my husband Björn, who motivated and supported me at all stages of this adventure, sharing with me his positive energy and his personal library.

Part I. Communicative efficiency: Theoretical preliminaries

Chapter 1. Efficiency in human communication: principles and manifestations

1.1. Basic concepts

1.1.1. The Principle of Communicative Efficiency

The aim of the present study is to develop a theory of communicative efficiency in human languages, based on general pragmatic principles. There are many cases in human communication when the speaker has different options with regard to the amount of effort needed to pronounce or write something. Some examples are given in (1). In (1a), the speaker may choose between the first, shorter form *sang* and the second, longer expression. In (1b), one can use the lexical causative *stop* or the periphrastic causative *get X to stop*. Example (1c) contains different referential expressions: the longer proper name *Mary*, and the shorter pronominal forms *she*. In (1d), the difference between the sentences is in the use or absence of the complementizer *that*. In (1e), the speaker can choose between the clipped form *math* and the full form *mathematics*. The example in (1f) contrasts the analytic and synthetic comparative forms of adjectives, which can sometimes be used interchangeably in English. The example in (1g) illustrates variation in pronunciation of *I don't know*. The variants differ in the total length, the presence or absence of the pronominal subject and amount of articulatory detail. Finally, (1h) shows an instance of the genitive alternation, where the older Saxon genitive with *-'s* is shorter than the newer Norman genitive with *of*.

(1) a. *Mary **sang** Jingle Bells. – Mary **produced sounds that reminded o**f Jingle Bells.*

b. *John **stopped** the car. – John **got** the car **to stop**.*

c. ***Mary** entered the room. **She** saw a stranger in a black coat.*

d. *She believes you are here. – She believes **that** you are here.*

e. *I'm studying **maths**. – I'm studying **mathematics**.*

f. *Jane is **cleverer** than Mary. – Jane is **more clever** than Mary.*

g. *Dunno [ˈdəˈnəʊ]. - I don't know [aɪ dəʊn (t) ˈnəʊ].*

h. *the emperor**'s** family - the family **of the** emperor*

In all these pairs, the amount of the speaker's articulatory effort is different. The asymmetry is due to different segmental lengths. In addition to length differences, one can also observe different amount of phonological detail, as in (1g), the use or absence of particular units, as in (1d), or alternative expressions of varying length.

As for the functional differences between the short and long variants in (1), they are not always easy to pinpoint. It may be stylistic properties (e.g. *maths* is an informal variant of *mathematics* in British English), more or less conventionalized conversational implicatures, as in (1a) and (1b), or a set of factors related to processing, distribution of semantic classes, phonology, etc., as in the use or absence of *that* (cf. Jaeger 2010) and the genitive alternation (cf. Szmrecsanyi 2010). One can say that the meanings are more or less comparable.

One can also find formal asymmetries in more or less conventionalized pairs with contrastive grammatical meanings. A typical example is the singular – plural distinction. It is well known that singular forms are less often marked formally than plural forms (Greenberg 1966), as illustrated by the pair *book – books* in (2a). In (2b), the shorter form *furniture* has a collective use, whereas the longer form *a piece of furniture* has a singulative meaning. In (2c), the shorter form with the indefinite article *a finger* implicates conventionally the speaker's own finger, whereas other possessors should be introduced explicitly. In (2d), the embedded infinitival clause construction with the same subject is shorter than the expression with different subjects, where a finite clause is required.

(2) a. *(one) book-**Ø** – (five) book-**s***

b. *furniture – **a piece of** furniture*

c. *I broke **a** finger. – I broke **Peter's** finger.*

d. *I promised **to** do this task. – I promised **that Jane would** do this task.*

Finally, one can find signal asymmetries even when the words are neither semantically, nor formally related. One can take, for example, words like *I*, *in* and *be*, which are shorter than such words as words *harpsichord*, *archaeopteryx* and *gongoozle* 'to watch the passage of boats'. According to Zipf's (1935 [1965]) Law of Abbreviation, more frequent words tend to be shorter than less frequent ones. Although one can also find quite a few pairs where this does not hold (e.g. the word *understand* is more popular, yet longer than a physics term *quark*), the Law of Abbreviation predicts the correct relationship between length and frequency in the majority of cases (see more information in Section 1.2.7).

What do these examples have in common? In this study I will argue that in all these cases language users exhibit efficient communicative behavior when choosing between the shorter and longer forms. This approach continues a long intellectual tradition. Among the precursors are the following laws, principles and hypotheses:

- Zipf's Law of Abbreviation and the more general Principle of Least Effort (Zipf 1935 [1965], 1949);
- Gricean maxims of Quantity (and partly of Manner) and Neo-Gricean maxims (Grice 1975; Horn 1984; Levinson 2000, etc.);
- Haiman's (1983) principle of economy;
- Du Bois's (1985) dictum "Grammars code best what speakers do most";
- Keller's (1994: 107) hypermaxim "Talk in such a way that you are socially successful, at the lowest possible cost" and maxim "Talk in such a way that you do not spend more energy than you need to attain your goal";
- Hawkins' (2014) principle "Minimize Forms";
- Givón's (2017: 157) code–quantity principle;
- Haspelmath's (Forthcoming-a) grammatical form–frequency correspondence hypothesis.

The novelty of this study is that it brings together phenomena described in different domains of research (most importantly, usage-based linguistics, psycholinguistics, pragmatics,

typology) in an attempt to provide a unified account, which is based on fundamental pragmatic principles and involves probabilistic information about linguistic meanings and forms.

In a very general form, one can formulate the following principle of human communication, which can be called the Principle of Communicative Efficiency:

(3) **The Principle of Communicative Efficiency**

Communicate in such a way as to maximize the benefit-to-cost ratio.

By benefits I mean various cognitive effects that the speaker wants to invoke in the hearer. Cognitive effects are a central notion of Relevance Theory (Sperber and Wilson 1986/1995). They represent new contextual implications derived from the message or lead to strengthening or revision of existing assumptions.[1] In this study, I assume a maximally broad interpretation of cognitive effects: changes in the intellectual state of the hearer as a result of obtaining some information, changes in his or her emotional state and changes in the attitude towards the speaker and their interaction. These changes correspond to Jakobson's referential, poetic and phatic functions of language use (Jakobson 1960 [1971]).

As for costs, these are are articulatory and cognitive efforts required for language production. In this study, I will focus on articulatory efforts, which are approximated by segmental length (see Section 1.1.2).

It can be claimed that efficiency in communication and in other domains is an inherent property of living organisms, which is a product of biological evolution. Natural selection favours the behaviour in which the benefit-to-cost ratio is maximized. The individuals which exhibit behaviour with higher benefit-to-cost ratios will leave more copies of their genes (Ha 2010). For example, Zach (1979) found efficient foraging behaviour in Northwestern crows, who feed on whelks (sea snails) by dropping them from a height in order to break them. The birds preferred the largest whelks, which have a higher caloric content and broke more readily than medium and small ones. Since ascending flight was energetically expensive, the crows minimized the total amount of ascending flight required for breaking whelks by choosing the optimal height of drop. As a result, they achieved a large positive difference between the

[1] In Relevance Theory, maximum relevance is achieved with maximal cognitive effects and minimal cognitive efforts. Human cognition is geared to the maximization of relevance. Every act of ostensive communication conveys the presumption of its optimal relevance (Sperber and Wilson 1986/1995: 260–72).

amount of calories gained from the large and breakable whelks and the amount of calories spent flying.

This study focuses on the amount of speaker's effort. How successfully the message comes across also depends on what kind of cues the speaker uses. An example is the use of T or V forms. When used inappropriately, they may lead to undesirable cognitive effects, e.g. inferences like "The speaker does not respect me!" or "The speaker is acting distant. I must have done something wrong!", which can lead to disruption of communication. Another important aspect is the order of cues. For instance, Hawkins (1994) argued that certain word orders of constituents are easier to process than others. These types of efficiency are outside the scope of this study.

This study also does not investigate how to maximize the cognitive effects for a given amount of effort. Instead, we will focus on the situations where the cognitive effects are pre-determined by the speaker's needs, e.g. to obtain desired objects, get useful information, find a mate or delegate a task. In order to reach these goals, the speaker needs to cause specific cognitive effects in the hearer. The communication is efficient when the speaker spends not more and not less energy than it is necessary to cause these cognitive effects. If a language user spends too much effort, this behaviour will not be efficient. If he or she spends too little, the cognitive effects may not be evoked and the goals will not be reached (cf. the Q- and R-principles in Horn 1984, see also Section 1.2.1). Since the speaker and the hearer mutually assume that the other is acting rationally, one can formulate two general heuristics, or rules of thumb, which work both for the hearer and for the speaker:

(4) **Low-Cost Heuristic**

"Low costs – Small cognitive effects":

If the speaker uses a low-cost way of communicating some information, the latter is not supposed to change the previous cognitive state of the hearer substantially.

High-Cost Heuristic

"High Costs – Large cognitive effects":

If the speaker uses a high-cost way of communicating some information, the latter is supposed to change the previous cognitive state of the hearer substantially.

The cognitive state of the hearer is changed to the extent that the conveyed information differs from his or her current mental state, which is defined by recently activated contextual information and the resting activation of relevant exemplars and concepts in the long-term memory. The more informative (in the information-theoretic sense), or less predictable, accessible, activated, expected, etc. some message is, the greater the change of the hearer's cognitive state. Therefore, if the hearer evaluates the chosen linguistic form as costly, he or she infers that the message requires a larger deviation from his or her actual cognitive state. If the hearer evaluates the form as 'cheap' in terms of effort, then he or she infers that no large deviation from the current state is required.

One also needs to mention situations when the information that is omitted is simply irrelevant. Speakers do not waste their effort on transfer of information that will not lead to any useful cognitive effects in the hearer's mind. This corresponds to Givón's (2017: 3) principle of so-called cataphoric zeros: "Unimportant information need not be mentioned". An example is omission of the agent in passive constructions (see Section 1.2.2). In this study, however, I will mostly focus on those cases, where relevant information can be expressed by different amount of coding.

1.1.2. Cheap inference and expensive articulation

This subsection discusses the most important ways of minimizing communicative costs. As pointed out by Levinson (2000: 28), human speech encoding is, due to the physiological constraints on articulation, the slowest stage in human communication. It also consumes muscular energy. All other aspects of speech production and comprehension can run at a much higher rate, including inference. So,

> inference is cheap, articulation expensive, and thus the design requirements are for a system that maximizes inference. (Levinson 2000: 29)

In order to achieve his or her own communicative goals, the speaker needs to exploit the hearer's cognitive resources. Since inferences are cheap, the hearer agrees to take part in the interaction and spend this effort. Why would the hearer agree to participate in this game? He or she may be interested in getting useful information, and in maintaining the social connection

with the speaker, hoping that the speaker will cooperate in return. The biological bias towards helping an organism with similar genes to survive and propagate, which can be seen as the evolutionary basis for altruism, may also be a driving factor.

Thus, in order to be efficient, the speaker needs some help from the hearer, and some amount of cooperation. To what extent they cooperate, is an interesting and yet unresolved question. According to the Principle of Least Collaborative Effort, which was proposed by Clark and Wilkes-Gibbs (1986),

> participants in a contribution try to minimise the total effort spent on that contribution – in both the presentation and acceptance phases. (Clark and Schaefer 1989: 269)

Not everyone agrees with this view. According to Davies (2007), there is not enough evidence that the effort is indeed saved by the interactors jointly. In general, it is very difficult to measure the joint effort. In my opinion, it is more likely that each individual participant tries to minimize his or her own efforts. Yet, this individualistic, "selfish" efficiency ultimately leads to a more efficient communicative system for everyone.

How can one minimize the articulatory effort, at the same time achieving the intended cognitive effects in the hearer? Obviously, the speaker can spare effort when the information that one wants to convey is already available to the hearer or can be inferred based on the context and general knowledge. In particular, the following types of information are relevant:

(i) extra-linguistic contextual and linguistic co-textual information;

(ii) encyclopaedic information about the world (e.g. typical scenarios), based on the language user's experience;

(iii) probabilistic associations between formal units, linguistic categories and forms-concepts, which come from the language users' experience with language in use.

This information represents the common ground, i.e. the information that the participants believe they share (Clark 1996). It increases during the interaction, when the communicators coordinate their beliefs and communicative tools.

Let us focus on (i): Some information may come from the immediate context. A good illustration is reference management. For example, in Lao the initial references are done with

full noun phrases, and subsequent references are made by reduced nominal forms such as pronouns or zero anaphora, i.e. complete ellipsis of phonological material in the argument slot. This is the default reference management device in many languages. Consider an example from Lao (cf. Enfield 2007: Section 8.3). The first time the full noun is used to mention Mone, the son of the couple addressed. In the second sentence, a personal pronoun is used. Finally, we observe a zero anaphora:

(5) (Lao: Tai-Kadai, Enfield 2007: 488)

daj4-ñin2 ***bakø-mòòn3***

hear M.B.-M

'(I) heard (from) Mone.'

man2 *mùa2* *qaw3* *ngen2* *nam2* *khòòj5*

3.B return take money with 1SG.POL

'He went to get money from me.'

vaa1 *qiø-phòò1* *kaø* *tok2 khan2daj3 vaa1 san4*

say F.B.-Fa T.LNK fall stairs say thus

'(He) said, Dad fell (from) the stairs, so (he) said.'

The referential expressions are smaller when the referent is accessible from previous context. More information about these strategies is provided in Section 1.2.2.

There are also more subtle reduction effects. In particular, words that have been used in discourse tend to have reduced pronunciation more often than those that are introduced for the first time in the given text. For instance, Fowler and Housum (1987) found that 71 percent of second occurrences of nouns in spontaneous narratives were shorter than the first occurrence of the same noun. Fowler and Housum explained the results in terms of predictability: words that are predictable from the previous context are durationally shorter than new words, which are less predictable. Similar effects on final *-t* and -d deletion in English are reported by

Gregory et al. (1999). Other possible causes of phonetic reduction are discussed in Section 1.2.6.

Another source of information that the speaker can rely on is encyclopaedic knowledge (ii). For example, it allows one to use bridging implicatures, as in the example below:

(6) *I bought a new bicycle. Unfortunately, the saddle is very uncomfortable.*

Here, the speaker relies on the hearer's knowledge about bicycles, which have saddles. This knowledge allows for the omission of a connecting sentence 'The bicycle has a saddle'.

Similarly, one needs knowledge of the world in order to interpret some nominal compounds (at least, when hearing them for the first time), e.g. *paper clip* (i.e. a clip for paper), *paper factory* (i.e. a factory that produces paper) and *paper plate* (i.e. a plate made of paper) (cf. Hawkins 2014: 16).

Relevant information comes also from the language users' long-term experience with language, listed under (iii). More exactly, this is information about the probabilities of occurrence of diverse forms, meanings and their combinations. An example from Levinson (2000: 138) is provided below:

(7) a. *Her mansion is on the corner.*

b. *Her immodest, pretentious house is on the corner.*

In (7a), the use of the word *mansion* instead of the typical word *house* signals the hearer that the speaker implies an unusual interpretation, something similar to (7b). More about that will follow in Section 1.2.1. It is argued here that the expectations based on previous discourse experience play a decisive role in making language efficient.

All these types of information can interact. For instance, information from immediate co-text can be intertwined with expectations based on previous linguistic experience. An example is the omission of the referent when it coincides with the subject of the matrix clause (Comrie 1986: 89–90).

(8) a. *Peter intends to do the task himself.*

b. *Peter intends that Mary should do the task herself.*

In (8a), the subject (Peter) of the embedded clause can be restored easily because the referent has just been activated. However, one also needs the knowledge about the use of the matrix verb: intentions are usually with respect to one's own actions. If the intention relates to someone else's actions (i.e. Mary's doing the task), as in (8b), then the subject of the embedded clause should be explicit. If we take the verb *persuade*, the picture is reversed:

(9) a. *Mary persuaded Peter to do the task himself.*

b. *Mary persuaded Peter that she should do the task herself.*

The shorter form in (9a) is used because it is typical that the object of persuasion (i.e. Peter) performs the action specified by the second verb. There is no need to refer to Peter twice. The longer form in (9b) is used in less typical situations, when the action is not related to the person being persuaded. That is why the subject of the embedded clause is overt.

Another illustration of interaction between different types of information involves semantic relatedness of previous words to the target word, which increases its predictability and therefore its chances of being reduced. For example, Gregory et al. (1999) used Latent Semantic Analysis (Landauer and Dumais 1997) to obtain a word association score between the target and all of the words in the conversation prior to it. The scores ranged from -1 to 1. For instance, the target word *food* in a conversation about restaurants had semantic relatedness 0.45, and in a conversation about vacations, it had the score 0.27. They show that higher semantic relatedness decreases the duration of words ending in *-t* and *-d* in English. These words are more predictable, or primed, because of the previous context and the semantic and encyclopaedic knowledge it evokes.

1.1.3. Different perspectives on communicative efficiency

Efficiency can be seen from different perspectives. In the previous section, I focused mostly on the pragmatic aspects. Taking a cognitive perspective, one can say that the speaker can use less coding when the information is easily accessible in the hearer's mind. The crucial role is played here by the frequency and recency of the experience associated with the exemplars of relevant linguistic and non-linguistic categories (Pierrehumbert 2001). If the exemplars have been activated recently, they will have stronger representations. Therefore, the associated information will be more accessible. In addition, frequent exemplars have the higher resting level of activation than less frequent ones.

Moreover, some information may also be inherently easier to access than other (e.g. human referents vs. inanimate objects). This is known as salience. The total amount of activated information matters, as well. For instance, if there are several competing referents, each of them will be less accessible when there is only one referent (see Section 1.2.2).

In neurolinguistics, one speaks about activation and inhibition of biological neurons or nodes in a neural network. It is easier to activate a neuron if it has been primed. Priming happens due to transmission of signal in the neighbouring nodes and prepares the connected node for possible activation. In addition, priming is transmitted more rapidly if the links between the nodes are strong. This strength depends on the frequency with which a particular node has been primed and activated via a particular connection (MacKay 1987: 9–12).

Therefore, one can say the information conveyed by a linguistic unit is more accessible a) if this unit has a high frequency; b) if it co-occurs frequently with the currently activated units, and c) if it is located in the network near other activated units (e.g. due to semantic relatedness). Importantly, the speaker can spare the articulatory effort when the information is more accessible and increase the effort when it is less accessible. For example, Ariel (2001, 2008: Section 2.2; see also Section 1.2.2) shows that the degree of accessibility of mental representations plays an important role in the choice of anaphoric expressions. The higher the accessibility, the shorter the referring expression (see Section 1.2.2).

When speaking about the knowledge of the world and previous experience with language, an important psycholinguistic concept is cue validity. According to the classical definition, cue validity is the probability that a certain object belongs to a particular category given the cue (Rosch and Mervis 1975). It is designated as a conditional probability $\mathbb{P}$ (category|cue). For example, the probability that a creature with feathers (a cue) is a bird (the

category) is higher than the probability that a creature that can fly is a bird, since bats, numerous insects and some other species can fly, as well:

(10) $\mathbb{P}$ (bird|feathers) > $\mathbb{P}$ (bird|fly)

To compute the conditional probability $\mathbb{P}$ (bird|feathers), one needs to obtain the number of all creatures having feathers and the number of birds. We ignore here the difference between the scientific and lay person's categories, for the sake of simplicity. Next, one can divide the number of birds by the number of creatures with feathers. Obviously, it will be close to one (among a few exceptions will be extinct dinosaurs and eccentric fashionistas). One can also say that the conditional probability of a feathered creature being a bird is higher than the same creature being a dinosaur:

(11) $\mathbb{P}$ (bird|feathers) > $\mathbb{P}$ (dinosaurs|feathers)

Why would that be important? In the present study, cues can be different words, morphemes, constituents, constructions, etc., and categories are various interpretations, such as lexical meanings, syntactic functions of arguments or word identities in phonetic processing. For example, one can say that the conditional probability that an animate core argument is a subject is greater than the conditional probability that it is an object (see Chapter 6):

(12) $\mathbb{P}$ (Subject|Animate) > $\mathbb{P}$ (Object|Animate)

When processing language, the hearer has some linguistic cues – in the form of an ambiguous expression, or context, and tries to come to the intended interpretation (which linguistic unit is produced, or which formal role it plays, or which chunk of referents is meant). In the most general form, the relevant probabilistic information is $\mathbb{P}$ (Interpretation|Cue), or the conditional probability of the intended interpretation given some linguistic cue. When it is higher than an alternative interpretation given the same cue, the speaker can spare the effort:

(13) $\mathbb{P}$ (Interpretation$_i$|Cue) > $\mathbb{P}$ (Interpretation$_j$|Cue)

I will argue that this type of probability plays a key role in explaining the grammatical phenomena described in this study.

The exact type of probability depends on the phenomenon in question. In phonological categorization, for instance, the type of the probabilities that matters for production is $\mathbb{P}$ (Phonetic_Unit|Context), where the context is the surrounding units (words, syllables, phones). As will be shown in Section 1.2.6, there is ample evidence that more predictable units undergo reduction more often than less predictable ones.

For semantic disambiguation, the probability is $\mathbb{P}$ (Meaning|Form) or $\mathbb{P}$ (Semantic_Property|Form), when it comes to a specific property of the referent. For example, the word *nurse* is often used to represent female nurses because $\mathbb{P}$ (Female|*nurse*) > $\mathbb{P}$ (Male|*nurse*) in most language users' experience. See more information in Section 1.2.1.

For morphosyntactic processing, a relevant probability is $\mathbb{P}$ (Function|Form), when the hearer should predict a grammatical function or feature (e.g. object or subject, singular or plural) of a grammatical form or construction.

For interpretation of some constructions, one may need the probability $\mathbb{P}$ (Enriched_intepretation|Construction). Let us take possessive constructions as an illustration (cf. Levinson 2000: 146). For instance, *Mary's dress* normally means a dress that Mary owns or the one that she is wearing at the moment. The probability of this interpretation is higher than the probability that this is the dress that Mary created herself, although this interpretation can be more likely in some situations. Compare this with *Chanel's dress*.[2] Here, the expression is likely to represent a dress created by Coco Chanel.

Another relevant probability is $\mathbb{P}$ (Function|Context). For instance, as will be shown in Section 1.2.5, the clauses with the verbs frequently followed by a complement clause CC, i.e. with a high $\mathbb{P}$ (CC|Verb), are more frequently used without the complementizer *that*.

It seems very likely that the predictability is co-determined by multiple contextual cues simultaneously: lexical, syntactic, semantic, etc. There is evidence, for example, that this is the case in word phonetic reduction and in the use of the complementizer *that* (Jaeger and Buz 2017).

[2] See, for example, http://www.beniciamagazine.com/October-2011/Fashion-Noir/ (last access 25.11.2018).

In addition to cue validity, one also speaks about category validity, i.e. the probability that a object that belongs to a certain category has a certain feature (cue), ℙ (cue|category). For instance, the probability that a bird has feathers is higher than the probability that it can swim.

(14) ℙ (feathers|bird) > ℙ (swim|bird)

In linguistics, one can say that the typical features of a grammatical subject are topicality, animacy and givenness (see Chapter 6). These will be examples of category validity.

The validity of linguistic cues and categories can be approximated by conditional probabilities derived from corpora. My working hypothesis is that corpus-based probabilities are similar cross-linguistically when we speak about fundamental grammatical categories, such as different types of causation or typical agents and patients. This working hypothesis is supported by empirical evidence, which is presented in Chapters 3 and 6. Moreover, Greenberg (1966) showed that different languages display very similar relative frequencies of basic grammatical categories. Importantly, these frequencies do not represent the frequencies of occurrence of different referents or situations in the world. Instead, they represent the frequencies of what language users talk about. These frequencies are "a consequence of (joint) salience for the members of the relevant speech community" (Croft 2000: 76).

In many studies, the probabilities of linguistic features, meanings, forms and other categorical outcomes of a random variable are represented in a transformed form. Usually, a logarithm is taken. In information theory, the negative logarithm of a probability of a given categorical outcome represents its Shannon information content (MacKay 2003: 67), or surprisal. If the base of the logarithm is 2, the information is measured in bits. The role of information content is discussed in Part IV of this thesis.

To summarize, one can represent the correspondences between amount of effort and the pragmatic, cognitive, probabilistic and information-theoretic properties of the intended message as shown in Table 1.1.

Table 1.1. Correspondences between amount of signal and concepts from various fields

Discipline	**Small amount of effort**	**Large amount of effort**
Pragmatics	The intended information is present or easily inferable from the context, general knowledge and previous experience with language.	The intended information is neither present, nor easily inferable from the context, general knowledge or previous experience with language.
Cognitive science/psychology	The accessibility of the relevant mental representation is high.	The accessibility of the relevant mental representation is low.
Psychology of categorization	The cue validity of the available feature with regard to the intended category is high.	The cue validity of the available feature with regard to the intended category is low.
Probability theory	The probability of the intended interpetation is high.	The probability of the intended interpretation is low.
Information theory	The information content/surprisal of the intended message is low.	The information content/surprisal of the intended message is high.

A relevant question is then whether one should take the simple probability of an interpretation given some cue, or also consider the association between this interpretation and other cues. The latter would lead us to more complex mathematical measures. One of the options is pointwise mutual information (PMI). For example, Bouma (2016) finds an effect of PMI between the matrix predicate and the head of the infinitival complement with and without the particle *om* in Dutch. The higher the mutual attraction, the lower the chances of the longer form with *om*. Similarly, Gregory et al. (1999) find that mutual information of the target word with the next word increases the chances of tapping and deletion of final *-t* and *-d* and decreases the duration of a word with these final consonants in a corpus of spoken English. Other possible measures are delta *P*, a measure used in contingency learning (Ellis 2006), and collostructional strength (Stefanowitsch and Gries 2003 and later work). In this study, I use the simplest option in the form of conditional probabilities because this approach has been used successfully in numerous studies. Moreover, Gregory et al. (1999) demonstrated that mutual information produces very similar results to the conditional probability of the target word given the neighboring word.

In the rest of this chapter, I will examine a few examples of efficiency in language. I will start from the least conventionalized cases (conversational implicatures), and will gradually move on to more conventionalized and automatized, less conscious manifestations of efficiency. Finally, Section 1.3 provides a summary and a short discussion of the results.

1.2. Efficient asymmetries of the speaker's effort in language: an overview

1.2.1. Implicatures of Quantity and Manner

Grice (1975) postulated the Cooperative Principle and four types of maxims, which govern the communicative behaviour of language users. Two of these types are particularly relevant for the present discussion, namely, the maxims of Quantity and the maxims of Manner. The first Maxim of Quantity says, "Make your contribution as informative as required (for the current purposes of the exchange)", whereas the second one says, "Do not make your contribution more informative than is required" (Grice 1975: 45). This means that one should provide as much information as needed, not more and not less. The supermaxim of Manner, "Be perspicuous", is related to how something is said, not what is said, and includes several submaxims: "Avoid obscurity of expression", "Avoid ambiguity", "Be brief (avoid unnecessary prolixity)", which is particularly relevant for the present study, and "Be orderly" (Grice 1975: 46).

Let us begin with the maxims of Quantity. Horn (1984) re-formulated them as the Q-principle: "say as much as you can" (given R), and the R-principle: "say no more than you must" (given Q). The Q-principle can be also called the Principle of Sufficient Effort, and the R-principle is analogous to Zipf's Principle of Least Effort (1949).[3]

[3] According to Horn, the Q-principle represents Auditor's Economy, which corresponds to the Force of Diversification by Zipf (1949), whereas the R-principle represents Speaker's economy, or the Force of Unification. At the same time, the parallel is not perfect. Zipf (1949: 20ff) focuses primarily on the speaker's efforts of learning and maintaining a large vocabulary, and selecting the words for the message. The maximal economy is achieved if the speaker has only one word in the lexicon that can mean everything. This is a case of maximal paradigmatic economy. He does not say anything about syntagmatic economy, i.e. articulation effort (although his Law of Abbreviation implies it). In Horn's theory, both syntagmatic economy (e.g. the Avoid Pronoun principle) and paradigmatic economy (e.g. autohyponymy) are explained by the R-principle.

The most relevant for the purposes of the present study is the account proposed by Levinson (2000) because it involves the notion of typicality, which can be directly linked to the probabilistic account developed here. One of his fundamental principles is called the I-heuristic: "What is expressed simply is stereotypically exemplified" (2000: 37), which is related to Grice's second Maxim of Quantity (cf. Horn's R-principle). Consider the following example:

(15) John: "*I cut a finger*".

Under normal circumstances, this utterance communicates that the finger belongs to John, although this is not encoded in the utterance. This is an instance of generalized conversational implicatures, which can be triggered normally (in the absence of special circumstances) by the use of certain forms in an utterance (Grice 1975). When the finger belongs to someone else, a longer expression will be used (e.g. *I cut my brother's finger*). We are speaking about an implicature here because it still depends on the context. One can imagine a situation when the most natural interpretation could be that John cut someone else's finger. That would be the case, for instance, if John were a manicurist speaking to his colleague (Levinson 2000: 17). But one has to imagine very special circumstances if one's goal is to override the stereotypical interpretation.

Other examples are provided below, where an arrow stands for 'implicates':

(15) a. *I'll call Dad* (→ the speaker's dad) vs. *I'll call your Dad*.

b. *a nurse* (→ female nurse) vs. *a male nurse*, *a secretary* (→ female secretary) vs. *a male secretary*

c. *a bread knife* (→ a knife for cutting bread), *a steel knife* (→ a knife made of steel), *a kitchen knife* (→ a knife used in the kitchen) vs. *a knife made of ice* (since 'an ice knife' may be ambiguous)

From the efficiency perspective, the I-implicatures can be interpreted with the help of the Low-Cost Heuristic. In the examples above, there is a default interpretation that is easy to access and that involves some typical relationship or scenario. Probabilistically speaking, the

intended interpretation, which belongs to the set of alternatives {Interpretation$_1$, Interpretation$_2$, … Interpetation$_i$} has the highest $\mathbb{P}$ (Interpetation$_i$|Cue). For example, female nurses constitute the overwhelming majority of the entire population of nurses, which is known both to the hearer and the speaker. This is why this interpretation is highly probable, which allows for sparing the effort.

Levinson also formulated the Q-heuristic: "What isn't said, isn't" (Levinson 2000: 35), which closely corresponds to Horn's Q-principle and Grice's first Maxim of Quantity "Make your contribution as informative as required". This means, in other words, that the lack of extra information or a stronger statement is informative. The Q-heuristic helps the speaker to spare effort because it enables one to omit additional restrictions. In the examples below, the (a) version, where this principle is exploited, is shorter than the (b) version, which the speaker would need to produce if language users could not rely on the Q-heuristic.

(16) a. *Her dress was red.*

(→ not red and blue or red and any other colour)

b. *Her dress was red, and only red.*

(17) a. *I've eaten some chocolates.*

(→ There're still some left).

b. *I've eaten some chocolates, but not all.*

The knowledge required for inferring these implicatures is the knowledge of the existing alternative expressions, such as 'red and X' for (16a), where X stands for any other colour, and 'all' for (17a). The hearer derives the implicatures because he or she understands that the more informative expressions are not selected. In this sense, it is metalinguistic (Levinson 2000: 40–41). The Q-heuristic also plays an important role in the grammaticalization of some asymmetries, as will be shown below.[4]

[4] Levinson describes only the Q- and M-implicatures as metalinguistic, i.e. rely on the knowledge of alternative expressions. At the same time, McCawley (1978: 245) writes, "[w]hat is conversationally implicated by an utterance depends not only on the utterance, but on what other utterances the speaker could have produced but did not." It is an interesting theoretical question if the I-heuristic, which allows one to use the simple expression to

Finally, Levinson formulates the M-heuristic, which is related to Grice's maxims of Manner, especially to his first maxim "avoid obscurity of expression" and the third maxim "avoid prolixity". In the short version, it is expressed as follows: "What's said in an abnormal way isn't normal" (Levinson 2000: 38). There is also a longer version, which may be somewhat easier to understand:

(18) The M-heuristic

Speaker's maxim: Indicate an abnormal, nonstereotypical situation by using marked expressions that contrast with those you would use to describe the corresponding normal, stereotypical situation.

Recipient's corollary: What is said in an abnormal way indicates an abnormal situation, or marked messages indicated marked situations (…) (Levinson 2000: 136)

The notion of markedness used here is a generalization of the Prague School concept: marked forms are more morphologically complex and less lexicalized than the corresponding unmarked forms; they are also more prolix or periphrastic, less frequent or usual, and less neutral in register. One can see that these are very diverse features (cf. Haspelmath 2006). As far as the meaning is concerned, marked forms imply some additional meaning or connotation absent from the corresponding unmarked forms (Levinson 2000: 137). Consider some examples, where both I- and M-implicatures are present:

(19) a. *Sue smiled.*

(I-implicature → Sue produced a nice happy expression.)

b. *The corners of Sue's lips turned slightly upward.*

(M-implicature → Sue produced a smirk or grimace).

Another pair of examples illustrates the contrast between the forms of the type *go to school* and *go to the school*, which involves highly conventionalized inferences:

refer to stereotypical interpretations does not imply a comparison with some alternative form. After all, "simple" is also defined with regard to something "complex".

(20) a. *She went to school/church/university/bed/hospital/sea/town...*

(Conventionalized I-implicatures → She went to do the stereotypical activity association with this location.)

b. *She went to the school/church/university/bed/hospital/sea/town...*

(M-implicatures → She went to the place, but not necessarily to do the associated stereotypical activity).

Other examples of a contrast between more typical and less typical expressions, which involve these implicatures of stereotypical and non-stereotypical situations, include litotes (e.g. *happy* vs. *not unhappy*, where the latter implicates 'less than happy'), lexicalized forms of periphrasis (*pink* vs. *pale read*, i.e. 'an untypical pink'), nominal compounds (e.g. *a matchbox* vs. *a box for matches*, i.e. 'an untypical one'), some prepositions (e.g. *on the table* vs. *on top of the table*). Another well-known case is the contrast between lexical and analytic causatives, as in the example below:

(21) a. *John stopped the car.*

(I-implicature → in the usual way, i.e. by putting his foot on the brake pedal).

b. *John got the car to stop.*

(M-implicature → in an unusual way, e.g. by using the emergency brake or crashing into a lamppost).

For the M-implicature to be derived, it is crucial to have a conventionalized, typical expression. If a periphrastic causative does not have a corresponding lexical causative, this implicature does not emerge. For example, the expression 'make someone laugh' does not generate such an implicature (McCawley 1978: 250). Causatives are discussed in detail in Part II of this thesis.

M-implicatures can be regarded as inferences based on the High-Cost Heuristic. When confronted with a costly form, the hearer chooses the interpretation that changes his or her current mental representations the most. This is the interpretation with a low probability $\mathbb{P}$ (Interpetation_i) out of set of alternative interpretations. In the overwhelming majority of the

examples provided by Grice (1975), Levinson (2000), Huang (2007) and in other works on implicatures, the marked form is longer than the unmarked form. Therefore, we observe efficient asymmetries in the speaker's effort. Exceptions can be found in some pairs of cross-register doublets, as in the following example from Levinson (2000: 139):

(22) a. *He was reading a **book**.*

(I-communicates → He was reading an ordinary book).

b. *He was reading a **tome**.*

(M-communicates → He was reading some massive, weighty volume).

Another relevant pair is *horse* and *steed*, as in *a brave knight and his noble steed* (Levinson 2000: 139), although usually such words also have length differences (e.g. *gift – donation*, *house – mansion*/*residence*). In principle, the stylistic differences can be also interpreted in terms of efficiency because the stylistically marked words may be more difficult to extract from the memory. This type of effort is outside the scope of the present study, however, as was mentioned in Section 1.1.1.

Unlike Levinson's heuristics and implicatures, the efficiency-based approach is directly linked with measurable costs and does not involve such fuzzy concepts, such as simplicity (or complexity), stereotypicality or markedness. One cannot exclude that some other probabilistic matching happens between interpretations and cues, which is based on purely semantic or stylistic information.

1.2.2. Ellipsis and anaphora

This subsection provides an account of efficient asymmetries that involve referential expressions. One can formulate a hierarchy of explicitness of such expressions (Ariel 1990, 2001; Arnold 2010, see also Givón 1983, 2017), as shown in (23):

(23) Hierarchy of explicitness:[5]

Most explicit

Semantically rich expressions (*the teacher who is late*)

Shorter nominal expressions (*Mary*, *the teacher*)

Pronouns (*she*)

Zeros

Least explicit

This variation is constrained by the degree of accessibility and activation of the mental representations of the referent in the memory.[6] Highly accessible representations are represented by shorter forms than non-activated and less accessible ones. The level of accessibility and activation depends on several factors, which are mentioned below (based on Ariel 1990, 2001, 2008 and Arnold 2010) and which can interact in complex ways (see Ariel 2001).

1. Previous mention in discourse. The referents that have been introduced in discourse have more activated representations than the referents that have not been mentioned. This is why full nouns are typically used to introduce new referents, while pronouns are usually reserved for the referents already introduced in discourse.

2. Density of mention. High density of mention of the referents in previous discourse means higher activation of their mental representations and therefore increases the chances of pronominal expressions (Levy and McNeill 1992).

3. Recency in discourse. The more recent the mention of the referent, the more accessible the mental representation is. For example, Arnold (2010) provides data to show that the chances of the pronominal reference decrease with the distance from the last mention of the referent (measured in clauses).

4. Paragraphs and episode boundaries, which decrease accessibility.

5. Topicality. Topical referents are more accessible and therefore expressed by less explicit forms than non-topical ones.

[5] See a more detailed accessibility marking scale in Ariel (1990: 73).

[6] The notions *activation* and *accessibility* are very similar. The difference is in the potential character of accessibility. If a representation is highly accessible, it is either activated in discourse, or it can be easy to access due to high frequency, salience, relatedness to the activated information, etc. If a representation has low accessibility, it is normally not activated. For example, the name *Merkel* is probably neither activated, nor easily accessible for the reader of this chapter. I thank Mira Ariel (p.c.) for this explanation. All possible misconceptions are solely mine.

6. Syntactic function of the referent. The referent is more accessible when it has been mentioned previously in the same syntactic function. This parallelism makes it easier for the hearer to identify the referent. This explains why reduced forms are more likely if the referring expression and the previous mention of the referent are in the same syntactic position (Levy and McNeill 1992). Consider an example:

(24) *Mary invited Jane to the conference.*

a. *She asked Jane to present her new research on metaphors.*

b. *Jane asked her to tell more about the event.*

According to Arnold (2010), the preference for a pronoun that refers to Mary should be stronger in (24a), *She asked Jane*, than in (24b), *Jane asked her*.

7. The presence of competing referents in the context, even when there is no direct need for disambiguation. For example, Arnold and Griffin (2007) performed an experiment with cartoons, on which the subjects could see either one character or two different-gender characters. The first line of the story was, for example, *Daisy went for a boat ride {with Mickey} on the lake*. Next, the second picture was shown, which displayed one character doing something (e.g. Daisy rowing away). The second character was either present or absent. Participants generated another line for the story (e.g. *Daisy left Mickey behind*; or *She rowed into the sunset*). Interestingly, pronouns were more common in the one-character than two-character stories, despite the obvious fact that there was no risk of confusability, since the characters had different genders. Arnold and Griffin (2007) explain this finding by greater cognitive load in the situation with two characters and the differences in the degree of activation of the character in question in two different conditions. The preference for the more explicit form in the situations with more than one characters results from competition between the entities in the speaker's mental model, which results in a lower level of activation for each entity.

8. Probabilistic expectations based on the previous experience with language. For example, Arnold (2001) shows that goals of the verbs of transfer, e.g. *give/send/bring to Mary*, are more frequently referred to by shorter pronominal forms than sources, e.g. *accept/get/borrow from Mary*. Language users also refer more to goal referents than to source referents in discourse, as Arnold's story-telling experiment and corpus analyses revealed:

> This indicates that from the comprehender's point of view, it is more probable that the speaker will refer to the goal referent than the source referent (Arnold 2001: 228).

The higher probability of goals means their higher accessibility and therefore shorter expression. Notably, this frequency asymmetry is also observed for inanimate goals and sources (e.g. *to London/the market/a village* is more frequent than *from London/the market/a village*), which accounts for the cross-linguistic differences in the length of marking of goals and sources (Michaelis 2017). Another case is the use of reflexive pronouns. Although their referents are highly accessible due to coreferentiality (e.g. *She adores herself*), the reflexive form is long because the verb *adore* usually takes an object that is different from the subject. See a more detailed discussion of reflexive pronouns in Chapter 2, Section 2.3.

9. Degree of interaction between the speaker and the hearer. Wilkes-Gibbs and Clark (1992) show that descriptive nominal expressions tend to become shorter when the speaker and the hearer develop and expand their common ground – the information they believe they share. Interestingly, even the subtle differences in the status of the hearer, e.g. from being able to overhear or watch the previous interactions to being totally new to the scene, determine the speed and amount of coding in the subsequent interaction.

To summarize, the more accessible or activated a referent is due to the immediate context, interaction settings or previous experience with language, the less costly the referential expression. As pointed out by Ariel (1990), the choice of the specific form helps listeners to identify the location of the referent in their mental representation. The use of the shorter variant signals that the referent is accessible, in compliance with the Low-Cost Heuristic. Longer forms signal low accessibility, according to the High-Cost Heuristic. There is experimental evidence (Gordon and Chan 1995) showing that reading times increase when a repeated long referential expression is used for a highly accessible referent. Consider the following example.

(25) *Susan decided to give Fred a hamster. She/Susan was questioned at length by Fred about what to feed it.*

Self-paced reading times were longer when the referential expression was *Susan* than when the pronoun *she* was used. This can be interpreted by the Principle of Communicative Efficiency and the High-Cost Heuristic. When the short form *she* is not used, which would normally represent the highly accessible referent, the proper name is supposed to introduce a new

referent, but this is not the case. The readers get confused, which slows down their reading times.

Let us now focus on zero anaphora. Different languages have different rules with regard to which constituents can or should be omitted in discourse. Yet, there are a few general tendencies.

First, given and topical referents, which are restorable from previous discourse, are more frequently omitted than new and focal ones. As defined by Lambrecht (1994: 218), the focus relation relates "the pragmatically non-recoverable to the recoverable component of a proposition and thereby creates a new state of information in the mind of the addressee." Focal elements are thus not predictable. Therefore, they need to be overtly expressed. In many languages, including Chinese, Japanese, Korean, Hindi, Hungarian and Lao any given, non-focal argument can be omitted (Goldberg 2005), while in English this is only possible with some verbs (Fillmore 1986; see also below).

Second, agreement is a strong factor that enables the omission of arguments. In a cross-linguistic survey by Gilligan (1987: Section 3.4), languages where the verb agrees with a specific argument nearly always allow for omission of that argument. An example is subject agreement in Pashto:

(26) (Pashto: Indo-European, Huang 2007: 142)

Ø	*mana*	*xwr-əm.*
	apple	eat-1.M.SG

'(I) ate the apple.'

Languages without agreement allow pro-drop less frequently. An example of a language without agreement that allows omission of the subject is Mandarin Chinese:

(27) (Mandarin Chinese: Sino-Tibetan, Huang 2007: 143)

Ø	*qu*	*guo*	*Beijing*	*ma?*
	go	EXP	Beijing	Q

'Have (e.g. you) been to Beijing?'

As for the languages without agreement, imperative subjects can be omitted in nearly all languages in Gilligan's sample, followed by thematic subjects and direct objects. The explanation is as follows. Imperative subjects are restorable from the context, and therefore

redundant. Languages that allow for omission of subjects are more frequent than languages than license the omission of objects because subjects are more thematic, given and therefore recoverable from the context than objects (Lambrecht 1994: 262; see also Chapter 6). Various other constituents (indirect objects, possessive pronouns and adpositional objects) are very rarely omittable.

Similarly, Siewierska (2004: 43–46) observes that in the vast majority of languages that she examined (89%), more phonologically reduced and/or dependent person markers are used for arguments higher on the argument prominence hierarchy than those lower on the hierarchy. We observe a correlation between the two scales within a language, as shown in (28):

(28) a. Scale of phonological reduction/dependence of person markers:
Zero > Bound > Clitic > Weak

b. Scale of argument prominence:
Subject > Direct object/Theme > Indirect object > Oblique

For example, if pronominal direct objects can be omitted, then subjects can do it, too. Siewierska explains this correlation by the differences in accessibility of typical arguments in different syntactic positions:

> since dependent person markers involve less encoding than independent ones, the expectation is that they should be characteristic of syntactic funcitons which tend to realize highly accessible referents (Siewierska 2004: 46).

As far as zero objects are concerned, English represents an interesting case. Although it generally does not allow for object omission, there are exceptions, which have been regarded as lexically specific. In particular, the object can be omitted if the verb is used in a particular sense, but cannot be omitted if the verb is used in some other sense (Fillmore 1986):

- *He won* can be said when the person in question won the election/game/race, but not if he won the gold medal or the first prize.
- *She lost*, again, can be said if she lost some competition, but not if she lost her wallet or keys.
- *We've already eaten* can be said in the situation when we have had a meal, but not when we have eaten something specific.

- *I forgot*, e.g. to fix something, but not if the speaker forgot the keys.

Importantly, the object cannot be omitted even if it is previously mentioned or clear from the context (e.g. *Where're the keys? I forgot *(them)*). In these cases, one might be tempted to conclude that abstract entities and events are more commonly omitted than concrete physical objects. However, this is not true. If we take verbs of motion with a specific destination or point of departure, the object can be omitted if it is a physical location, and cannot be omitted if it is abstract and metaphorical:

- *She was approaching*, e.g. the speaker, the town, but not if she approached the solution.
- *She arrived*, e.g. at the summit, but not if she arrived at the answer.

According to the efficiency theory, the elliptical use is based on conventionalized inferences explained by the Low-Cost Heuristic. This is obvious in the case of motion verbs, where the interpretation of the physical motion (approaching a location and arriving at a certain place) is the stereotypical interpretation, and the metaphorical extensions (approaching a solution or arriving at an answer) are less probable. In the other cases, the interpretation that allows the ellipsis is on average more probable than the interpretation that does not.

Consider the verb *win*. In a random sample of 100 examples of the verb from COCA (Davies 2008–), 90 were instances of *win* as a verb followed by a direct object or used without any complement. The majority of these instances (61) were about winning some competition (elections, sports, social conflicts, etc.), as in (29a), and only 26 were about winning something for oneself (a prize, confidence, support, more rights, a Senate seat, etc.) as a result of a competition or after applying some effort, as in (29b). In three instances, it was difficult to classify the examples semantically.

(29) a. *Everything counts, everything has to be perfect for you to win the game.* (COCA, News, Denver, 2005)

b. *Guess what? You can win a cruise at home as well.* (COCA, Spoken, NBC: Today Show, 2017)

The meaning of winning some competition is more common and therefore more restorable from context than winning some objects or other benefits.

Next, I zoomed in on the instances with explicit objects. The majority of the contexts with winning something for oneself had an indefinite object (e.g. win a cruise, a Senate seat or a legislative approval). There were 20 instances out of the total of 26. The indefinite objects were not given in the previous context and not restorable from the context. In contrast, only 8 contexts of winning a competition had indefinite objects (e.g. win a tournament), against 21 instances with the definite object (e.g. win the presidential election). These are given and accessible referents. Usually, the information about winning a competition is mentioned previously or clear from the context. Consider two examples:

(30) a. *If this is a big chess game, did you win or lose?* (COCA, Spoken, CBS_48Hours, 2007)

b. *How are you doing in the polls? How are you going to win in New Hampshire?* (COCA, Spoken, CBS_Early, 1999)

Although more research is obviously needed, one can hypothesize that both the probability of the intended interpretation and typical accessibility of the referent in the context play a role. If the given sense is the most probable interpretation and its referents are usually accessible, it is likely that the intransitive use of the verb in this specific sense will be lexicalized.

The reasons for omission can also be due to reasons different from economy. For example, taboo objects, such as bodily emissions (spit, piss) are usually omitted, although they can also be restored, as in the examples below (from Goldberg 2005):

(31) a. *Pat sneezed (mucus) onto the computer screen.*

b. *The hopeful man ejaculated (his sperm) into the petri dish.*

c. *Pat vomited (her lunch) into the sink.*

These are cases of the so-called Implicit Theme Construction (Goldberg 2005). The object can be easily restored by the hearer based on general knowledge, although this may not be the most pleasant information. The object can also be irrelevant, or deprofiled (Goldberg 2005):

(32) a. *Tigers only kill at night.*

b. *She gave and gave, and he took and took.*

c. *She picked up the carving knife and began to chop.*

These are instances of the so-called Deprofiled Object Construction (Goldberg 2005). The objects are omitted because the emphasis is on the action itself, rather than on the object. This agrees with Givón's (2017: 3) principle of cataphoric zeros: "Unimportant information need not be mentioned". Another illustration of this principle is omission of agents in passive constructions, as in the following example:

(33) *An English tourist was robbed of his Rolex watch (by Ø).*

Goldberg also explains conventionalized habitual uses like *She drinks/smokes/writes* as a result of such deprofiling of the object and subsequent conventionalization and lexicalization of the intransitive use. A similar perspective is taken by Givón (2017: 198). Indeed, what is important is that the person in question is an alcoholic, a smoker or a writer. This information can allow the hearer to make some useful inferences. Although this is by all means correct, it seems that predictability of the object also plays a role. In particular, Huang (2007: 48-49) classifies the uses like *John doesn't drink* in the sense 'John doesn't drink alcohol' as cases of lexical narrowing based on an I-implicature (see the previous section). Alcohol is available as a probable option if one speaks about a habit (cf. the Present Simple form). This is an implicature, which can be cancelled if, for instance, John is a patient in a hospital, and he is not able to drink any liquids. Similarly, one can say *John smokes*, implying that he smokes tobacco (cigarettes, cigars or a pipe). Smoking other substances would be a less likely interpretation. One might wonder, however, if the implicature 'John smokes tobacco' will be inferred, for example, in a Rastafarian community.

My hypothesis is that the omission of the object is more likely when the action represented by the verb is stereotypical and the non-overt object or agent is predictable. Compare the examples in (32) with the following sentences:

(34) a. ?*Hedgehogs only kill at night.*
b. ?*Stalin gave and gave, and Mother Theresa took and took.*
c. ?*She picked up her scalpel and began to chop.*

The omission is less felicitous because the events are less probable and the objects are less predictable than in (32).

1.2.3. Grammatical coding asymmetries and splits

Grammatical coding asymmetries consist of members of contrasting grammatical categories, which are expressed by markers of different length (Greenberg 1966; Haspelmath Forthcoming-a). Below are some examples:

(35) a. single vs. plural nouns (e.g. *book – books*)

b. positive vs. comparative and superlative degrees of comparison of adjectives, e.g. *nice – nicer – the nicest*)

c. cardinal vs. ordinal numerals (e.g. *ten – tenth*)

d. indicative vs. subjunctive (e.g. *I go – I would go*)

e. active vs. passive verb forms (*I called X – I was called by X*).

The first member in these pairs is shorter than the second (and third) one. This is a robust cross-linguistic tendency. These phenomena became important in structuralist linguistics after Roman Jakobson (1932 [1971]) extended the notion of markedness from phonology to grammar. In such pairs as the ones above, the shorter member is considered the unmarked one, whereas the longer one is referred to as marked. The unmarked member appears in neutralization contexts. For instance, in the opposition between singular and plural, as in *cat – cats*, the singular form appears in neutralization contexts (e.g. the generic use in *The cat is a night wanderer*) and is unmarked. With time, the notion of markedness has become so broad, being understood as non-naturalness, cognitive complexity, language-specific or cross-linguistic rarity, poor cross-linguistic attestation, etc., that it can hardly be considered a useful scientific notion (see Haspelmath 2006).

In fact, as argued in Haspelmath (2006), all these phenomena can be reduced to frequency effects, which provide a more parsimonious explanation and a causal mechanism for a large number of associated empirical tendencies. One of such tendencies is the fact that the unmarked members in the examples above usually have higher inflectional and syntagmatic potential than the marked members, and, importantly for the present discussion, they are expressed by zero or shorter forms (Greenberg 1996; Croft 2003: Chapter 4). These tendencies

can be explained by the fact that the unmarked members are more frequent than the marked ones (some corpus evidence was provided already in Greenberg 1966). According to Haspelmath, the unmarked categories are more frequent, and therefore, their meaning is more predictable:

> "Speakers can afford to use short shapes or zero coding for predictable meanings, but they have to make a greater coding effort for unpredictable meaning" (Haspelmath Forthcoming-a: 19).

Using the notation introduced above, the probability of a singular interpretation of a noun is greater than that of a plural interpretation: $\mathbb{P}$ (Singular|Noun) > $\mathbb{P}$ (Plural|Noun), which represents a special case of $\mathbb{P}$ (Interpretation|Cue) where the hearer needs to infer the information about the number of a noun.

This allows language users to spare effort when speaking about singular referents, based on the Low-Cost Heuristic. When the longer construction designating the rare category becomes sufficiently frequent and obligatory for expressing the meaning associated with it (as the *-s* plural in English), Levinson's Q-heuristic comes into play, "What isn't said, isn't" (cf. García and van Putte 1989). The shorter form will become associated exclusively with the more frequent category, e.g. singular. According to Bybee (1994), this mechanism explains the grammaticalization of zero (e.g. *book-Ø – books*). If the overt plural is always used when plural is intended, then by inference, the unmarked noun comes to be interpreted only as singular. The meaningful zero is "parasitic" on the longer form (García and van Putte 1989).

This does not happen to all asymmetries. In some cases, the Q-implicatures of this sort are not inferred. For example, consider the contrast between the simple past and habitual past constructions in English, e.g. *walked* vs. *used to walk*. The simple past tense in English can represent non-habitual, perfective events in the past (*On that day, she walked in the park, fed pigeons and went home.*), but it can also represent habitual events (*She walked in the park every day*). As Bybee (1994: 239) argues, if the habitual past constructions (e.g. *She used to walk in the park*) had become obligatory, the aspectual meaning of the simple past, in the absence of *used to,* would be restricted to non-habitual meanings.

The examples in (35) illustrated general markedness, where the distinction is applied to all members of the grammatical class (verbs, nouns, numerals, adjectives, etc.) across the board.

Local markedness, in contrast, displays a markedness reversal for some members of these classes. Tiersma (1982) discusses such exceptions in the paradigm levelling in Frisian and some other languages. Markedness theory predicts that the leveling of paradigmatic alternation will favor the unmarked form. However, as some nouns in Frisian undergo change, the originally "marked" plural form becomes the basis for the singular form, rather than the "unmarked" singular. For example, *goes/gwozzen* 'goose/geeze' becomes *gwos/gwozzen*. Thus, the plural stem can be seen as unmarked. Tiersma shows that this markedness reversal happens to those nouns that are frequently used in the plural ('arm', 'goose', 'horn', 'stocking', etc.).

In some cases, the frequency effects can be even stronger and trigger a reversal in the marking itself. There are a few languages, for example, that can have both overt plural marking (e.g. *day – days*) and overt singular marking (e.g. Welsh *pys-en* 'pea' - *pys* 'peas'), depending on the noun. Haspelmath and Karjus (2017) demonstrate that "individualist nouns" (those that tend to occur with uniplex meaning, e.g. *day*) have overt plural marking, while "gregarious" nouns (those that are usually associated with multiplex meaning, e.g. *pea*) have overt singular marking. Other examples of the latter are the names of some fruits and vegetables, e.g. Russian *kartofel'* 'potatoes (mass noun)' – *kartofelina* 'potato'; small animals, e.g. Welsh *adar* 'birds/flock of birds' - *aderyn* 'bird'; and body parts, e.g. Cushitic *farró* 'fingers'- *farri-t* 'finger'. Corpus data from different languages demonstrate that the nouns that tend to have overt singular cross-linguistically are also used predominantly in the multiplex sense.

In these cases of markedness reversal, there is a difference between the general tendency for most nouns, which have higher conditional probabilities $\mathbb{P}$ (Singular|Noun) than $\mathbb{P}$ (Plural|Noun), and the group of gregarious nouns, for which $\mathbb{P}$ (Singular|Noun) is lower than $\mathbb{P}$ (Plural|Noun). This is why it is more efficient to mark the singular of gregarious nouns. At the same time, there is a strong competing factor, namely, the systemic pressure, which explains why such efficient form-meaning mappings are not very frequent cross-linguistically. For example, English individualist and gregarious nouns behave similarly, e.g. *day – days*, *pea – peas*, *potato – potatoes*, *bee – bees*, *eye – eyes*.

Another relevant case is coding splits. A famous example is differential argument marking. If there is a coding split, prominent (e.g. animate, definite) objects, which are less typical, tend to be formally marked, while less prominent (inanimate and indefinite) are usually unmarked. More information is provided in Chapter 6. Coding splits are also observed in locative marking. If there is variation in locative marking in a certain language, then place names will be unmarked, inanimates will be either unmarked or marked, and animates will tend

to be marked. Again, the explanation is that place names represent typical locations, while animates are untypical locations. Another example is adnominal possessive constructions, e.g. *John's house* (Haspelmath 2017). In some languages, different possessive constructions are used, depending on whether possession is alienable or inalienable. For example, in Abun, a West Papuan language, there is a following contrast:

(36) (Abun: West Papuan, Berry and Berry 1999: 77-82, cited from Haspelmath 2017: 194)

a. alienable possession

ji *bi* *nggwe*

I GEN garden

'my garden'

b. inalienable possession

ji *syim*

I arm

'my arm'

This example illustrates a cross-linguistic tendency for inalienable possession constructions, as in (36b), to have shorter coding than alienable possession constructions, as in (36a). Haspelmath's corpus data demonstrate that objects that are usually inalienable (kinship terms, body parts) more frequently occur in the possessive constructions (e.g. my hand, his sister) than alienable possessed entities, e.g. a house, a garden or a knife. More formally speaking, $\mathbb{P}$ (Possessed_object|*hand*) > $\mathbb{P}$ (Possessed_object|*house*). From this follows that the nouns that are more frequently mentioned as possessed objects receive less formal marking than the ones that are less frequently mentioned as such. The same pragmatic mechanism based on the Principle of Communicative Efficiency is responsible for this variation. More details about the diachronic development of such patterns follow in Chapter 2.

One should also mention here multifactorial grammatical alternations. In such situations, the coding asymmetry is associated with more or less likely frames or scenarios, rather than with the more or less probable grammatical functions of one word. Consider the English dative alternation between the double-object dative (e.g. *Mary gives her colleague the*

memory stick) and the prepositional dative (e.g. *Mary gives the memory stick to her colleague*). The two constructions have different word ordes, namely, Recipient + Theme in the double object construction and Theme + Recipient in the prepositional dative (although there can be exceptions, cf. Hawkins 1994: 214). There is substantial evidence that language users switch between the constructions in order to manage the flow of information and optimize processing. In particular, the double object construction is preferred when the recipient is given, short and pronominal, and the theme is new, long, and nominal, whereas the prepositional dative is preferred in the reverse situations (Hawkins 1994: 212–214; Goldberg 1995: 91ff, Bresnan et al. 2007). Moreover, the prepositional dative construction is a metaphorical extension of the caused motion construction (cf. *I sent the letter to my parents* vs. *I sent the letter to his old address*) (Goldberg 1995: Chapters 5–7). While the double object construction means 'X causes Y to receive Z', the *to*-dative means 'X causes Z to move to Y' (*Ibid*). This semantic difference is also supported by the distinctive collexeme analysis in Gries and Stefanowitsch (2004).

Yet, the constructions differ not only with regard to the order of their constituents, but, crucially, also in the amount of formal coding. Haspelmath (Forthcoming-b) argues that the shorter variant in alternations is normally used when the referential prominence of arguments corresponds to their roles, while the longer variant is used when there is some deviation from such canonical relationships. In particular, a typical recipient – i.e. animate, given, definite, pronominal – in a ditransitive (dative) construction is more prominent referentially than a typical theme, which is normally inanimate, new, indefinite and nominal. Although Haspelmath does not spell this out, one can see immediately that the results of Bresnan et al. (2007) support this interpretation. Their model shows that the double object construction is preferred when the recipient is animate, definite, given and pronominal, whereas the theme is non-given, non-pronominal and indefinite and has a low rank on the animacy hierarchy. The *to*-dative is particularly favoured when the arguments have the opposite characteristics. One can then regard the form-meaning mapping in the dative alternation as a manifestation of efficient communication because the construction with more formal coding expresses the less probable configuration of participants.

1.2.4. The use and omission of function words and grammatical morphemes

In Relevance Theory (Sperber and Wilson 1986/1995), an important distinction is made between conceptual (representational) and procedural (computational) information. The former is information about concepts or conceptual representations to be manipulated, and the latter is information about how to manipulate them (e.g. Blakemore 1987; Wilson and Sperber 1993). For instance, the conjunction *so* plays such a role:

(37) *Peter's got a PhD, so he'll be able to fill in this form.*

Such connectors indicate the type of inference process that the hearer is expected to go through. In (37), the connector *so* indicates that the second clause should be interpreted as a conclusion. As Blakemore points out, such expressions contribute to relevance by guiding the hearer towards the intended cognitive effects. In Grice (1975), these inferences are called conventional implicatures. The connector *so* conventionally implicates, according to Grice, that the first clause explains the second. Regardless of the differences between the interpretations proposed by different authors, the common idea is that the speaker guides the hearer's inferential process by providing a cue, which helps the hearer to make inferences based on the propositions in the first and second clause. Other examples of such cues are the connectors *but, and, therefore, on the other hand* and *after all*.

Importantly, connectors can be omitted when the intended inference is easy to make. For example, Blumenthal-Dramé and Kortmann (2017) investigate the use and absence of causal and concessive adverbial connectors, as in the following examples from their study:

(38) a. *John didn't read the essay questions properly and **therefore** failed the exam last January.*

b. *John didn't read the essay questions properly and failed the exam last January.*

c. *Peter studied a lot and **still** failed the exam last January.*

d. *Peter studied a lot and failed the exam last January.*

They argue that there is a general tendency for concessive relations to be marked overtly, as in (38c), while causal relations are more often left implicit, as in (38b). The reason is that concessive relationships are more complex. As a result, implicit concessivity is more disruptive to discourse processing than implicit causality. Therefore, concessive connectors provide more cognitive benefits than causal ones.

Taking the efficiency perspective, one can say that the pragmatic basis for omission of the connectors is the Low-Cost Heuristic, while the High-Cost Heuristic is responsible for their use. Causal relations are more frequent than concessive ones in discourse. This claim is supported by the counts from the Penn Discourse Treebank obtained by Asr and Demberg (2012), who also show that causal relations are much more often implicit (62% to 69%, depending on the order of cause and effect) than concessive relations (8% to 19%). Therefore, the omission of a connector signals that the more probable (causal) meaning is intended. One also cannot exclude that humans have a cognitive bias towards establishing causal links between events, even these events they are not causally related, e.g. the logical fallacy *post hoc ergo propter hoc.*[7] Whether causal relations enjoy a special status in cognition is an open question.

The same reasoning can be applied to other clause-linking elements, such as complementizers and relativizers. They help the hearer to identify the syntactic and semantic role of elements in discourse. If the use of such function words is optional in a language, the speaker can use them in the situations when the function of the clause they introduce is more difficult to identify, and omit them when the function is more obvious. An important role is played here by their heads – i.e. nominal phrases and predicates. If they are often followed by a clause, the interpretation is easier to activate, which allows the speaker to omit the function word. For instance, as shown by Wasow et al. (2011), the relativizer *that* or *which* in non-subject relative clauses is more likely to be omitted when the nominal phrase is definite (e.g. *the colleague I'm replacing*) or contains a superlative adjective (e.g. *the most interesting subject I've ever studied*) because such nominal phrases are more commonly followed by a relative clause than indefinite nominal phrases (e.g. *a secret that I don't want to tell anyone*).

[7] This fallacy is observed when a causal relationship is established on the basis of temporal succession, e.g. the rain started because the shaman had performed his rituals well; the child was diagnosed with autism because it had been vaccinated; a woman cannot have children because she had had an abortion, etc.

Similar results have been obtained for the complementizer *that* after different predicates, which can be used with complement clauses (Jaeger 2010):

(39) a. *I think (that) alternatives exist.*

b. *I'll show ?(that) alternatives exist.*

Jaeger finds that the odds of *that* are higher with matrix verbs that are rarely followed by a complement clause (e.g. *teach*, *see*, *show*), and lower when the matrix verb is frequently followed by a complement clause (*think*, *guess*, *suppose*, etc.). Thus, the use of *that* in (39a) is less likely than in (39b).

This variation has been explained by the Universal Information Density hypothesis, which predicts that speakers aim to transmit information uniformly close to, but not exceeding, the channel capacity (Levy and Jaeger 2007). Adding extra markers in more informative contexts helps to keep the information flow even and uniform, avoiding peaks and canyons. Mentioning the complementizer *that* at the onset of a complement clause distributes the same amount of information over one more word, thereby lowering information density. This idea is also similar to smooth signal redundancy hypothesis in phonology (Aylett and Turk 2004). In fact, it goes back to August and Gertraud Fenk (1980: 402):

> Von einem Verständigungssystem, welches die Übermittlung von Nachrichten ohne Verluste erlauben soll, ist daher nicht nur ein *durchschnittliches Redundanzniveau* [emphasis by the authors - NL] zu fordern, welches die Kurzspeicherkapazität nicht übersteigt, sondern auch, dass sich die Information möglichst gleichmäßig auf kleine Zeitabschnitte verteilt.

There are several problems with this account. First, there is no clear reason of why it is advantageous for the speaker and hearer to have the same density of information (Ferrer-i-Cancho 2017: Section 3). Even if we accept the idea of Fenk and Fenk (1980) that this is due to buffer memory restrictions, which allows to process approximately one syllable (in about half a second), one may wonder then about the predictability effects based on longer dependencies, such as the ones involved in tracking down the referents (see Section 1.2.2) and

processing of syntactic constructions. It is also not clear how to explain the well-documented effects of backward transitional probabilities – i.e. the probabilities of a target unit given the following units – on phonetic duration of the target units (see Section 1.2.6).

As Ferrer-i-Cancho et al. (2013) argues, uniform information density represents a particular case of constant entropy rate. Unfortunately, real texts do not meet this requirement. Constant entropy rate can be only found in texts with randomly scrambled words or letters, and homogeneous and periodic sequences, e.g. *aaaaaaaaa* or *abcabcabcabc*. Obviously, this has nothing to do with real language.

Moreover, it is not clear what the unit of analysis should be in order to falsify the Uniform Information Density hypothesis. To put is differently, for which units of speech should the information content be measured and compared? A word is a poor candidate because words have internal structure and different lengths, especially in languages with complex morphology.

In my view, the explanation of all these effects in terms of the Low-Cost Heuristic is perfectly sufficient. The hearer provides additional formal cues to help the hearer to make inferences in those situations when the interpretation is less probable, and omits them when it is more probable. The additional requirement of uniform information density does not add anything to this explanation.

The same probabilistic information seems to be at work in the cases of variable case marking, e.g. optional object marker *-o* in Japanese. The object may or not be marked. Kurumada and Jaeger (2015) show that Japanese speakers tend to mark the object when it is animate or occurs in a less plausible configuration of A and P, as in (40a). In contrast, in more stereotypical contexts, such as the one in (40b), the marker is often omitted:

(40) a. *The police officer attacked the criminal in the middle of the night.*

b. *The criminal attacked the police officer in the middle of the night.*

In probabilistic terms, the probability of interpreting the criminal as the object of the sentence in (40a) is lower than the probability of doing so in (40b). By using the object marker, the speaker signals that the situation does not correspond to the stereotypical situation. Note that this happens even when the other cues are sufficient for disambiguation (e.g. word order).

Another illustration is the use of resumptive pronouns in relative clauses. Keenan and Comrie (1977) found that languages use relative clauses according to the following hierarchy:

(41) Subject > Direct Object > Indirect Object > Oblique > Genitive > Obj. of Comparison

For example, if a language can relativize oblique nominal phrases, it can also relativize subjects, direct objects and indirect objects. More directly relevant for this study, however, is another finding by Keenan and Comrie, namely, that the same hierarchy constrains the use of resumptive pronouns in relative clauses. Consider an example from Hebrew:

(42) (Hebrew: Afro-Asiatic, Keenan and Comrie 1977: 92)

ha-isha	*she-David*	*natan*	*la*	*et*	*ha-sefer*
the woman	that David	gave	to-her	DO	the book

'the woman that David gave the book to'

Here, *la* is a resumptive pronoun in the indirect object position. According to the hierarchy, if a language has resumptive pronouns in the subject position, the pronouns will also be used in all other positions. If a language requires or allows them in the indirect object position (but not in the subject and direct object positions), it will also require or allow them for obliques, genitives and objects of comparison. How can one explain this observation? Keenan (1975) provides corpus data from English to demonstrate that the order in the hierarchy correlates with the frequency with which different positions occur. In a sample of above 2200 relative clauses, subjects are the most commonly relativized (e.g. *the girl who is playing a computer game*), and objects of comparison are never relativized. There are only a few examples of relativized genitives (e.g. *the gate of which the hinges were rusty*). From this we can conclude that the probability that a relative clause relativizes the subject, or $\mathbb{P}$ (Relativized_Subject|RelClause), is greater than the probability that a clause relativizes the object, or $\mathbb{P}$ (Relativized_DirObject|RelClause), and so on down the hierarchy. This is why relativized subjects need less additional coding than the other positions, which fits the pragmatic account presented here. Note that individual languages usually make cut-off points on the scale, so that

the use or absence of a pronoun in a given position is a categorical rather than probabilistic decision. As it happens very often, fine-grained probabilistic distributions are grammaticalized into coarse-grained distinctions (see also Chapter 6 on differential case marking). However, these splits are still functionally motivated (cf. Ariel 2008: 136–137).

We have discussed several instances of efficiency in morphosyntactic variation where the speaker relies on the hearer's previous experience with language. Another important source of information is the previous context, which makes some information more or less accessible. An interesting piece of evidence is provided by Lee and Choi (2010). In their experiment, they ask Korean speakers to rate focal subjects and objects with and without case marking, which is often optional in Korean. They distinguish between more and less predictable focal objects. The object is considered less predictable when the focus is replacing. Consider the micro-dialogue below. The object 'cell phone', which is not predictable from the previous context, is the replacing focus of the second utterance. The word can be used with and without the object marker.

(43) (Korean: Isolate, Lee and Choi 2010: 215)

A: *Cinmi-ka* *computer(-lul)* *sa-ss-e.*
Jinmi-NOM computer(-ACC) buy-PST-IND
'Jinmi bought a computer.'

B: *Aniya,* *hywutaephon(-ul)* *sa-ss-e.*
no, cell phone(-ACC) buy-PST-IND
'No, (she) bought a cell phone.'

Compare this example with a selecting focus, when both alternatives are already present in the first speaker's question, so that the word 'computer' in the second speaker's response is more accessible. Here, the ellipsis of the accusative marker is even favoured by native speakers:

(44) (Korean: Isolate, Lee and Choi 2010: 215)

A: *Cinmi-ka computer(-lul) sa-ss-e, hywutaephon(-ul) sa-ss-e?*
Jinmi-NOM computer(-ACC) buy-PST-IND, cell phone(-ACC) buy-PST-IND
'Did Jinmi buy a computer or a cell phone?'

B: *Computer/?computer-lul sa-ss-e. molla-ss-e?*
computer/computer-ACC buy- PST-IND didn't know
'(She) bought a computer. Didn't you know?'

Notably, native speakers of Korean produce higher acceptability ratings when an object with the replacing focus, as in (43), is marked than when it is unmarked. In such contexts, the mental representation of the object less accessible. In contract, they rate objects with the selecting focus, as in (44), higher when they are unmarked than when they are marked. In these contexts, the mental representation is more accessible because the object has been already introduced. Thus, higher contextual accessibility leads to zero marking, whereas lower accessibility means that the marker is more likely to be used.

As another illustration, consider the use or absence of the particle *to* after *help*. According to Rohdenburg (1996), the chances of the *to*-form increase with linguistic distance (in words) between *help* and the infinitive. For examples, the use of *to* is more likely in (45b) than in (45a):

(45) a. *You should help him (to) overcome his fears.*

b. *You should help this troubled teenager with many complexes and difficult childhood ?(to) overcome his fears.*

More information about this alternation is provided in Chapter 7. According to Rohdenburg, this variation can be explained by the principle of (reduction of) cognitive complexity:

(46) The principle of cognitive complexity:

> In the case of more or less explicit grammatical options the more explicit one(s) will tend to be favored in cognitively more complex environments (Rohdenburg 1996: 151).

Other formal asymmetries considered by Rohdenburg include inflected and uninflected present tense forms in non-standard varieties of English (e.g. *My mother and father drink/drinks*), optional prepositions (e.g. *time spent (in) doing something*) and prepositional substitutions (e.g. *She was prevailed on/upon to write another letter*). In addition to the linguistic distance, which was discussed above, higher complexity is also attributed to passive constructions.

I believe that the phenomena explained by Rohdenburg's principle can also be explained by the Principle of Communicative Efficiency. The cases like (45), which involve effects of linguistic distance, can be explained by the probability of the interpretation of *help* and the infinitive as one construction, which reduces with the linguistic distance between the components of this construction. As the linguistic distance increases and there are more and more words added between the matrix verb and the infinitive, the mental representation of the matrix verb becomes less accessible, which makes it more difficult to interpret the infinitival complement as a part of the construction with *help*. At the same time, the hearer may have less experience of using and processing such constructions in discourse because the structures like (45b) are quite rare. All this makes the intended interpretation (*help* + Infinitival complement) less probable as linguistic distance increases. According to the High-Cost Heuristic, the speaker should choose the costlier expression.

Importantly, there may be an overlap between the procedural and conceptual contents if the optional marker is not completely desemantisized. As Escandell-Vidal and Leonetti (2011) note, this is not uncommon: conceptual and procedural features can co-exist within a single unit without mixing up. For instance, some linguists have claimed that there is a conceptual difference between *help* + Infinitive and *help* + *to*-infinitive: it has been proposed that the variant with the bare infinitive designates a more active involvement of the Helper in carrying out the event expressed by the infinitival complement (Dixon 1991: 199). Consider the following examples:

(47) a. *John helped Mary eat the pudding* (he ate half).

b. *John helped Mary to eat the pudding* (by guiding the spoon to her mouth,

since she was still an invalid).

When *to* is omitted, as in (47a), the sentence is likely to describe a cooperative effort where Mary and John ate the pudding together; when *to* is included, as in (47b), the sentence means that John acted as a facilitator for Mary, who actually ate the pudding herself (Dixon 1991: 199; 230). Not all linguists agree with that distinction, however (see more details in Chapter 7).

1.2.5. Analytic support: variation of analytic and synthetic forms

The cognitive complexity principle, which was mentioned in the previous section, is also used by Mondorf (2014), who provides similar accounts of the use of synthetic and analytic forms in the following cases:

- English adjectival forms of comparison (e.g. *cleverer – more clever*, *fuller – more full*),
- the English genitive alternation (e.g. *the topic's relevance – the relevance of the topic*),
- English subjunctive alternation (*if he agree-Ø* vs. *if he agrees* vs. *if he should agree*),
- German past time alternation (*er brauchte – hat gebraucht*),
- English future tense alternation (*will – going to*), since *will* is often contracted to *'ll*.
- Spanish future tense alternation (e.g. *comeré* vs. *voy a comer*).

Here we are dealing with alternative forms that differ in the degree of autonomy, rather than in the presence or absence of a marker, as in the previous section.

Mondorf argues that the use of these forms can be explained in terms of processing demands. Analytic forms are used in the situations that require more processing efforts, while synthetic forms are used in easy-to-process environments. Complexity is multifactorial, and depends on many properties of contexts, such as the following:

- phonological factors (e.g. consonant clusters, as in *strict* or *apt*, can be complex);
- semantic factors, e.g. abstract and figurative concepts are regarded as more complex than concrete and literal ones. Compare the figurative use of *bitter* in *the more bitter takeover battles of the past* with a literal use: *the beer is bitterer;*
- negation, which is believed to add complexity to the context. Negated contexts generally increase the ratio of the longer variant, as has been found for the analytic and

synthetic future in Mexican Spanish (Lastra and Butragueño 2010) and the English subjunctive with zero-inflected verbs vs. would-subjunctive (Schlüter 2009);

- syntactic factors, e.g. longer constituents are more difficult to process, e.g. *he would be more proud to win honestly* vs. *he would be prouder if he had won honestly.* Also, as shown by Szmrecsanyi (2003), the more analytic future form with *going to* is more frequently used in syntactically dependent environments, which are more cognitively complex, than the *will*-future;[8]
- frequency, e.g. low-frequency words in comparison with high-frequency words. For example, the analytic more-variant is chosen more often with the adjectives that are infrequently used in the comparative (Mondorf 2003: 260–261). Mondorf (2014) also argues that the German synthetic preterite (*er sagte* 'he said') is being replaced by the German periphrastic perfect (*er hat gesagt*). At the same time, the synthetic past is still predominant with a few extremely frequent verbs (e.g. *war*, the past form of *sein* 'be'). This may be explained, however, by the conserving effect of frequency (Bybee and Thompson 1997).
- Discourse activation. For instance, the synthetic comparative form of an adjective can still be used in a complex environment, as defined by the previous conditions, if the comparative form of any type has been previously activated in context (Mondorf 2003: 285–286).

My preliminary interpretation of this variation is that the more complex contexts contain less frequent forms and meanings (e.g. passive forms, negation, complex syntactic structures and some figurative meanings), and therefore do not provide very reliable contextual cues for interpretation of the forms. The use of the longer form is therefore an example of the High-Cost Heuristic at work.

One cannot exclude the possibility that some processing optimization is going on, as well. This interpretation raises some difficult questions, however. First of all, note that all examples provided in this subsection display differences in length. One might ask then, whether analyticity itself provide any cognitive advantages for processing, or it is the length only. Second, whose task is facilitated – the speaker's, the hearer's, or maybe the tasks of both? For example, Hopper and Traugott (1993: 65) argue that the form *going to* is more substantive and

[8] There is also evidence that the periphrastic future was preferred more in subordinate clauses, in the 19th century Spanish (Poplack 2011).

therefore more accessible to hearers than *'ll* or even *will*. Developing this idea, Szmrecsanyi (2003: 23) concludes,

> Because BE GOING TO typically contains more material than WILL/SHALL, it provides a sort of redundancy that will ease online processing for hearers by making the predication more accessible.

At the same time, planning time is a scarce resource in complex environments. Therefore, it may be also advantageous for the speaker to use the longer form. For example, by using the longer form BE GOING TO, speakers can "stall" for planning time (Szmrecsanyi 2003: 23). Hopefully, these puzzles will be solved in future research.

1.2.6. Reduction and enhancement of linguistic units in speech production[9]

There is ample evidence in the literature that more predictable linguistic units (words, syllables and individual sounds) undergo reduction more frequently than less predictable ones. Already Bolinger (1963) observed that words are durationally shorter when they are predictable either due to their high frequency or the frequency of combinations where they occur, e.g. the relatively new word *robot* is pronounced longer than the more familiar *rowboat*, and verbs can be pronounced shorter when followed by more typical complements or adjuncts.

In pronunciation, articulation varies in order to reduce effort while at the same time producing a signal which shows sufficient acoustic distinctiveness for the listener to correctly identify the linguistic content of the message (Lindblom 1990). The probabilities that determine the predictability can be of different kinds. They may be context-free, determined only by the frequency of a given unit in discourse, and contextual, often defined as the conditional probability of a unit given the left or right context, e.g. *n* words on the left or right. The conditional probability can be measured in a specific context where the unit of interest is used, or it can be averaged across all contexts where the unit occurs.

Bell et al. (2009) studied the relationships between pronounced durations of words in a spoken corpus and several factors: frequency, contextual probability and repetition. They looked separately at content and function words. Both in content and in function words, there

[9] I thank Vsevolod Kapatsinski for sharing some useful information on this topic.

is a significant effect of different types of conditional probability – given the previous context or the next context. Moreover, word frequency (i.e. context-free probability) and repetition lead to reduction of content words. Similarly, Fowler and Housum (1987), as was mentioned in Section 1.1.2, found the effects of repetition on the duration of content words in a narration.

Phonological reduction can manifest itself not only in formal shortening, but also in the loss of phonological detail. For instance, Aylett and Turk (2004) report that phrase-medial syllables with high language redundancy (i.e., which are highly predictable from lexical, syntactic, semantic, and pragmatic factors) are shorter than less predictable elements. At the same time, they observe a loss of articulatory detail. In particular, vowels undergo centralization on their F1/F2 values. As a result, the vowel space is reduced (Aylett and Turk 2006).

In addition to such effects of "online" predictability, there are also effects of "offline" average predictability, measured across different contexts in which the unit occurs. As shown by Seyfarth (2014), both play a role in reducing the acoustic duration of a notional word, many other factors being controlled for. Therefore, form reduction is stored in the lexicon. Similarly, Cohen Priva (2008) tests if oral and nasal stop deletion in English is influenced by the phones' informativity, i.e. the local negative log predictability given all the phones that precede it in the same word, averaged across every instance of the phone in spoken natural language. Higher predictability means low informativity. Cohen Priva observes that average phone informativity decreases the probability of it being deleted in medial onset and coda contexts, even when frequency and online predictability are controlled for. Therefore, "informativity becomes part of the knowledge kept about each phone" (Cohen Priva 2008: 97). This is how the results of the use of a unit in a particular context percolate into language structure.

Pierrehumbert (2001) proposes an exemplar-based model in order to explain why high-frequency words undergo reduction faster than low-frequency words. For example, the middle schwa is deleted before /r/ and /n/ in high-frequency words, such as *evening* and *every*, but is retained in rare words, such as *mammary* and *artillery* (Hooper 1976). According to Pierrehumbert, this difference can be explained by the systematic production bias towards lenition (Lindblom 1984), or "undershooting" the phonetic target to the extent that it does not disrupt understanding. Since high frequency words are used more often than low frequency words, their stored exemplar representations are affected by this persistent bias more. This explains why high frequency words are more reduced than low frequency words synchronically and why the former undergo this reduction faster than the latter in diachrony. It does not seem

very plausible to me, though, that there is a certain constant rate of lenition, which is applied to every use of a word or sound in every context. A more realistic exemplar model should take into account predictability of exemplars in different contexts.

Another manifestation of shortening is word clipping. Mahowald et al. (2013) examine such pairs as *exam – examination*, *chimp – chimpanzee* and *math – mathematics*. Their corpus-based study and experiment with forced-choice sentence completion demonstrates that the online probability given the left context and the offline average predictability are associated with the shorter variants.

An example of such reduction at the level of a phrase is provided by Ariel (2008: 184). In contemporary Hebrew, there is a tendency to delete the adjective *tov* 'good' from greetings, such as *Boker tov* 'Good morning' and *Laila tov* 'Good night'. However, this does not happen with *(axar ha) cohoraim* 'afternoon' and *erev* evening. Ariel argues that the deletion in the first two greetings is possible because they are sufficiently frequent in discourse and are salient enough to license this omission.

Speakers also enhance the linguistic form in some circumstances, e.g. when they believe that the hearer may need help to disambiguate between two similarly sounding words, e.g. *dose – doze*. This has been shown in studies of hyperarticulation. In particular, they tend to increase the voicing of the final consonant in words like *doze* when the hearer chooses between two similarly sounding words. The voicing is then more likely in this condition than in the situations when such ambiguity is not present (Seyfarth et al. 2016). They also hyperarticulate when their communication partners misunderstand the instructions (e.g. Stent et al. 2008). This effect increases immediately after the speaker finds out that he or she was misunderstood, and then decays gradually over several turns in the absence of further misrecognitions (*Ibid.*).

Explanation of these effects has been a controversial issue. First, such reduction and enhancement can be explained by audience design (Bell 1984). Generally speaking, this means that language users pro-creatively adjust their message in order to increase their communicative success, while at the same time reducing their efforts any time they can. This is an instance of efficient communication based on the Principle of Communicative Efficiency. The interaction between the interlocutors plays an important role here. They are constantly updating their common ground, coordinating their language, reducing or enhancing the forms in the process (Clark 1996; see also Vajrabhaya 2016).

But this is not the only explanation that can be found in the literature, nor the most popular one. It is often argued that the predictability of a word helps it to be selected and articulated faster (e.g. Bell et al. 2003). A related popular view in usage-based linguistics involves the phenomenon of chunking of neighbouring units. According to Bybee, for example, each instance of use further automates and increases the fluency of the sequence, leading to fusion of the units (Bybee 2007: 324). A frequently repeated stretch of speech becomes automated as a processing unit due to the neuromotor routines. Further repetition leads to the reduction and overlapping of articulatory gestures. All this shortens the duration. For instance, Bybee and Scheibman (1999) found that reduction of the vowel and the consonants in *don't* in spoken English is particularly frequent after the pronoun *I* and before the verbs *know* and *think* because this contraction particularly frequently occurs in phrases *I don't know* and *I don't think*. This process of automatization is not restricted to language alone and is largely unconscious.

An interesting question is then, whether this is the joint probability of two or more neighbouring units, or the conditional probability of one unit given the other that determines the chances of reduction. Some empirical evidence that the latter may be more important than the former can be found in the literature on language production. In particular, Bell et al. (2003) investigated the effects of conditional probabilities and joint probabilities (i.e. the probabilities of the target word and the preceding or following word together) on the duration and phonetic reduction of function words in spoken English, and found that the conditional probabilities have either the strongest or the only significant effect in the predicted direction (i.e. more predictable target words are more frequently reduced than less predictable ones). Joint probabilities, which basically represent the frequencies of possible chunks, sometimes have an effect in the opposite direction. This finding can be regarded as a piece of evidence that predictability is more important than chunking.

Another piece of evidence comes from Gahl and Garnsey's (2004) experimental study, where they show that the effects of syntactic predictability of a direct object or a complement clause on the duration of the preceding verb and the final *-t/d* deletion cannot be accounted for by automatization and chunking.[10] Gahl and Garnsey's carefully selected stimuli differed in the probabilities of the syntactic units, rather than the probabilities of the nouns following the verb, e.g. *He accepted the money* vs. *He accepted the money is evil*. This means that reduction due to chunking cannot explain all instances of phonetic reduction.

[10] I thank Karsten Schmidtke-Bode for making me aware of this important piece of evidence.

Yet another production-based explanation is that the speaker buys time for planning by using a longer expression. As was already mentioned, this was one of the explanations offered by Szmrecsanyi (2003) to provide an account for the use of the construction *be going to* in syntactically complex environments, which are more demanding in terms of processing resources. As shown by Bell et al. (2003), planning problems, which are represented by disfluencies either preceding or following a function word, increase the chances of longer or fuller variants of words in language production. While such effects are not excluded, many instances of reduction are difficult to explain by planning issues only. For example, Jaeger and Buz (2017) argue that the link between the contextual predictability of a linguistic form and its own realization is not very clear if one accepts the 'buying-time' explanation. There is also evidence that backward transitional probabilities (i.e. those that predict the target unit given the following context) play a role that is at least as important as the role of forward transitional probabilities (i.e. the ones that predict the target unit from the preceding context), if not even more important (Seyfarth 2014). Moreover, speakers adapt subsequent productions towards less reduced variants if previous use of more reduced variants resulted in communicative failure (Stent et al. 2008; Buz et al. 2016). As Jaeger and Buz (2017) argue, this is incompatible that the idea that reduction is solely due to production ease. See also an overview of different perspectives in Vajrabhaya (2016).

One cannot not exclude the possibility that routinization, "stalling for time" and other production-related phenomena all play a role. It may also be that the importance of the production factors is higher in the least conscious processes, such as phonological reduction, and lower when the choice of the linguistic form is conscious, as in the use of non-conventional implicatures. The evidence, however, is not sufficient for excluding the Principle of Communicative Efficiency entirely from the list of factors that trigger phonetic reduction or enhancement. In my view, the effect of production factors should be ultimately constrained by the communicative need of the speaker, who wants to get the message across, although some of the lower-level reduction or enhancement processes can be caused by the cognitive processes unrelated to the hearer's needs (cf. Lindblom 1990). Some sophisticated experiments are needed in order to obtain a conclusive answer and to disentangle different cognitive and social effects.

A final word of warning should be said against a potential misunderstanding that an account based on audience design should only display online effects. There is no irreconcilable conflict between this account and the evidence of entrenchment effects, which can last for a

while, or even become conventionalized. For example, the voice-onset time of words with initial voiceless stops that have minimal pairs, e.g. *cod – god*, is greater in comparison with the words without such a pair, e.g. *cop – *gop*. Baese-Berk and Goldrick (2009) found that this difference is observed even if the minimal pair is not present in the context (i.e. there is no need of disambiguation). They conclude that this effect is not driven by what they call 'listener-modeling'. We know from Cohen-Priva (2008), Seyfaert (2014), which were mentioned above, and other studies, that units that frequently occur in reducing contexts also become more reduced in general, i.e. usage percolates into the system. Therefore, units that are sufficiently frequently hyperarticulated or reduced in some contexts, may become hyperarticulated or reduced across the board. This may lead to short-term or long-term effects. For example, Stent et al. (2008) show that hyperarticulation is a targeted and flexible adaptation to a specific situation, which decays with time. At the same time, reduced or enhanced forms can be entrenched and conventionalized in their conjunction with specific communicative situations. As a result, whole special registers can emerge, e.g. child-directed speech, foreigner-directed speech, etc. (Jaeger and Buz 2017).

1.2.7. Zipf's Law of Abbreviation

Finally, we need to address perhaps the most famous manifestation of language efficiency, namely, the negative correlation between the average probability and length of linguistic units at the level of types (words, syllables, phrases, etc.). This correlation is known as Zipf's Law of Abbreviation (1935 [1965]): more frequent units tend to be shorter than less frequent ones. Recently, Bentz and Ferrer-i-Cancho (2016) have tested the law on 986 languages from 80 families, using massively parallel corpora. They find a significant negative correlation between word length in characters and word frequency for all languages (on Parallel Bible Translations). Thus, the Law of Abbreviation is an absolute language universal, although it is statistical in each separate language (the correlations are rather weak). Notably, Piatandosi et al. (2011) found out that the average informativity, i.e. basically the reverse of the conditional probability of a word given its previous context (1 to 3 words on the left), is even more strongly correlated with word length than simple frequency.

Zipf also raises the question about the direction of causality: is the length of a word a cause or a result of its usage frequency? According to him,

> on the whole the comparative length or shortness of a word cannot be the cause of its relative frequency of occurrence because a speaker selects his words not according to their lengths, but solely according to the meanings of the words and the ideas he wishes to convey" (1935 [1965]: 29).

He concedes that sometimes speakers, due to their youth or lack of experience, may seek to avoid long or unusual words. At the same time, some speakers may prefer longer and more unusual words, which would counterbalance the effect of the former.

As the main causal factor of this correlation Zipf (1935 [1965]) names saving time and effort. The linguistic mechanisms where this efficiency manifests itself are as follows:

(a) truncations (*movies* instead of *moving pictures*, *gas* instead of *gasoline*);

(b) substitutions, permanent or temporary. Temporary substitutions are anaphoric pronouns (see Section 1.2.2). Examples of permanent substitutions are *car,* which is used instead of *automobile* or, in more specialized domains, *juice* for *electricity* or *soup* for *nitroglycerine*.

Zipf also argues that the frequency of concepts is crucial. For example, the action of striking something with the chin is less common than striking with the foot. This is why English possesses a single word only for the second concept, i.e. *kick.* Another example is *brother* vs. *uncle's second wife's tenth child by her first marriage*. Similarly, Hawkins (2014: 17), who provides an example of a teacher and a teacher who is late for class, writes:

> The more frequently selected properties are conventionalized in single lexemes or unique categories and constructions in all these examples. Less frequently used properties must then be expressed through word and phrase combinations and their meanings must be derived by semantic composition.

Zipf and Hawkins leave out an important aspect, namely, how this correspondence between form and frequency emerges. I will argue that the Low-Cost and High-Cost Heuristics play an important role in the conventionalization of simple expressions, which designate frequent concepts, and in the emergence of complex compositional expressions, which convey rare concepts. The argumentation is presented in Chapter 3 (Section 3.5.2).

Similarly, as pointed out by Hawkins (2014: 17), there are semantic and syntactic properties that frequently occur across languages and that have priority in grammatical and lexical conventions, such as causation, agenthood, patienthood, frequent speech acts (asserting, commanding, questioning). These functions usually have distinct formal expressions across grammars. Note that less frequent speech acts (baptizing, bequeathing) are assigned separate lexical items, rather than being represented by unique syntactic constructions. This makes the expression of more frequently used meanings shorter, that of less frequently used meanings longer.

One should also consider formal erosion as a result of grammaticalization (e.g. Lehmann 2015: Section 4.2.1). For example, full verbs become auxiliaries (the Old English *willan* 'want' > *will* and even *'ll*), full pronouns become clitics (e.g. *them* and *'em*), *because* becomes '*cause* and even *coz*. A more detailed discussion of the diachronic mechanisms that lead to efficient formal asymmetries is provided in the next chapter.

Although Zipf's work has been extremely influential, his elaboration of the Principle of Least Effort (1949) is based on a somewhat misleading analogy between language and an artisan in a workshop. A language user is the artisan, words and morphemes are tools, and words' length corresponds to the tools' physical size and mass. The main problem with this analogy is that the artisan plans consciously how his tools are arranged and what size they have in order to minimize his effort over an extended time period. This creates an impression that language change is teleological, and that language users consciously optimize words in order to make the language maximally efficient. Obviously, this is incorrect. As will be argued in Chapter 2 (Section 2.5), linguistic efficiency is an unplanned, unintentional result of intentional actions based on the 'invisible hand' principle (Keller 1994).

There have also been some sceptical opinions about the interpretation of Zipf's Law of Abbreviation in terms of efficient organization of language. Miller (1957) noted that a correlation between word length and word frequency is also observed when one types characters randomly. At the same time, Howes (1964) argued that the assumptions of Miller's model (crucially, the equal probability of all letters) are not applicable to natural language. It has also been argued that the correlation between a word's conditional probability and its length found by Piantadosi et al. (2011) is not necessarily due to language efficiency and optimal choices: random typing yields a linear relationship between these variables, as well (Ferrer-i-Cancho and Moscoso del Prado Martín 2011). At the same time, it is difficult to deny that at least some of the Zipfian effects are real. Examples can be easily found in daily life. Consider

the recent political events in Germany, for instance. One of the candidates who have been competing to succeed Angela Merkel as the leader of the Christian Democratic Union is Annegret Kramp-Karrenbauer.[11] When she became a candidate for this key position and began to appear regularly in the news, the German and international media began to refer to her as AKK. Obviously, this is done for optimization of communication. A challenging task for the future is to separate the true Zipfian effects like this one from possible statistical artefacts.

1.3. Summary of the chapter and the structure of this study

This chapter has discussed the communicative and cognitive foundations for manifestations of efficiency in language. Communicative efficiency is defined as maximization of the benefit-to-cost ratio in communication. This principle has analogues in biological evolution. In human language, efficient communication requires spending not more and not less effort than it is necessary to trigger the intended cognitive effects in the hearer. Based on the assumption of mutual rationality, the speaker and hearer share the heuristics "Low costs – low benefits" (the Low-Cost Heuristic) and "High costs – high benefits" (the High-Cost Heuristic). The costs are defined by the articulatory effort, and the benefits can be measured as the change in the hearer's mental representations in comparison with the mental representations before the communicative cues were provided. On the neural level, this change corresponds to the change in the neural activation of particular nodes and brain areas. If they are already activated due to the recency or high frequency of linguistic units and links between them, the change is small. Therefore, the cognitive effects (the benefits) are small. In this situation, a low-cost expression is expected. If they had a low level of activation – due to the absence from the immediate context and low frequency of the corresponding exemplar – the change is substantial. In this case, the benefits are large, and a high-cost expression should be used, following the Principle of Communicative Efficiency.

The accessibility of information depends on the previous and immediate context, knowledge of the world and one's previous experience with language. On the cognitive and neural levels, one can speak about the strength of exemplars, activation, priming and resting

[11] https://en.wikipedia.org/wiki/Annegret_Kramp-Karrenbauer (last access 21.11.2018)

activation level in neural networks. In probability theory, these phenomena are reflected in the probabilities of forms and interpretations, including the conditional probabilities of interpretations given some forms. Probability is inversely correlated with the information-theoretic notion of information content (or surprisal).

Efficiency is not the only factor that shapes language structure and use, of course. Among other factors that can either inhibit or facilitate the emergence of efficient asymmetries one can name the following:

1. An important factor is systemic pressure, which is based on analogy. As shown by Haspelmath and Karjus (2017), the role of efficiency may be rather small in comparison with analogy. The strength of systemic pressure may be different for different phenomena. For instance, the singular and plural marking of nouns usually exhibits more systematic patterns than the expression of causal and non-causal events, which seems to be more sensitive to the usage profiles of individual lexical items, which have a tendency to represent spontaneous or caused events (Haspelmath 1993; Haspelmath et al. 2014; see also Chapter 3).

2. Cognitive processes within the speaker's mind, such as routinization and automation, memory access and planning, may facilitate or impede the production of reduced or enhanced forms, as discussed in Section 1.2.6. One of the relevant factors is persistence (Szmrecsanyi 2006), which explains why speakers tend to use the same grammatical and lexical variants in discourse, even if that might potentially clash with efficiency considerations.

3. Language learning processes influence the linguistic form. For example, complex morphology may be lost in the situations of intensive language contact with a large number of adult L2 learners (Trudgill 2011). Also, a more efficient system may be more difficult to learn, and the other way round.

4. Phonological factors may also contribute to reduction or enhancement. There is evidence, for example, that speakers tend to add optional function words in order to avoid potential stress class, i.e. a combination of adjacent stressed syllables (e.g. Wasow et al. 2015). It has also been observed that more reduced forms tend to occur before a consonant than before a vowel (e.g. Bell et al. 2003).

5. Social considerations, such as politeness and taboo, may lead to omission of some information (e.g. Goldberg 2005), as mentioned in Section 1.2.2.

6. Normative considerations should not be underestimated. Innovative reductions are often treated as a sign of linguistic and personal sloppiness. For instance, in standard Belgian Dutch

the verb *kijken* 'look, watch' takes a prepositional complement *naar* 'to' + NP. At the same time, there is a tendency to omit *naar* before TV, films, series, TV, Netflix, etc., following the Low-Cost Heuristic. Yet, some people regard this omission as degradation of the language and vociferously demand that the others keep using the preposition.[12]

Examples of efficient formal asymmetries include some types of Gricean and Neo-Gricean implicatures, anaphoric expressions, grammatical categories, the use or omission of function words, analytic and synthetic forms, phonological and morphological variants, and Zipf's Law of Abbreviation. Efficiency is thus observed at all levels of language structure. The choice between forms can be either categorical and fully conventionalized, as in the grammatical coding asymmetries (e.g. *book – books*) or probabilistic, depending on the context (e.g. phonetic reduction, the genitive and dative alternations or the use and omission of complementizers and relativizers in English). Semantically, some of the pairs have contrastive meanings, e.g. *book – books*, some are near-synonyms, e.g. *kill – cause to die*, and some have little noticeable difference, e.g. the use or absence of the complementizer *that* or variation of *help* + (*to*) Infinitive. Some of these pairs are related formally, e.g. the pairs with zero or non-zero marking, where one variant is simply a shorter form of the other one. The others are made of different linguistic material, e.g. *kill* vs. *cause to die*. Yet, I argue that the formal differences can be explained by contextual accessibility of the information and/or by the hearer's previous experience with language and the world that the speaker relies on. Moreover, these phenomena vary with regard to the level of conscious attention and control on the part of the language users. For instance, the implicature in *Mary produced sounds that reminded of Jingle Bells* requires the speaker's and the hearer's attention. On the opposite pole are phonetic reduction and the use or omission of function words, which are usually unconscious.

In the remaining part of this study I will focus on four phenomena:

1. Variation of causative constructions (e.g. *stop the car* vs. *get the car to stop*). This topic is addressed in Part II (Chapters 3–5).

2. Differential case marking, which is discussed in Part III (Chapter 6).

[12] Based on an article in De Morgen, a daily Belgian newspaper (in Dutch): https://www.demorgen.be/tvmedia/als-we-zomaar-woorden-beginnen-af-te-schaffen-spreken-we-binnen-de-kortste-keren-allemaal-jerommekestaal-ba500534/ (last access 19.11.2018). See also linguists' reaction to the article here (also in Dutch): https://www.arts.kuleuven.be/nieuws/taalcolumns-mogen-best-wat-beter-geinformeerd-zijn-opinie-freek-van-de-velde-en-dirk-pijpops (last access 19.11.2018).

3. Use and omission of function words in the following cases: *help* + *(to)* Infinitive, *be/sit/stay*, *etc. (at) home* and *go (and)* + Infinitive. These cases are examined in Part IV (Chapters 7–10).

4. Phonetic merge and reduction in the alternation *want to* vs. *wanna* + Infinitive. This alternation is also discussed in Part IV (Chapter 10).

In these case studies, efficiency will be operationalized and measured, in the most general form, as a correlation between the length (measured as the number of segments) and the relevant probabilities, which are computed from corpora and reflect the language users' experience with language.

Before we move to the case studies, it is necessary to discuss the diachronic scenarios and mechanisms which can cause efficient formal asymmetries to emerge. This will be done in the next chapter.

Chapter 2. Efficiency in diachrony

2.1. Aims of this chapter

The previous chapter discussed different types of efficient formal asymmetries which display efficiency. In line with the Principle of Communicative Efficiency, less costly forms are used to convey more probable, predictable, accessible, etc. information, while costlier forms represent less probable information. These formal asymmetries emerge due to the Low-Cost and High-Cost Heuristics. The asymmetries can propagate in the language system and become conventionalized. The present chapter provides illustrations and a general discussion of such changes.

The Low-Cost Heuristic leads to reduction of forms as an adjustment to the high probability of the information they convey. The change may involve a phonologically or morphologically modified version of the original expression (e.g. *going to* > *gonna* or *mathematics* > *maths*) or come from some other source (e.g. *car* replacing *automobile*). This is onomasiological change, which is driven by the changes in the probability of the meaning. In addition, one can speak about semantic innovations, when a form becomes associated with a more probable meaning. This is semasiological change, which occurs when the form becomes less costly.

The High-Cost Heuristic leads to enhancement of a form as an adjustment to low probability of the information it expresses. Enhancement manifests itself in hyperarticulation of the original features (as in *It's a **pin**, not a bin*) or addition of some new phonological or morphological material. A short expression can also be replaced with a longer one, as in the process of renewal, when a new and longer construction takes over the functions of an older and reduced one. This is onomasiological change. Taking the semasiological perspective, one can imagine a situation when an existing long form begins to be associated with a less probable interpretation than the original meaning.

A cause for onomasiological changes is the changing probability of the meaning expressed by a form. This change may happen due to cultural reasons. This is the case with *car*, which replaced *automobile* (Zipf 1935 [1965]: 33) due to the increasing popularity of this kind of transport. In German, the shorter variant *Auto* becomes the default, as opposed to the original

form *Automobil*. The replacement of *Automobil* by *Auto* is an onomasiological change based on the Low-Cost Heuristic. At the same time, the longer variant *Automobil* has changed its usage from neutral contexts to more pragmatically marked elevated or ironic contexts.[13] This is a semasiological change based on the High-Cost Heuristic.

As mentioned by Zipf (1935 [1965]), the innovation often begins in a special interest community. An example is *app,* from *application program*. This abbreviation lived its quiet life in the community of software developers in the 1980s and 1990s.[14] Since the creation of Apple's App Store and similar platforms, it became a popular word known to the general public.

Consider another example from Tenepaja Tzeltal, a Mayan language spoken in Mexico (Witkowski and Brown 1983). Both reduction and enhancement are observed here. Before the conquest, the word *čih* designated deer. After the conquest, sheep were imported. The new and exotic animals were named by a longer expression, *tunim čih* 'cotton deer'. This corresponds to the High-Cost Heuristic because the costlier expression corresponds to the less probable meaning. As sheep were becoming increasingly popular, the adjective 'cotton' was dropped, and the word *čih* began to designate sheep. This is an instance of reduction based on the Low-Cost Heuristic. Presumably, the word without the adjective was still polysemous with 'deer'. Finally, the expression for 'deer' apparently got enhanced in order to avoid the ambiguity, which lead to *teʔtikil čih* 'wild sheep' – again, a result of the High-Cost Heuristic at work. Thus, we observe three stages:

(1) (Tenepaja Tzeltal: Mayan, Witkowski and Brown 1983: 571)

	DEER	SHEEP
Stage 1 (pre-conquest)	*čih* 'deer'	-
Stage 2 (early post-conquest)	*čih* 'deer'	*tunim čih* 'cotton deer'
Stage 3 (contemporary)	*teʔtikil čih* 'wild sheep'	*čih* 'sheep'

Tenepaja Tzeltal is spoken in the Chiapas highlands, where sheep and the manufacture of woolen products are important. In contrast, in a closely related language, Bachajón Tzeltal,

[13] https://www.duden.de/rechtschreibung/Automobil (last access 07.11.2018).
[14] https://blog.oxforddictionaries.com/2011/10/14/the-rise-of-the-app/ (last access 07.11.2018).

which is spoken in the lowlands, where sheep are uncommon, only the two first stages have happened. The name for ‘sheep’ is still ‘cotton deer’.

This description is onomasiological. We examine first the changes in the probabilities of the meanings and then look at the formal changes. If we take the semasiological perspective, moving from forms to meanings, we can see that the shorter expression *čih* in Tenepaja Tzeltal conveys first the originally more typical ‘deer’, but later, after a period of ambiguity, begins to designate the increasingly popular concept ‘sheep’.

It is usually very difficult to separate onomasiological changes from semasiological ones because of constant mutual coordination of linguistic tools used by the speaker and the hearer in the process of communication. For example, in the process of grammaticalization the meaning of a gram gravitates towards more general and therefore more probable functions, whereas the form undergoes gradual reduction (see Section 2.2).

In addition, we often do not have enough data to tell whether we deal with reduction, enhancement or both. Even in the simple case of the English complementizer *that*, we cannot say for sure whether people omit it or add it (Jaeger and Buz 2017). Similarly, when discussing changes in word duration, it is often not clear whether we deal with enhancement of less probable units, or reduction of more probable ones (Vajrabhaya 2016). Therefore, the mechanisms described here represent to some extent theoretical abstractions.

The processes based on the Low-Cost and High-Cost Heuristics are discussed in Sections 2.2 and 2.3, respectively. Section 2.4 discusses an alternative, source-based type of explanations. Section 2.5 contains a note on teleology, which precludes potential misunderstandings of the approach developed in this thesis. A summary and discussion are provided in Section 2.6.

2.2. Changes triggered by the Low-Cost Heuristic

The onomasiological manifestation of changes based on the Low-Cost Heuristic is formal reduction. It is central to language change in general. As put by Langacker, “It would not be entirely inappropriate to regard languages in their diachronic aspect as gigantic expression-compacting machines” (Langacker 1977: 106).

Reduction happens in the situations when the speaker assumes that the hearer can easily recover the intended meaning even if the form is reduced. As discussed in the previous chapter, the speaker does not need to say something that is obvious from the context, common knowledge, etc. The hearer recognizes that the intended information should be highly probable or easily accessible. This leads to efficient formal asymmetries, for example, when the form that corresponds to the more frequent category undergoes erosion or loss of marking, while the form that corresponds to the less frequent category remains the same (Croft 2003: 116).

As an illustration, consider the (in)alienability distinction in possessive constructions. In Old Italian, the shorter inalienable pattern arose by phonological reduction due to high frequency, whereas the longer alienable pattern did not undergo this process (Haspelmath 2017):

(3) (Old Italian < Latin, Rohlfs 1949–1954, cited from Haspelmath 2017: 222)

a. *moglia-ma* < *mulier mea* 'my wife' (inalienable)

fratel-to < *fratellus tuus* 'your brother' (inalienable)

b. *terra mia* < *terra mea* 'my land' (alienable)

A similar difference can be observed in English dialects:

(4) (Lancashire English, Hollmann and Siewierska 2007: 407, cited from Haspelmath 2017: 222)

a. *m[ɪ] brother* (inalienable)

b. *m[aɪ] football shoes* (alienable)

As was mentioned in Section 1.2.3, the basis for this reduction is the fact that inalienable entities are more typical in the role of possessees than alienable objects.

Another example is the gradual loss of the final *-n* by English possessive determiners *mine* > *my* and *thine* > *thy* from Middle English to the 18th century (Hilpert 2012). Unlike the possessive determiners, their predicative counterparts (e.g. *The book is mine*) did not undergo

reduction. The resulting formal asymmetry is efficient, since the dependent forms (i.e. the determiners) are used more frequently than independent forms (i.e. the pronouns). According to Michaelis (In press), this asymmetry is widely attested in languages of the world. An example is Juba Arabic, a lingua franca spoken in Sudan, where the original form *bita-i* [POSS-1SG] 'my/mine' has been reanalysed as the dependent possessive and reduced to *tái* 'my', e.g. *ída tái* 'my hand', whereas the non-shortened form *bita-i* continues to be used as the independent possessive form, i.e. 'mine' (Michaelis, In press).

As was mentioned in Section 2.1, grammaticalization involves both onomasiological and semasiological change. When a unit is used frequently, it automatically becomes more probable. According to the Low-Cost Heuristic, it will undergo formal reduction (e.g. *going to* > *gonna* before an infinitive). At the same time, an innovative and less costly form will signal the hearer that the meaning is more probable.[15] This can trigger semantic generalization, or bleaching. For example, in the case of *going to/gonna*, the documented path of change is from directed motion to intention, and from intention to prediction of future events. Consider the examples below:

(5) a. *Where are you going? – I'm going to see my aunt.* [Directed motion]
b. *I'm going to write a letter.* [Intention]
c. *I think I'm going to sneeze*. [Prediction]

Directed motion with a human subject normally implies intention, as in (5a), while intention does not entail directed motion, as in (5b). Similarly, intention often implies prediction, as in (5b), while prediction does not entail intention, as in (5c) (Croft 2000: 162). Thus, intention is semantically broader and therefore more probable than directed motion, and future is semantically broader and more probable than intention.

The higher probability of the new meaning also explains why the unit becomes more frequently used after this semantic shift is made – simply because this form-meaning pairing is appropriate in more situations. This increases further its predictability. As a result, the form is even more reduced, and so on. Figure 2.1 demonstrates this cyclic process. Note that these changes are slow and gradual, rather than fast and abrupt.

[15] Langacker (2011: 83) gives a hint to a similar idea, speaking about the parallelism between 'less meaning' and 'less form' as an explanation of why formal reduction is accompanied by semantic reduction in the process of grammaticalization. However, he seems to attribute this correspondence to iconic motivation.

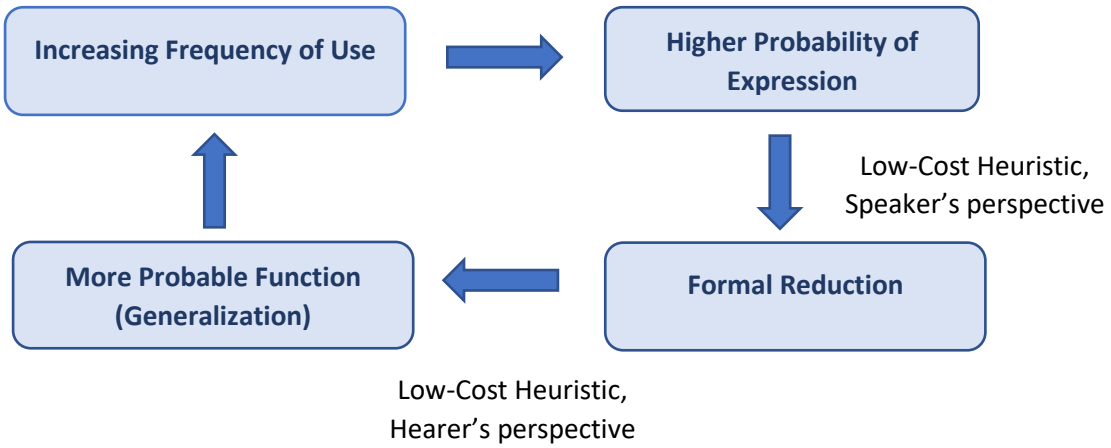

Figure 2.1. The cyclic process of formal reduction and semantic generalization (bleaching) based on the Low-Cost Heuristic

This mechanism accounts for two most important manifestations of grammaticalization: formal reduction and semantic bleaching or generalization (Haiman 1991). It does not exclude other typical processes, such as chunking and automatization. However, as was argued in the previous chapter, these processes are ultimately subordinate to the higher-level pragmatic goals and are allowed only to the extent that they do not hinder comprehension. In addition, reanalysis can determine the type and locus of formal and semantic reduction. The common paths of grammaticalization (e.g. motion > intention > future) emerge because of typical co-occurrence of semantic features in discourse. As was argued above, directed motion with a human subject usually implies intention, and speaking about one's intentions implies speaking about the future (Croft 2000: 162).

While the model in Figure 2.1 is similar to some previous accounts, in particular, with regard to the role of frequency as both a cause and an effect in a self-feeding process (Bybee 2010: 107; Langacker 2011: 83), it also explains a crucial aspect that have not been made sufficiently clear in the previous descriptions of grammaticalization, namely, why the meaning becomes more abstract and grammatical. A popular explanation of semantic bleaching has been habituation, by which a repeated stimulus loses some of its semantic force because people cease to respond to it at the same level (Haiman 1994; Bybee 2003). This happens, for example, to swear words, when they are used very frequently, ritual apologies and greetings, or van Gogh's paintings when they are reproduced everywhere – on posters, dishes, scarves and umbrellas. Habituation does not require reduction. It links directly frequency (or probability of the

expression) and semantic bleaching. However, a semantic change when a linguistic unit allegedly becomes trite and stale is different from the process where a lexical meaning becomes more grammatical and abstract, e.g. from directed motion to intention, and from intention to prediction, as in the development of *going to* as a future marker. Therefore, habituation itself cannot explain the development of grammatical meanings, although it may serve as a facilitating factor, which weakens the more specific aspects of a word's meaning (cf. Haspelmath 1999: 1062; Bybee 2003: Section 6.3).

Another explanation why frequent units undergo semantic generalization was offered by Zipf (1949). His metaphor of a language user as an artisan and words as instruments on his workbench was discussed in Section 1.2.7. According to Zipf, the artisan is more likely to pick up a more accessible tool that is located closer to him, even if it does not fit the task perfectly, than to bring a more appropriate tool, which is farther away. In other words, we might use a knife to open a food package if the scizzors are in another room. Frequent words are more accessible than rare ones for new meanings in production. Therefore, they will be preferred to more semantically appropriate but rare words. This hypothesis was tested by Harmon and Kapatsinski (2017) in an artificial language learning experiment. Indeed, speakers tend to use the more frequent forms for new categories. In other words, more frequent units are more onomasiologically salient than their less frequent alternatives (cf. Geeraerts et al. 1994). However, it remains unclear what drives the qualitative change in the meaning, which becomes more grammatical and abstract, and why these developments follow common diachronic paths (e.g. motion > intention > future).

2.3. Changes triggered by the High-Cost Heuristic

In cases of efficient formal enhancement, we expect the more probable meaning to be expressed by the old and shorter form, while the new and longer form expresses the less probable meaning.

These changes sometimes arise for the purposes of overriding the default interpretation of a linguistic cue. An example is the rise of reflexive pronouns from emphatic forms (e.g. König and Vezzosi 2004; Ariel 2008: Chapter 6). As shown by Ariel, it is much more common to describe activities or situations in which a participant engages with other participants, rather than with him-/herself. For example, *John hit Peter/his enemy/his scientific opponent, etc.* is a more typical scenario than *John hit himself*. Coreferential arguments of the same predicate are

a minority. This explains why in many languages reflexive pronouns are longer than non-reflexive ones, e.g. *John hit him – John hit himself.* Historically, they often originate from emphatic forms. For example, the emphatic *self* was added in Old English and then became a part of the reflexive pronouns. Ariel described this process as follows:

> …Old English speakers at some point started adjoining an independent emphatic form (*self*) to their pronouns in order to counteract the default pragmatic inference to disjointness. (Ariel 2008: 222)

Similar processes have taken place in other languages, where the origin of reflexive pronouns can be traced back to emphatic markers, e.g. Turkic, Finno-Ugric, Caucasian, Persian, Japanese, Indic and Semitic languages (Ariel 2008: 223).

Importantly, König and Vezzosi (2004) point out that the onset contexts for the development of such reflexive anaphors are sentences with other-directed transitive verbs (e.g. *help* and *deliver*) and third-person singular subjects. It is in these contexts, they argue, that the need for disambiguation and reinforcement is the greatest. If the action expressed by a verb is other-directed, this means that the probability of a non-coreferential object is higher than the probability of a non-coreferential object: $\mathbb{P}$ (Non-Coreferential_Object|Verb) > $\mathbb{P}$ (Coreferential_Object|Verb). Therefore, the less probable coreferential objects need additional formal marking. Note that if the action is very likely to be self-directed, as in grooming verbs, e.g. *He washed and dressed himself,* it can be expressed without any pronoun, e.g. *He washed and dressed*. This is an example of the Low-Cost Heuristic at work.

As for the person-related grammaticalization asymmetry, it can be explained by the fact that a 3rd person subject and a 3rd person object can either be coreferential (in a pseudo-English without reflexives, that would be *she$_1$ sees her$_1$*) or have disjoint reference (*she$_1$ sees her$_2$*), while a 1st or 2nd person subject and object (*I$_1$ see me$_1$*) are always coreferential (probably, with the exception of speakers suffering from the split personality disorder). Therefore, the contexts where one needs to provide extra coding are those where the referent is more difficult to identify (that is, in the 3rd person) because there may be more than one candidate present in discourse (see Section 1.2.2). This explains the cross-linguistic universal: if a language has a first person reflexive pronoun, it also has a third person reflexive pronoun (Faltz 1985: 43, 120).

An alternative explanation, proposed by Newmeyer (2003: 694–695), is that the the predominant tendency to grammaticalize the 3rd person distinction in languages of the world is due to the higher frequency of coreferential 3rd person subjects and objects. He provides corpus data which show that reflexive pronouns in English are more common in the 3rd person than in the 1st and 2nd person (see also Haspelmath 2008: Note 4).

> Given that more frequently appealed to concepts are more likely to be lexicalized than those that are less frequently appealed to, the implicational relationship among reflexive pronouns follows automatically. There is no need to appeal to ambiguity-reducing 'usefulness'. (Newmeyer 2003: 695)

If this is true and disambiguation indeed does not matter, the efficiency account would predict that the 3rd person coreferential pronouns should emerge after the 1st and 2nd person pronouns because language users are more used to coreferential subjects and objects in the 3rd person. That goes against the diachronic observations made by König and Vezzosi (2004) and typological evidence. If this is true, the efficiency theory will run into problems.

Let us look closer at the the frequency data reported by Newmeyer. He shows that the third person reflexive pronouns are predominant (*myself* – 169 occurrences, *yourself* – 94, *himself* – 511, *herself* – 203, *itself* – 272). These data come from the LOB corpus, which represents written English (e.g. books and periodicals). No wonder that the 3rd person pronouns are so frequent. It is well known that the frequencies of different person forms in a corpus depend on the type of texts it contains (e.g. Biber 1988). In written formal texts there are few references to the 1st and 2nd person, while in spontaneous conversations they are very common. I took a subset of the spoken part of the Russian National Corpus with informal dialogues and conversations,[16] and searched for the full reflexive pronoun *sebja* as a direct object (this form is the same for all grammatical persons). After a manual cleaning, I obtained 163 examples, where the object was coreferential with the subject.[17] The distribution of the person forms is very different from Newmeyer's: the reflexive pronoun coreferential with the 1st person is the most frequent (72 occurrences), followed by the 3rd person (54 occurrences) and the 2nd person

[16] http://www.ruscorpora.ru/en/ (last access 25.11.2018).

[17] I only took the examples where I could identify the coreference from the available context. I also discarded several expressions, e.g. *čuvstvovat' sebja* 'feel (oneself)' and *vesti sebja* 'behave (oneself)', where no non-coreferential substitutions of *sebja* are possible without a change in meaning.

(37 occurrences). Therefore, there is no evidence that the 3rd person coreferential objects are the most frequent in informal speech, which represents the primary mode of interaction between language users.

One should also add here that the presence of personal non-reflexive pronouns (e.g. *John chides him*) implies, by virtue of a Q-implicature, disjoint reference, since the use of reflexives with other-directed verbs is conventionalized in coreferential contexts (cf. Levinson 2000: 287).

An example of ongoing change can be found in Nigerian Pidgin English. There are several past tense forms: zero and with various additional markers. According to Poplack and Tagliamonte (1996), the zero form is the default form for neutral past tense reference, which can be used to state what happened without any emphasis. This may be due to substrate influence, since the conceptual space marked by zero is very similar to what is covered by the unmarked past in some West African languages. The additional markers are used to divert from this neutral interpretation and express sequential, continuous, anterior remote or other readings. These constructions have not reached a very high stage of grammaticalization yet. At least, the lack of additional markers does not conventionally mean the lack of these more specific meanings, i.e. the zero marking is not grammaticalized (cf. Bybee 1994). The exact interpretation is to be inferred from the context.

Yet another reason for enhancement is the desire to be expressive and attract attention. Haspelmath (1999) calls it the principle of extravagance, "speak in such a way that you are noticed". Petré (2017) argues that the increase in the use of the present progressive tense *be* + Ving in the Early Modern English period is due to the speakers' desire to make their expressions cognitively more salient, more noticeable in comparison with its neutral competitor, the simple present. Compare the examples below:

(6) (Early Modern English, Petré 2017: 236)

a. *You are now poysoning your souls by sin…*

b. *Thou pleasest thy throat, and poysonest thy soul.*

Extravagance of this kind is also a manifestation of the Principle of Communicative Efficiency. The cognitive effects achieved here with the help of enhanced forms are emotional and

interpersonal, rather than related to some objective information about the referential situation. The conventional progressive interpretation of *be* + Ving in the present may be a by-product of such extravagance, since the ongoingness is usually associated with high speaker involvement, as argued by Petré. It is also interesting, however, that the construction has never spread to designate the habitual aspect.[18] This is in fact common cross-linguistically (Bybee 1994: Section 6). Present progressive grams do not generalize their meanings to include habitual senses. While originally progressive grams may also cover imperfective in the past tense (e.g. in Turkish or Scots Gaelic), this is not observed in the present. At the same time, as argued by Bybee, habitual is the default meaning of the present tense. Therefore, the new expressive present progressive construction started its grammatical 'career' by covering the less typical meaning of ongoingness, in accordance with the predictions, while the old and "boring" construction has kept the most typical present tense function – i.e. the habitual meaning.

I believe that this is a typical scenario. In the examples where I could find some information about the distributional differences between the older and newer forms, the older form is first used in the most typical contexts, while the newer form emerges on the periphery, where the meaning is less typical. If the newer form gains in frequency, it takes over some of the territory of the older form, but the latter is still used in the most typical situations. Consider an example from Maltese (Koptjevskaja-Tamm 1996; Haspelmath 2017), which has a relatively novel construction to express alienable possession with *ta'* 'of', which has different personal forms:

(7) (Maltese: Semitic, Haspelmath 2017: 224)

a. *Inalienable possession*

id 'hand' > *id-i* [hand-1SG.POSS] 'my hand'

[18] With the exception of so-called subjective progressives, e.g. the famous McDonald's advertising slogan *I'm loving it* or *You're always losing your things!* In such examples, the emphatic function may still be alive.

b. *Alienable possession*

ktieb ‘book’ > **ktieb-i* [book-1SG.POSS]

il-ktieb tiegħ-i [ART-book of-1SG] ‘my book’

The analytic construction is newer and has largely replaced the classical Arabic suffix *-i*, with the exception of body parts and kinship terms (with which both forms are possible), and a few fossilized expressions (Eksell Harning 1980; Koptjevskaja-Tamm 1996). It represents the longer alternative. At first, it was used with rather untypical possessors, e.g. possession of an abstract noun or duration (e.g. *a two hours' journey*), as one can infer from the data from different Arabic dialects. As Koptjevskaja-Tamm (1996: 262) writes, “it emerges first of all in those uses where the need for it is most acute, most pronounced.” This means that the longer construction is used when the probability of the possessive interpretation is lower. This is in accordance with the High-Cost Heuristic. Later, however, the new analytic genitive took over a substantial part of the meaning of the initial synthetic genitive. At present, the synthetic genitive is restricted to relatively few classes, it still has a relatively high relative frequency (Koptjevskaja-Tamm 1996), due to the central position of these nouns as possessees and their high individual token frequency. These stereotypical contexts represent the last bastion of the older and shorter forms, due to the inference based on the Low-Cost Heuristic.

This logic can be extended to other instances of periphrasis, when more analytic expressions are used instead of the old synthetic ones in order to say approximately the same thing. With time, they may become grammaticalized and replace the old forms – the process known as renewal (Hopper and Traugott 1993: 65). When the old and new constructions co-exist in one functional domain, they express different meanings, which differ in their probability. A famous example of renewal is the development of future forms. The central function of the future tense is prediction (Bybee et al. 1994: Chapter 7). New and longer future constructions begin with peripheral functions, e.g. immediate future and intention, which are often related to motion and modal meanings (especially volition and obligation).

Whether a particular formal asymmetry emerges due to reduction or enhancement, is difficult to predict. Revisiting the possessive pronouns, which were discussed in the previous section, one finds quite a few instances of enhancement of the less frequent independent form, as in *The book is mine*. In her study based on creole languages, Michaelis (In press) names at least five different sources where additional coding material comes from: adpositions ‘of’ and

'for', dummy nouns 'part' or 'thing', intensifiers, nominalizers and determiners/demonstratives. For example, in Haitian Creole *pa m nan* 'mine' is the result of lengthening of the dependent form *m (nan)* 'my', as in *se m* 'my sister' by the noun *pa* 'part'. At the same time, in some creole languages one can find examples of reduction of the more frequent attributive form (see Section 2.2.).

2.4. Source-based explanations

Opponents of functional explanation of language change claim that the grammatical asymmetries of the type described in Section 1.2.3 develop because of the properties of the source constructions, rather than for functional purposes, such as disambiguation. Consider an example from Cristofaro (In press). She shows that plural markers can develop from partitive, e.g. 'many of them', as in Bengali:

(8) (Bengali: Indo-Aryan, Chatterji 1926: 735-736, cited from Cristofaro In press: Section 2)

a. āmhā-**rā** sābā
   we-GEN all
   'all of us'

b. chēlē-**rā**
   child-GEN

Other source constructions for plural are distributive expressions (house here and here) and expressions of multitude (all), as in Southern Paiute (Uto-Aztecan) and Maithili (Indo-Aryan). Discussing such cases, Cristofaro writes, "These various processes do not appear to be triggered by the higher need to disambiguate plural as opposed to singular." (In press, Section 2).

However, this is not the whole story. With regard to the Bengali construction, Chatterji (1926: 726–727) points out that the addition of some noun of multitude (e.g. 'all') to the noun

was a new device "[t]o indicate the plural, which had come to be indistinguishable from the singular" because of the loss of the original plural nominative affix – that is, exactly for the purpose of disambiguation. In other words, a costlier expression is provided in order to cancel the more probable interpretation. The construction could be used to represent plural in New Benghali (from 1800 on), e.g. *rājā**rā**-såbɸ* 'kings' (Chatterji 1926: 734), although the noun of multitude was already perceived as superfluous. It was first omitted with 1st and 2nd person pronouns ('we', 'ye'), and later with other pronouns and finally with nouns. To summarize, we observe here a construction, whose function was to highlight the non-stereotypical interpretation and which subsequently got reanalyzed and lost one of its elements due to the Low-Cost Heuristic – first in the most predictable situations, where it was redundant, and later in the less predictable ones.

The fact that the original constructions performed different functions does not represent counter-evidence to the efficiency-based explanation. Language users create new form-meaning pairings by adjusting semantically suitable constructions for their needs, not by deciding that some new phonetic sequence will from now on represent the plural. This kind of recycling is the norm. What is crucial is that the opposite situations – i.e. when when a morpheme or word becomes reanalyzed as the singular marker, while the plural form has zero expression – are very rare. One such examples is Imonda (a Papuan language), where the plural is unmarked, and some nouns take the singular marker that comes from the partitive construction (i.e. 'one from among the group of X'). However, this marker seems to be used only with five human nouns: women, men, girls, boys and enemies (Seiler 1984: 62–63).[19] The fact that plural markers regardless of their origin propagate more frequently than singular markers, as in Imonda, suggests that the solution with marked plural and unmarked singular forms is more efficient than the one with unmarked plural and marker singular forms.

Also, as argued by Schmidtke-Bode (In press), the functional-adaptive explanation is the only option that can explain efficient use of variable marking. For instance, the object marker in Japanese (see Section 1.2.4) can be used in those situations where the object is less predictable. Moreover, the differential coding splits, such as differential case marking or differential argument indexing, indicate the boundaries of propagation, where the functional expansion of the marker stops because its use is no longer efficient. More about that will follow in Chapter 6.

[19] Normally, the number is indexed on the verb, whereby singular is unmarked.

2.5. A note of teleology

It is important to say a few words about the causes of emergence of efficient language systems. Different trends of functionalism have different views, as pointed out by Bybee (1999). For example, Dressler (1990: 76) argued that language development is teleologic:

> …both linguistic universals and all language Systems have the teleology of overcoming substantial difficulties of language performances (including storage/memorization, retrieval, evaluation) for the purpose of the two basic functions of language: the communicative and the cognitive function.

In other words, languages change in order to become better systems for the purposes of communication and cognition. In contrast, Bybee has often argued against the view that language development has any long-term goal to aim at or any purposes to fulfil (e.g. Bybee et al. 1994: 297–298; Bybee 2010). The development of grammar results from mental and communicative processes (with the focus on "mental"), which are rather "mechanistic" in nature. For example, she wrote:

> the increase in efficiency in high-frequency words results from the way the general neuromotor system operates, and is neither restricted to language nor a conscious goal-directed process (Bybee 2010: 146).

There seems to be a lot of confusion around the terms, such as teleology, intentionality, or goal-orienteness, and consciousness (cf. Keller 1994). There seems to be a misconception that functional explanations of language (such as the one developed in this thesis) are necessarily teleological, and therefore flawed.

The notion of teleology goes back to Aristotle and his theory of four causes. One of the causes is called final: the properties of an explanandum can be explained by what it is for. For example, Aristotle argues that animals have sharp front teeth for biting, and flat molar teeth for

grinding the food. Since this is a regularity, and not a coincidence, the shape of the teeth should be explained by their function.[20] Such explanations are called teleological.

Leaving aside the relevance of that pre-Darwinian explanation, one can speak about three possible types of teleological causes: some external benevolent force (intelligent aliens, magic, deities, etc.), smart self-optimization or intentional actions of language users. The first two are obviously not applicable (as far as I know). A language is not a sentient being that can adjust its behaviour, and there is no supernatural force that would cause it to become more efficient.

As for intentional actions of language users, the answer is less obvious. Users perform their communicative tasks intentionally, in order to meet their practical needs. Their main concern is to get their message across, with all illocutionary and perlocutionary effects that they deem desirable (e.g. to influence others, to impress, to convince, to get important information, etc). These conscious actions are made up of fully or partly unconscious sub-actions which determine the linguistic shape of an utterance, e.g. using the present continuous tense, choosing a word out of a set of synonyms, deleting final *-t* or *-d*, etc. The potential for conscious attention is different for the different types of efficiency examined in Chapter 1. For example, creative implicatures, such as the one derived from *Mary produced sounds that reminded of Jingle Bells*, are very likely to be consciously produced, unlike the use of grammatical markers and function words or subtle reduction and enhancement of phonetic details in pronunciation. However, even the use of such units can occasionally become conscious, e.g. when someone repeats a word with an emphasis on a segment in case of misunderstanding. This hierarchical structure is typical of human actions: conscious goal-directed activities consist normally of automated unconscious sub-actions, such as pressing the keys when playing the piano, changing the gears when driving the car, or performing precise muscular movements when using a knife and fork. At the same time, they can be called intentional because they form part of intentional actions. The unconscious or conscious choices of individual users may propagate in the language system and become conventionalized as an unintended result of intentional actions in an 'invisible hand' process (Keller 1994). All these considerations are summarized in Table 2.1.

[20] See Aristotle. *The Organon and Other Works*. Opensource collection. Translated under the editorship of W.D. Ross. https://archive.org/details/AristotleOrganon. *Physics*, Book II, Sections 8–9.

Table 2.1. Different properties of language use and change

Phenomenon	**Driven by some external force or self-optimization**	**Intentional**	**Conscious**
Individual use of language	No	Yes	Yes and no
Language change	No (although it may appear so)	Usually no	Usually no

In her influential work, Bybee emphasizes the key role of fully unconscious cognitive processes within a language user's mind, such as entrenchment, schematization, routinization and chunking, in explaining language change. These are undoubtedly very important. However, what is missing in this view is a link between mental representations of an individual and language as a system of conventions. These are phenomena that belong to different 'worlds' (Popper 1972):

(9) Popper's three worlds

World 1: physical entities

World 2: individual mental states, including states of consciousness and psychological dispositions and unconscious states

World 3: world of the products of the human mind, such as scientific theories, works of art, social institutions and artefacts.

While the cognitive processes discussed by Bybee belong to World 2, language as a social phenomenon belongs to World 3, although it also normally has a physical manifestation in World 1 (Geeraerts 2016). The key question is how the mental representations (Word 2), which are central in the usage-based framework, can have an impact on the social reality (Word 3). In order to answer this question, one needs an interface between cognition and language as a social phenomenon (e.g. Schmid 2015, see also Divjak et al. 2016). Such an interface is provided by language usage, which requires joint attention, common ground and coordination of each other's mental representations and communicative tools (Clark 1996; Tomasello 2008),

as well as a belief in each other's rational behaviour and adherence to the communication maxims. The speaker performs reduction or enhancement of forms based on the shared knowledge of the actual situation, the world and the language, readjusting the forms depending on the hearer's reactions and the results of interaction.

Thus, the pragmatic perspective is a necessary and even central element for explaining various manifestations of efficiency in language and the natural link between the properties of a linguistic system and the cognitive and neural mechanisms. The speakers and the hearers are, by necessity, idealized, but there is some support from studies of actual language production (see Chapter 1), which allows us to extrapolate these findings to real communicators and the decisions that they make. Hopefully, these ideas will be supported by future experiments (one of them is presented in Chapter 5) and observational studies of human interaction.

2.6. Summary and discussion

This chapter has discussed how efficient formal asymmetries can emerge in grammar and lexicon based on the Low-Cost and High-Cost Heuristics. I have argued that the main driving force of these changes is the alignment of costs and benefits. More probable interpretations, which provide few cognitive effects, are expressed by less costly (shorter) forms, and less probable ones are conveyed by costlier (longer) forms.

When a reduced form spreads gradually over a range of functions or contexts, it should begin with more probable functions or contexts, and later spread to less predictable ones. As for enhancement, it is the other way round. The longer form appears first in the least predictable meanings, and ends in the most predictable ones. This hypothesis needs to be tested on diachronic data and in experiments.

The source-based explanations of efficient formal asymmetries do not provide a satisfactory account, in my opinion. They do not explain why the efficient formal asymmetries are so common cross-linguistically. In addition, they do not explain the existence of efficient coding splits and variable marking in languages of the world.

Emergence of efficient asymmetries in a language represents an unintended result of the language users' intentional actions. In this sense, it is an "invisible hand" phenomenon.

Therefore, language optimization (and change in general) is not a teleological process. It is not entirely mechanistic, either, because it is based on intentional actions, although some subroutines that involve reduction and enhancement are unconscious. Communicative efficiency has a sociocognitive nature and involves common ground and coordination of the speaker and the hearer's mental states and linguistic tools.

A causal explanation of efficient language should account for the actions of the individual participants. Explanations of "invisible hand" phenomena are valid only when they consider the actions of individual language users:

> A theory of the history of language has explanatory adequacy in so far as it succeeds in correlating reconstructed historical data with descriptive adequacy to the linguistic actions whose consequences they are; that is to say, by demonstrating that they are the necessary and unintended consequences of individual actions carried out according to specific maxims of action under specific ecological circumstances (Keller 1994: 159).

Such an account has been presented in this thesis. The next parts will zoom in on different types of efficient formal asymmetries in languages of the world.

Part II. Formal asymmetries between near-synonyms: Causative constructions

Chapter 3. Causative constructions: Form, meaning and frequency

3.1. Aims of this chapter

This chapter focuses on causative constructions in languages of the world. Probably the most famous cross-linguistic generalization about causatives is that formal compactness of causative constructions correlates with the directness of causation. Consider a famous example from Fodor (1970):

(1) a. *John caused Bill to die (on Sunday by stabbing him on Saturday).*

b. *John killed Bill (*on Sunday by stabbing him on Saturday).*

In (1a), the causing and caused events are not spatiotemporally integrated, and the causation is indirect. In contrast, in (1b) the events should occur in the same time and space, and the causation is direct. Formally, the periphrastic causative contruction *cause to die* in (1a) represents the causing and caused events separately, whereas the lexical causative verb *kill* in (1b) merges them in one word. Generally speaking, more conceptually bound events are expressed by more formally bound forms, as in (1b), and less conceptually bound events are expressed by less formally bound forms, as in (1a). Some researchers, including Haiman (1983, 1985), have explained this correlation between form and meaning as a manifestation of iconicity. Following Haspelmath (2008b), I will refer to this type of iconic relationships as iconicity of cohesion, in order to distinguish it from other types of iconicity.

In this chapter and two subsequent ones, I will argue that efficiency provides a better explanation for cross-linguistic variation of causatives than iconicity of cohesion, developing

the line of argumentation presented in Haspelmath (2008b). My explanation involves asymmetries in the probabilities of the meanings expressed by the constructions. These asymmetries also explain the differences in the degree of conventionalization and grammaticalization of the constructions. Developing the standard account of causatives in pragmatics, I will argue that the Low-Cost and High-Cost Heuristics, which are based on the probabilistic information about different types of causative situations, play a central role in the emergence of such efficient form-meaning pairings.

The arguments that support this idea are the following.

1. The cross-linguistic variation of causative constructions with regard to their compactness is not restricted to (in)directness. There are other semantic parameters that are correlated with different degrees of formal compactness. I will argue that all of these correlations, including the one related to (in)directness, can be explained by the differences in the probabilities of the corresponding causative situations. This argument is developed in the present chapter.

2. The cross-linguistic variation of causative constructions that express direct and indirect causation or other closely related distinctions, correlates more strongly with length differences than with the formal autonomy of the elements that express the cause and the effect, or the distance between them. This is demonstrated in the typological study presented in the next chapter.

3. Finally, one can model the development of efficient formal asymmetries in causatives without any iconic correlations. This is demonstrated in an artificial language learning experiment reported in Chapter 5.

The present chapter builds upon Dixon's (2000) semantic parameters of causation, which includes not only (in)directness, but also such parameters as involvement of the Causer in the caused event and the Causer's intentions. I argue that these parameters are not reducible to (in)directness and present the data from my typological database, which supports Dixon's observations and finds new distinctions. Some of Dixon's semantic parameters are made more specific or corrected. Next, I present spoken corpus data from three languages (English, Lao and Russian), which show very clearly that the functions that are expressed cross-linguistically by more compact forms are much more probable than the ones that are expressed by less compact forms. I will argue that the formal differences between different causatives are related to the probabilities of the situations that they express.

The structure of the chapter is as follows. Section 3.2 presents an overview of the previous literature on the form-meaning mapping in causative constructions. Section 3.3 presents the results of the typological survey. In Section 3.4, I present the probabilities of the meanings based on the spoken corpora. Section 3.5 discusses how various types of causatives can emerge and develop. Finally, Section 3.6 summarizes the results.

3.2. Theoretical background: correspondences between meaning and form

Causatives have received a lot attention in typological literature (e.g. Comrie 1981: Ch. 8; Song 1996; Dixon 2000; Shibatani and Pardeshi 2002; Haspelmath et al. 2014). Normally, they are classified into several types, which form the so-called causative continuum (Comrie 1981):

(2) Lexical – Morphological – Analytic (or Periphrastic)

where lexical causatives are where the cause and effect are expressed in one morpheme, e.g. *kill, break, give*. Morphological causatives contain a causative morpheme, e.g. Turkish *öl-dür-* 'kill' from *öl-* 'die'. Finally, analytic or periphrastic causatives are those in which the causative meaning is expressed by a combination of words, whereby the causing and caused events are expressed separately, e.g. *make* + NP + *dead, cause* + NP + *to die.* The latter type will be referred to as syntactic causatives in this study. These categories form a continuum because, first of all, the differences between lexicon, morphology and syntax are not clear-cut. In particular, one may argue about the class membership of non-productive morphological causatives, e.g. English *wid-en* and *solidi-fy,* which may still exhibit morphological boundaries. Such causatives would be located between the prototypical lexical ones (e.g. *kill*) and the prototypical morphological ones (see the Turkish example above). The second reason is that there may be different degrees of 'separateness' of the cause and the effect. For example, French causatives with *faire* 'make' are usually immediately followed by the infinitive, whereas constructions with *demander* 'ask, request' are followed by a nominal phrase, and then the infinitive, as in (3).

(3) (French; Comrie 1981: 162)

a. *J'ai fait manger les pommes à Paul.*

'I made Paul to eat the apples.'

b. *J'ai demandé à Paul de manger les pommes*

'I asked Paul to eat the apples.'

In this case, the cause and effect in the construction with *demander* will be less integrated. Therefore, the construction will be more 'syntactic' than the construction with *faire*. One can also speak about monoclausal and biclausal syntactic causatives, although this distinction is not clear-cut, either (Kulikov 2001: 887).

It has been observed that the continuum in (2) corresponds to the semantic continuum of direct and indirect causation, as shown below:

(4) Lexical – Morphological – Syntactic

more direct < ……> less direct

As Comrie (1981: 165) puts it, "the kind of formal distinction found across languages is identical: the continuum from analytic via morphological to lexical causative correlates with the continuum from less direct to more direct causation". This correlation is traditionally explained by an iconic correspondence between form and function: "[t]he linguistic distance between expressions corresponds to the conceptual distance between them" (Haiman 1983:782). Haiman's (1985: 105) scale of linguistic distance is shown in (5).

(5) a. X # A # B # Y

b. X # A # Y

c. X + A # Y

d. X # Y

e. X + Y

f. Z

In this cline, X and Y are the linguistic expressions of interest, which express the cause and the effect, A and B are other intervening units, # represents a word boundary, + stands for a morpheme boundary, and Z is a morpheme where X and Y are fused. It is important to note that Haiman's cline in (5) involves not only formal distance *per se*, but also autonomy of X and Y. For example, the types (5a) X # A # B # Y and (5b) X # A # Y differ only in the distance between X and Y, whereas (5d) X # Y and (5e) X + Y differ only in autonomy, representing two words or two morphemes, respectively. Obviously, words are more autonomous than morphemes. As for the types (5e) X + Y and (5f) Z, they differ both in distance and autonomy. The formal distance between the forms X and Y and/or their autonomy decreases from (5a) to (5e), until they are fully fused in (5f). This subtle difference between autonomy and distance will be important in the next chapter.

Directness and indirectness of causation can be defined in different ways, which presents a challenge for those studying this form-meaning correspondence. One can speak, for instance, about the spatiotemporal integration of events, as in the example (1) above, or about the physical contact between the participants (Haiman 1983). An illustration is provided in (6), taken from Haiman (1983: 784). In (6a), an instance of indirect causation, the Causer employs some unnatural force (e.g. magic or telekinesis) to produce the result. In contrast, in (6b) the Causer uses his or her own physical force.

(6) a. *I caused the cup to rise to my lips*.

b. *I raised the cup to my lips*.

Moreover, direct causation has been defined as causation in which the Causer is the main source of energy responsible for the caused event (cf. Verhagen and Kemmer 1997). In indirect causation, there is some other source. For example, stabbing someone dead represents an instance of direct causation because the energy comes from the Causer. In contrast, imagine that someone tempers with one's gun, so that the owner gets shot on a duel. This represents indirect causation because the energy necessary for killing comes from his opponent's gun. Using magic powers is another instance of indirect causation.

With animate Causees, the energy can come from the Causee. The Causer can make the causation indirect by giving directions (directive causation). However, this is only possible when the Causee is agentive and responds to the causing event (e.g. the Causer's command or request) by performing an action, as in (7a). When this agency is not present, e.g. when the Causee is asleep, as in (7b), the causation is direct despite the fact that the Causee is animate.

(7) a. *He made the children lie down.*

b. *He laid the children down.*

This account also agrees with Givón's, who predicts that periphrastic causatives are more likely to code causation with a human-agentive 'manipulee' (i.e. Causee), whereas morphological and lexical ones are more likely to code causation with an inanimate manipulee (Givón 1990: 556). Consider also so-called curative causation, where the Causer has something done by the Causee. A typical example is when the action is a part of the service provided by the Causee professionally:

(8) *I had my hair cut (by the hairdresser).*

The Causee is backgrounded, since it is not important who performs the action. The nominal phrase is the Affectee (Verhagen and Kemmer 1997). Indirect causation is also associated with transitivity of the predicate that expresses the caused event, as in the next example, where the Causee (*the mechanic*) serves as an intermediary in bringing out the change in the Affectee (*my transmission*).

(9) *I had the mechanic fix my transmission.*

Moreover, one can also regard letting and permission as indirect causation. In Cognitive Semantics, letting is defined as non-impingement, or cessation of impingement. The Causee's intrinsic tendency towards rest or motion is not changed by the Causer (Talmy 2000: 417–421). Compare direct causation in (10a) with non-impingement in (10b) and cessation of impingement in (10c):

(10) a. *Walt rolled his barrel of cash through the desert* (i.e. horizontally, using mostly his own energy).

b. *Walt let Jane choke to death* (i.e. by non-interference).

b. *Walt removed the stone and let his barrel of cash roll down the hill* (i.e. exploiting the barrel's tendency towards motion due to the force of gravity).

To summarize, one can make a list of semantic sub-parameters that can be interpreted as different aspects of (in)directness:

(11) Sub-parameters of (in)directness of causation:

	Direct causation	Indirect causation
a.	Spatiotemporal integration	Lack of spatiotemporal integration
b.	Causer is the main source of energy	Causee or another force (e.g. magic) is the main source of energy
c.	Causee is affected	Causee is an intermediary or agent
d.	Short causation chain (two participants, intransitive predicate expressing the effect)	Long causation chain (three participants, transitive predicate)
e.	Impingement (making)	Lack or cessation of impingement (letting) or weak non-physical impingement (having, getting)

Yet, there are several other parameters which are also correlated with formal variation of causatives and which cannot be reduced to (in)directness, even if we take into account all the sub-parameters listed in (11). These parameters are discussed by Dixon (2000). First of all, he introduces a scale of formal compactness, which includes four classes. The compactness decreases from lexical causatives (12a) to periphrastic causatives (12d):

(12) a. Lexical causatives, e.g., *break*$_{TR}$ or *walk*$_{TR}$;

b. Morphological causatives, e.g., internal or tone change, reduplication, or affixation;

c. Complex Predicates, e.g. serial verbs, French *faire* 'make' + V_{INF}, or causative particles;

d. Periphrastic causatives, where the causatives are represented by verbs that belong to separate clauses, e.g., French *laisser* + NP + Infinitive or Portuguese *fazer* + (NP) + Infinitive.

According to Dixon, the degree of compactness is correlated with the semantic and syntactic features which are shown in Table 3.1 (Dixon 2000: 76). If a language has two different causative forms, a more compact and a less compact one, they will differ along one or more parameters.

Table 3.1. Parameters of variation of causatives and their correlation with formal compactness, according to Dixon (2000)

	More compact forms	**Less compact forms**
1.	non-causal verb describing a state	non-causal verb describing an action
2.	intransitive (or intransitive and simple transitive) non-causal verb	transitive (ditransitive) non-causal verb
3.	causee lacking control	causee having control
4.	causee willing ('let')	causee unwilling ('make')
5.	causee partially affected	causee fully affected
6.	direct causation	indirect causation
7.	intentional causation	accidental causation
8.	causation occurring naturally	causation occurring with effort

The ninth semantic parameter Dixon discussed in his study is involvement of the Causer in the caused event. Yet, Dixon did not find any correlations between this parameter and the degree of compactness. Note that the 4th parameter in the table predicts more compact forms for willing Causees (letting), and less compact forms for unwilling Causees (making). It seems that two different distinctions are mixed up here. The first distinction is whether the Causee resists the Causer's action or not and which should be treated as the 8th parameter, i.e. whether the causation occurs naturally or with effort. The second distinction is that between making and letting (permissive causation). Here, the typological data, which are presented below, and previous corpus-based research (Levshina 2016) show clearly that making is expressed by more compact forms than letting.

Consider now the distinction between the intentionally and non-intentionally acting Causer. The sentence in (13a) is an example of intentional Causation, whereas (13b) exemplifies accidental causation.

(13) a. *The thief broke the window and got in.*

b. *Oops, I've just broken your Ming vase!*

According to Dixon, intentional causation is often expressed by more compact forms, while accidental causation is expressed by less compact forms cross-linguistically. An example can be found in Kammu, an Austro-Asiatic language spoken in Laos. In Kammu, the prefix *p(n)-* expresses intentional causation, whereas the particle *tòk* expresses unintentional causation:

(14) (Kammu: Austro-Asiatic; Svantesson 1983: 103–111, cited from Dixon 2000: 70)

a. *kəə* *p-háan* *tráak*

3SG+M CAUS-die buffalo

'He slaughtered the buffalo.'

b. *kəə* *tòk* *háan* *múuc*

3SG+M CAUS die ant

'He happened to kill the ant (e.g. by accidentally treading on it).'

Other meanings that are associated with less compact forms include forceful and comitative causation, which are difficult to interpret in terms of (in)directness, without making this distinction vacuous. For example, the causer acting accidentally is not necessarily acting indirectly, as in (15a), while acting intentionally does not mean acting directly, as in (15b).

(15) a. *Sorry, I've broken your iPad by pressing the screen too hard.*

b. *So how do you make it so that he does want to text you back?*[21]

I will argue that these correlations between form and meaning, including, but not limited to (in)directness, can be explained by efficiency. Below I test provide typological data and corpus evidence that support this theory.

3.3. Typological data: multifunctionality

3.3.1. Language sample

In this section I provide data from my typological database of causative constructions, which shows, in addition to Dixon's parameters, what kind of causation is expressed by forms with different degrees of compactness. I took a sample of 59 languages, each from a different language family, in which more than two causative constructions were described. The geographical distribution of the languages is shown in Figure 3.1. Lexical causatives were excluded (usually there is very little information about their common meaning in grammars).

[21] https://www.vixendaily.com/love/how-to-get-a-guy-to-text-you-back/ (last access 17.11.2018).

The data come from reference grammars. The list of languages and references are provided in Appendix 1.[22]

These constructions were then analyzed semantically and formally, and all pairs of causatives were compared within each language. I found information about the semantic differences (which was either provided explicitly by the grammars or could be inferred from the examples) in the pairs from 53 languages (see more information in Appendix 1). Only these constructions are analyzed in this chapter.

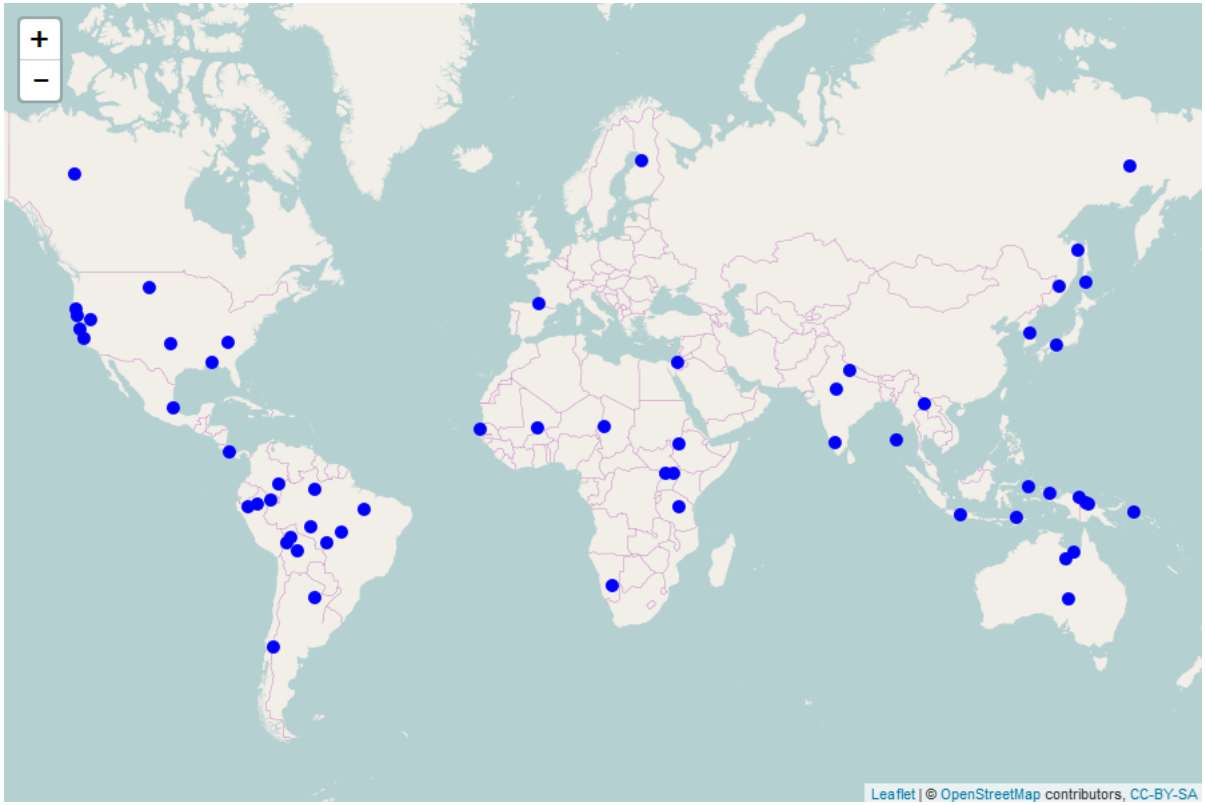

Figure 3.1. Geographical distribution of languages in the cross-linguistic sample

3.3.2. Form and functions of causatives

Compactness was determined according to Dixon's scale in (11). Labile verbs (e.g. *burn* or *melt*) were considered more compact than morphological causatives, whereas light verb constructions were considered more compact than serial verb constructions. Causatives with

[22] The database is available on GitHub: https://github.com/levshina/TypoCaus.

clitics were considered more compact than analytic causatives, but less compact than morphological causatives. Note that these notions are language-specific categories, which were determined according to the descriptions provided by the authors of the grammars.

Table 3.2 presents the semantic features of the less compact form in a pair of causatives. This is done because in many cases only the less compact form has a special semantic description in a grammar, whereas the more compact form is treated merely as a valency-increasing device, the default causative. An example is given below from Trumai, a language isolate from South America. In (16a), the default causative with the particle *ka* is used. In (16b), one can see the periphrastic causative with the verb *tao* 'order/give order', which means that the periphrastic construction represents causing someone to do something by order:

(16) Trumai (Trumai; Guirardello 1999: 356, 360)

a. *hai-ts Yakairu-ø sa ka.*

1-ERG Yakairu-ABS dance CAUS

'I made Yakairu dance.'

b. *hai-ts ka_in [Atawaka-ø pa] tao.*

1-ERG FOC/TNS Atawaka-ABS marry order

'I ordered Atawaka to marry.'

This lack of semantic specificity of the more compact causatives can be explained by the pragmatic approach developed in this study. The meaning of direct, intentional, non-forceful, factitive, etc. causation is inferred from the 'default', purely valency-increasing causatives on the basis of the Low-Cost Heuristic. As will be shown in Section 3.4, these features are much more probable than indirect, non-intentional, forceful, permissive, etc. causation. If the second construction becomes so frequent that the meaning of the default causative is compared with it, then Levinson's (2000) Q-implicature is inferred and subsequently conventionalized. The more compact causative then becomes conventionally associated with direct, intentional, etc. causation. More on the emergence and development of different form-meaning pairings follows in Section 3.5.

Table 3.2. Different types of causation in the typological sample, the meaning of the less compact form. The information about the languages is provided in Appendix 1.

The less compact form expresses more/more often…	**Languages in the sample**	**Number of languages**
Indirect causation	Ma'di, Gumuz, Humburi Senni, Kayardild, Kusunda, Chimariko, Hebrew, Humburi Senni, Basque, Betta Kurumba, Yukaghir (Kolyma), Creek, Japanese, Urarina	15
Directive causation (as opposed to manipulative)	Diyari	1
Agentive or volitional Causee	Aguaruna, Cherokee, Lakhota, Motuna	4
Causation by communication (e.g. ordering)	Trumai, Great Andamanese	2
Mediated causation	Hindi	1
Factitive causation with a human intermediary	Noon	1
'Indefinite' causation (have something done) with a backgrounded Causee	Ainu	1
Weaker integration of events	Apinayé, Takelma	2
Distant causation (vs. contact causation)	Nivkh	1
'Mild' causation	Caddo	1
Causee as beneficiary	Tubu/Dazaga	1
Formed from dynamic verbs, actions (vs. states)	Wappo, Garrwa, Finnish	3
Letting, permissive (vs. making, factitive)	Ma'di, Kusunda, Finnish, Trumai, Hebrew, Kusunda, Teribe	7
Forceful	Basque, Wappo, Ik, Finnish	4
Non-volitional, not intentionally acting Causer	Tidore, Adang, Apinayé	3
Involved Causer	Cavineña	1
Distributive causation	Yukaghir (Kolyma)	1
Iterative causation	Yukaghir (Kolyma)	1
'Resultative' causation (keep X in a certain state)	Yukaghir (Kolyma)	1
Ballistic causation	Hup	1

Note that some languages are mentioned in Table 3.2 more than once because they have multiple pairs of causatives that can be compared.

Moreover, I have encountered several combinations of features of the less compact form:

- Making/letting/compelling (Khoekhoe: Khoe-Kwadi)
- Permissive and not implicative (Waimiri-Atroarí: Cariban)
- Permission or coercion (Lahu: Sino-Tibetan, Slave: Na-Dene)
- Indirect and/or non-implicative (Korean: isolate)
- direct intentional vs. indirect and/or unintentional (Indonesian: Austronesian, Motuna: East Bougainville, Filomeno: Totonacan)
- 'Weak' causation with the semantics of motion, i.e. send (Yagua: Peba-Yaguan)

In addition, in Manambu, a Sepic language, verbal cause-effect compounds express the specific type of causing event, e.g. *vya-puti-* (hit-fall.off) 'shake something off by hitting, e.g. dust from a mat or a sheet'. Compare those with caused motion constructions and resultative constructions in English, e.g. *throw the ball into the street* or *paint the door green* (e.g. Hampe 2011). Since specific types of causation should be less frequent and therefore less compact than non-specific, generic causation types, this supports our predictions. There are also two violations of the form–meaning correlation related to (in)directness of causation (Kayardild: Tangkic, and Mutsun: Penutian). They are discussed in the next chapter.

Therefore, the semantic differences between more and less compact forms cannot be reduced to (in)directness. This finding is corroborated by a corpus-based study (Levshina 2016) of causatives in fifteen European languages. It shows that the formal distinction between analytic and lexical causatives is associated with numerous semantic and syntactic distinctions, which cannot be reduced to (in)directness alone.

3.4. Frequencies of different types of causation

3.4.1. Corpus data

This section presents corpus data, which show that the semantic functions expressed by more compact forms are more probable than the ones expressed by less compact forms. In order to

obtain the probabilities, I took different spoken corpora in three languages, English, Lao and Russian.

For English, I took samples of text from fourteen spontaneous informal conversations in the Santa-Barbara Corpus of Spoken American English (Du Bois et al. 2000–2005). I searched manually for all kinds of causative meanings, where one could distinguish the Causer, the Causee and the causing and caused effects, at least, potentially. The constructions were transitive verbs (lexical causatives, such as *break* and *kill*), syntactic causatives (e.g. *make/let/force/order/help* + (to) Infinitive), and resultative constructions (e.g. *keep* X in a certain state). In total, I obtained 205 causative situations.

For Lao, I took the transcripts of Enfield's (2007) dialogues from the appendix of his grammar of Lao. These are five dialogues about family, agriculture, fishing and work. I found only 60 instances in the entire corpus.

For Russian, I took one large text from Zemskaja and Kapanadze (1978), which contained the transcripts (with additional contextual information) of one day in a Soviet family. It includes all interactions between the wife, the husband, their son and the husband's mother during one day. It gives an idea of a typical linguistic behaviour of educated Russian speakers in the 1970s. The family members speak about food, health, childcare and home-making. The total number of causative examples was 90.

3.4.2. The coding schema

The examples from the corpora were coded for several variables, which represent different types of causation expressed cross-linguistically by the less compact form. First, there is a block of variables related to (in)directness. As shown in previous research, these parameters are highly correlated (Levshina 2016).

1. 'No Overlap': There is no temporal or spatial overlap between the Causer's actions (or lack thereof) and the event that corresponds to what happened with the Causee. Example: *John caused Bill to die on Sunday by stabbing him on Wednesday.*

2. 'Human Causee': The Causee is human. Example: *John caused Bill to die,* or *John killed Bill*.

3. ‘Controlling Causee’: The Causee is in control of the caused event. In other words, the Causee can choose, in principle, whether to perform what the Causer makes or lets the Causee do. Example: *Jane had her students write long term papers,* where the students could choose, theoretically, whether they comply or not (and bear the consequences).

4. ‘Caused Action’: ’The Causee performs an action (rather than gets into or keeps being in a certain state). Example: *The general had her troops run 10 miles.*

5. ‘Communication’: The Causer uses only communication in order to achieve the outcome. Example: *John got his grandparents to sponsor his album.*

6. ‘Human intermediary’: The situation implies a human intermediary, who must participate, so that the caused event could take place. Examples: *She made him dig a hole in the ground* or *She bought a new laptop* (i.e. with the help of a shop assistant).

7. ‘Letting’: The causing event is permissive (letting): The Causer’s action contains the verbs *let, allow* or *permit* or their equivalents in the other languages. Example: *He let the child play in the yard.*

The remaining types of events represent other types of causation, which are associated with less compact forms in the typological data.

8. ‘Forceful’: Forceful causation, as opposed to natural. Causation requires more effort from the Causer than usual. It is also possible to paraphrase the Causer’s action with ‘force’. Example: *Jane forced Peter to sign the agreement.*

9. ‘Non-intentional’: The Causer affects the Causee unintentionally, or is incapable of intentional actions (e.g. inanimate). Example: *John broke the window when he was playing football.*

10. ‘Involved Causer’: The Causer is involved in the caused event. In other words, the Causee performs the caused action or is in the caused state together with the Causee. Example: *Jane brought her friends to the party* (and came herself).

11. ‘Causee Beneficiary’: The Causee benefits from the caused event. Example: *John fed the child.*

12. ‘Non-implicative’: There is a possibility that the caused event doesn’t actually happen. Example: *John ordered Bill to surrender* (but Bill didn’t do it).

13. ‘Distributive’: The caused event occurs several times, each time with a different Causee. Example: *John killed Bill on Wednesday and Jane on Thursday*.

14. ‘Keeping’: The event can be paraphrased as ‘keep X in a certain state’. Example: *Jane kept all her savings under the mattress.*

15. ‘Iterative’: The causation repeats several times (with the same Causee). Example: *The gamer had to kill the villain again and again, until the villain had no more lives left.*

16. ‘Assistive’: The Causer helps the Causee to perform the caused event. Example: *John walked the child into the room.*

In a few cases, it was difficult to determine the value due to lack of additional context, but the proportion of missing values was never greater than 4% of the total number of examples in each of the three languages.

3.4.3. Quantitative results

Figure 3.3 presents the proportions of the functions, which are cross-linguistically expressed by less compact forms and which are related to indirectness of causation. One can see that none of them accounts for more than a third of all instances of causation in any of the corpora. This means that compact forms express more frequent functions. Interestingly, the Causee is more frequently human, controlling, acting and serving as an intermediary in Lao. This can be explained by the fact that the longest of the dialogues contains a discussion of employment, in particular, situations when the boss has the servant do something. Still, this causation is not sufficiently frequent.

Figure 3.3 shows the remaining nine variables. The relative frequencies of the remaining causation types are even lower. Therefore, less compact forms in all situations express less frequent events. The frequency of beneficiary Causees in the Russian data is relatively high because the language users often speak about childcare (feeding, dressing, putting to sleep, etc.).

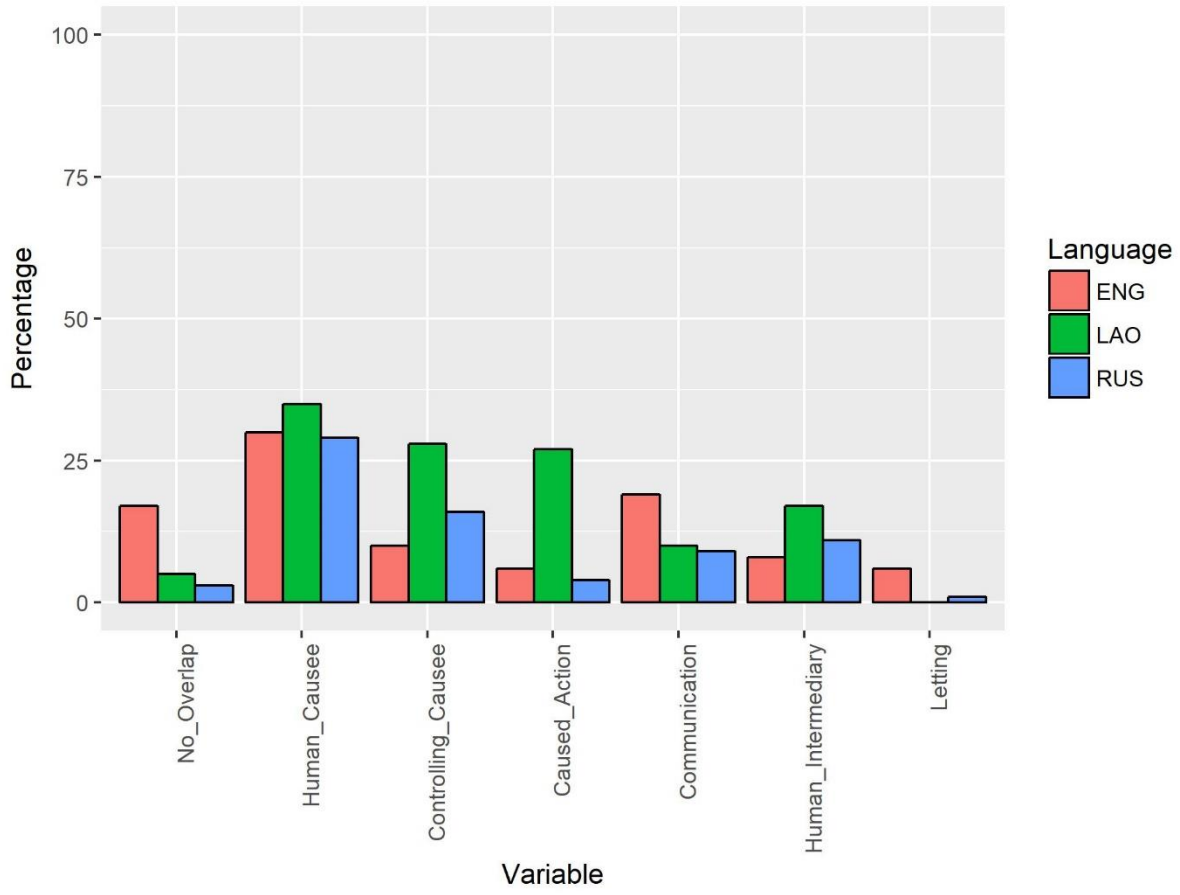

Figure 3.2. Percentage of the total number of causative situations in corpora of three languages: Frequencies of the features related to indirectness of causation

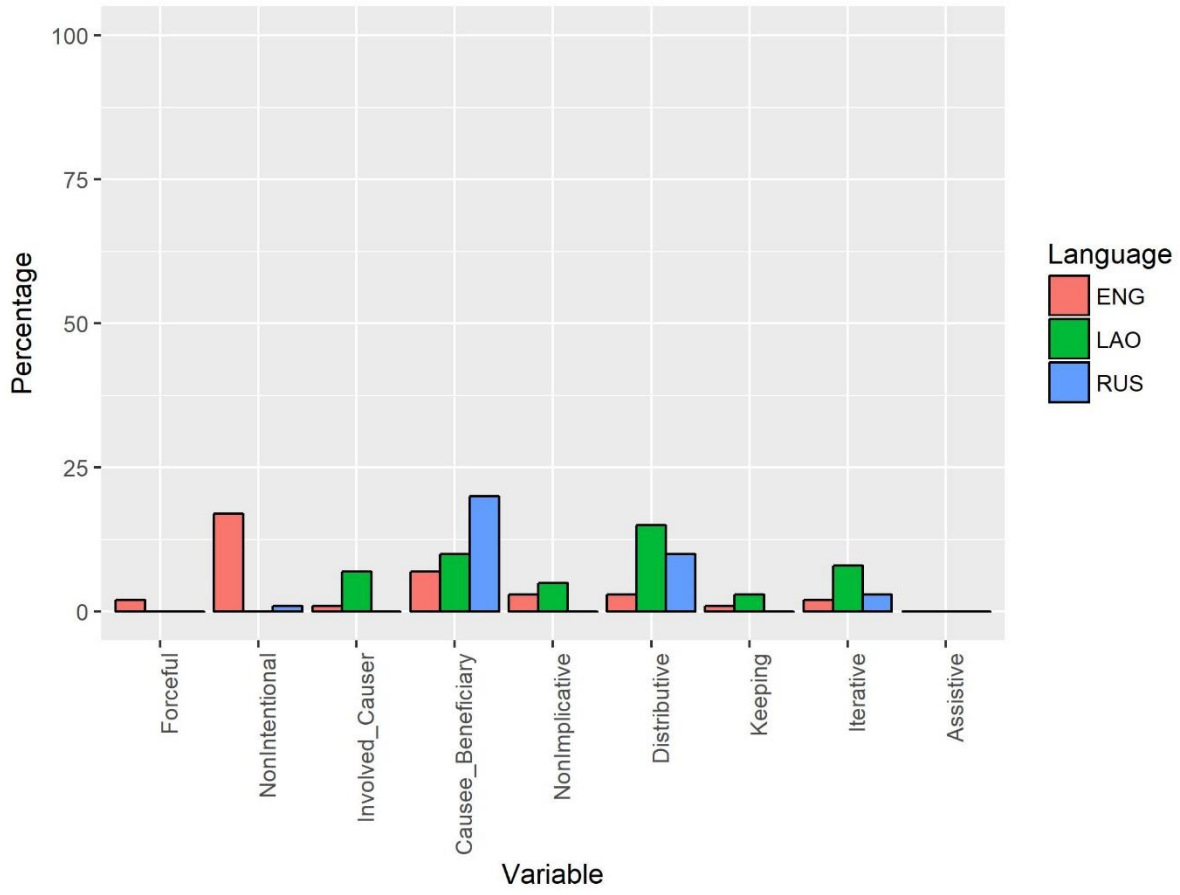

Figure 3.3. Percentage of the total number of causative situations in corpora of three languages: Frequencies of the other features

3.5. Possible diachronic scenarios

3.5.1. Diachronic sources and processes

There is not much direct evidence, unfortunately, how causatives have emerged and developed in different languages, especially as far as the more compact forms are concerned. Sometimes one can find information about the diachronic sources of causative constructions and their colexifications with non-causative expressions. The sources of causative markers and auxiliaries, that I found in the grammars are listed below:

- ‘make’ and ‘do’, e.g. suffix *-fy* in English, which comes from Latin *ficāre* ‘to do, make’; Dutch analytic causative with *doen* ‘do’;

- verbs of communication, e.g. ‘order’ (Trumai: isolate), ‘say’ (Skou: Skou) and ‘ask’ (Great Andamanese family);

- verbs of possession: ‘have’ and ‘get’ (English), ‘take’ (Hup: Nadahup), ‘give’ (Finnish: Uralic), ‘hold, grasp’ (Kayardild: Tangkic)

- motion verbs, e.g. ‘send’ (Yagua: Peba-Yaguan);

- position verbs, e.g. ‘stand’ (Hup: Nahahup);

- verbs of caused motion: ‘bring’ (Humburi Senni: Songhay), ‘put’ (Kayardild: Tangkic), ‘pull’ (Tubu/Dazaga: Saharan), ‘push’ (South Eastern Huastec: Mayan);

- abstract verbs, e.g. ‘cause’ (English) and ‘affect’ (Adang: Timor-Alor-Pantar) and ‘force’ (Ik: Eastern Sudanic), ‘treat in a certain way’ (Yuracaré: isolate);

- verbs of physical contact: ‘hit’, ‘step on’ and ‘bite’ (Manambu: Sepic);

- instrumental and manner affixes, e.g. ‘by hand’ (Northern Paiute: Uto-Aztecan) or ‘using a sawing action’ (Nishnaabemwin: Algic).

In addition, causative morphemes coincide with the following grams:

- transitivizers and verbalizers (Yapese: Austronesian);

- directional (allative) case markers, e.g. ‘towards’ (Ijo/Izon: Niger-Congo);

- intensifying affixes (e.g. Chichewa: Niger-Congo);

- aspectual affixes, e.g. punctual action (Mari: Uralic);

- passive markers (e.g. Southern Min: Sino-Tibetan);

- applicatives (e.g. Uto-Aztecan languages);

- benefactive affixes (e.g. Khasi: Mon-Khmer);

- complementizers, including purpose (e.g. Thai:Tai-Kadai).

In most cases, it is very difficult or even impossible to determine the direction of historical development. Still, one can find some common diachronic mechanisms, which can be explained by the Principle of Communicative Efficiency.

First of all, unidirectionality of grammaticalization means that more compact causatives emerge from less compact ones. If a construction becomes more frequent, it also becomes more probable. Due to the Low-Cost Heuristic, it becomes more compact (see Section 2.2). Moreover, shortness of coding leads to bound expression, because, as noted by Haspelmath (2008b: Section 6) short elements do not have enough bulk to stand on their own and need a host. An alternative explanation of low autonomy is that a more predictable causativizing element is already tightly associated with their host. This is why it is highly predictable in the first place. As a result of this high predictability, the causativizing element becomes reduced.[23]

Formal reduction can be accompanied by reanalysis. For instance, Song (1996) argues that the so-called purposive markers (in particular, those that express allative functions and represent a goal) become reanalyzed and grammaticalized as causative markers, while the element that expresses the causing event disappears.

Let us consider the development of the English causative with *make*. In old English, it was followed by *that*-clauses, as in the following example:

(17) (Old English Heptateuch, Exodus 96: 14; cited from Lowrey 2012)[24]

Ge habbaþ us gedon laþe Pharaone and eallum his folce and ***gemacod*** *þæt hig wyllað us mid hyra sweordum ofslean*

'You have made us hateful to Pharaoh and to all his people, and made them want to slay us with their swords'.

In Middle English and Early Modern English the *to*-infinitive was predominant, but the bare infinitive occurred, as well (Hollmann 2003: 166–167; Moriya 2017). Consider two examples from the King James Bible (1611):

(18) (Early Modern English, King James Bible, cited from Moriya 2017: 44)

[23] I thank Karsten Schmidtke-Bode for drawing my attention to this explanation.
[24] Heptateuch: Marsden, R., ed. 2008. *The Old English Heptateuch and Ælfric's Libellus de Veteri Testamento et Novo* (vol. I), EETS 330. Oxford: OUP

a. *And wherefore haue ye made vs to come vp out of Egypt, to bring vs in vnto this euil place?* (Numb. 20.5)

b. *And hee doeth great wonders, so that hee maketh fire come downe from heauen on the earth in the sight of men,* … (Rev. 13.13)

Moriya (2017) argues that the preference for the bare or *to*-infinitive is guided by various factors. One of them is *horror aequi* (the presence of *to* before make). Another one is the linguistic distance between *make* and the second verb (in particular, the length of the nominal phrase in-between). The marker is used when the environment is complex, according to Rohdenburg's principle of cognitive complexity (see Section 1.2.4 of Chapter 1). In terms of the Principle of Communicative Efficiency, one can say that long linguistic distance makes the interpretation of the second part of the construction as such less probable. Although the evidence for the semantic differences between the two variants is not conclusive and a lot of variation looks random, there are also some instances of the *to*-infinitive being preferred with willing Causees, which represent a less typical kind of causation.

Hollmann (2003: 166–167) argues that the high token frequency of the construction led to the erosion of the infinitival marker in causative *make*. We can hypothesize that this process was first observed in in more stereotypical causative situations and subschemata of the construction. There is some evidence that supports this account. At the moment, the causative *make* is used with the bare infinitive (e.g. *This makes me laugh*), with the exception of the passive form (e.g. *He was made to sit on an uncomfortable stool*). Due to its low frequency, the passive form has low predictability, and therefore remains the last bastion of the marked infinitive. Interestingly, the passive form of either the matrix verb or the infinitive was associated with the *to*-form in Late Middle English infinitival complements in general (Fischer 1995).

Importantly, the gradual disappearance of *to* after *make* has not happened in the other factitive causatives (*cause*, *force*, *get*, *persuade*, etc.). This difference between the *make*-causative and the other causative constructions seems to be efficient, as well. Hollmann (2003: 151–158) finds that the former score higher on semantic boundedness (which is operationalized as a weighted score based on such parameters as directness, intentionality, punctuality, etc.)

higher. I assume that situations which are associated with greater semantic boundedness by Hollmann and others are simply more frequent in discourse (see Section 3.4). Unfortunately, there is too little data about the development of the causative with *have*, which is used with the bare infinitive.

Another example of formal reduction is provided by Song (1996: 88). In Ijo (Izon), a Niger-Congo language, there is a causative suffix *-mọ*, which is also identical with the directive case marker. There are also cases when a separate lexical element *mie* is added, which expresses the causing event. Compare (19a) and (19b):

(19) (Ijo [Izon]: Niger-Congo. Song 1996: 88)

a. *áràú toboú mìe búnu-mo-mi*

she child make sleep-CAUS-ASP

'She soothed the child to sleep.'

b. *áràú toboú búnu-mo-mi*

she child sleep-CAUS-ASP

'She laid the child down to sleep.'

Song argues that the original purposive construction with two predicates (*mie* and the verb expressing the caused event) and the originally purposive marker *mo-* is giving way in this language to the morphological causative. The first predicate is normally omitted. The shorter causative in (19b) expresses more direct causation than the longer one in (19a). This can be explained by the Low-Cost Heuristic, as well: shorter forms represent more stereotypical causations.

As shown by Mithun (2002), causative morphemes can emerge as a result of semantic generalization and reanalysis of more specific, concrete meanings. For example, in diverse Northern American languages, there are numerous manner and means prefixes, e.g. doing things with hands, feet, teeth, a knife, by pressure, etc. For instance, the manner prefixes *yu-*, *pa-* and *ka-* described different hand motions in Lakhota. They are highly frequent, since language users do many things with hands, e.g. *bláya* 'be level, plain' – *yubláya* 'open, spread

out, unfold, make level'. It frequently formed pairs, which included non-causal and causal verbs. If an object is moved with a hand, it is easy to interpret the meaning as causation with the hand. Also, hands are unmarked instruments of human actions. As a result, they were reanalyzed as causative, and the notion of a hand movement has disappeared, as in the pair *bléza* 'clear' – *yubléza* 'make clear'. Note that this only happens with the most typical concrete meanings of the source construction.

The general grammaticalization path of causatives seems to be the following: from less direct and more marginal functions and analytic forms to more direct and typical functions and more compact forms. The general casual mechanism based on the Low-Cost Heuristic was proposed in Section 2.2. As a construction becomes more frequent, the causativizing element becomes more predictable, which leads to its shortening and boundedness. Moreover, shorter constructions also become associated with more stereotypical situations thanks to the hearer's inference based on the Low-Cost Heuristic. This is why the meaning of causatives tends to shift towards more direct causation (or another more typical kind of causation), as grammaticalization proceeds. A reduced construction, which formely expressed only indirect causation can be used in a reduced form to represent more direct causation, as in (19). The link of the expression with the more typical meaning leads to a further increase in frequency and greater reduction, and so on.

Let us now turn to the processes based on the High-Cost Heuristic. An efficient formal asymmetry is created when a new causative emerges with a longer form and less probable meaning than the old one. Consider Old Dutch. After the Germanic morphological causatives with the suffix *-ja* stopped being productive (possibly, due to the loss of transparency in umlaut), there remained many lexical (ex-morphological) causatives. In the 12-13th centuries, analytic causatives with *doen* 'do' and *laten* 'let' emerge (van der Horst 1998). The earliest instances of the *doen*-causatives express curative causation (i.e. having someone do something), which implies an agentive Causee, as in the following example:

(20) (Middle Dutch, van der Horst 1998: 56)

si sullen sin hus doen breken

they want his house do burst

'they will have his house broken down'

The first attestations of the construction with *laten* have a permissive sense:

(21) (Middle Dutch, van der Horst 1998: 64)

lat	*dise*	*arme*	*kinde*	*leuen*
let	these	poor	children	live

'let these poor children live'

Although the constructions later change their usage and semantics (probably, due to mutual alignment) (e.g. Verhagen and Kemmer 1997; Levshina 2011), the first attestations show that these constructions are used in the relatively infrequent functions, while the more frequent ones are performed by lexical causatives (e.g. *breken* 'break, burst', *weuen* 'weave', *spreiden* 'spread' or *leggen* 'lay').

Thus, if there is a novel causative expression, it is likely to begin with indirect causation or other non-stereotypical functions. This is supported by the pragmatic mechanism of the High-Cost Heuristic. If a novel, 'strange' expression is used, one is tempted to attribute to it less probable meanings. The speaker and the hearer know that there are some typical and untypical ways of making someone dead, such that $\mathbb{P}$ (meaning-typical) is greater than $\mathbb{P}$ (meaning-untypical). They also know that the costs of the new longer form are higher for the speaker than the costs of the default short form. Following the High-Cost Heuristic, the costlier expression promises the hearer more cognitive effects due to the lower probability of the intended interpretation.

If the less compact causatives become frequent, this may cause the more compact form to trigger a Q-implicature (see Section 1.2.3). In particular, this will mean that the causation expressed by the compact form is not indirect, not non-intentional, etc. Compare the causatives in Russian and German. In Russian, the lexical causatives can still be used to express indirect curative causation:

(22) *Šnur vstavil zuby za 250000$.*[25]

Sh. inserted teeth for 250000$

'Shnur (a musician) had his teeth replaced (lit. replaced his teeth) for 250,000$.'

The exact meaning is inferred from the context. In some cases, one can name the actual Causee (at the dentist's) or the price, as in this example, to indicate that this was a service provided by a professional, but this information is not obligatory. In my opinion, this function of lexical causatives can be expressed by the low frequency of synthetic causatives, expressing indirect, forceful or permissive causation in Russian. The Q-implicature is not inferred.

In contrast, German has more frequent synthetic causatives, in particular, the causative with *lassen* (Levshina 2015). In the example below, the analytic auxiliary *lassen* cannot be omitted if one refers to the standard procedure performed by a dentist:

(23) *Wollen Sie sich während Ihres Aufenthalts in Rumänien dritte Zähne einsetzen *(lassen)?*[26]

The lexical causative *einsetzen* cannot be used to express indirect curative causation because the Q-implicature is stronger. The speakers of German are aware of the more conventional alternative with *lassen*. This blocks the use of the lexical causative.

3.5.2. Efficient storage and retrieval or efficient communication?

Haspelmath (2008b) argues that lexical causatives of Haiman's type Z, which he calls suppletive, exist because they are highly frequent and therefore can be stored in the memory. For instance, English has unproductive morphological causatives *sadd-en* 'make sad', *wid-en* 'make wide', *hard-en* 'make hard', but it is only for high-frequency adjectives like *good* and *small* that it has suppletive causatives (*improve* 'make good', *reduce* 'make small'). This

[25] https://youtu.be/P2AHqvdCwGQ (last access 12.11.2018).
[26] https://www.siebenbuerger.de/zeitung/pdfarchiv/suche/erhard%20gillich/ (last access 12.11.2018).

minimizes the costs of storing and extracting the constructions from the memory (Haspelmath 2008b: Section 6).[27]

I will allow myself a literary intermezzo here. Jorge Luis Borges has a short story *Funes the Memorious*, which tells about a young man who, as a result of a head injury, could remember anything and forget nothing. He created his own system of enumeration, in which each number up to at least 24,000 is given an arbitrary name, and was thinking of a language that could give an arbitrary name to any object at any given point of time (similar ideas were expressed by John Locke). This came, however, at the cost of being unable to make generalizations and abstractions, that is, to ignore the subtle individual differences between objects and events. In real life, a fully suppletive language of this type is impossible due to a multitude of reasons, even if we could store as much information as we wanted. First of all, if there are no mental categories, which come as a result of generalization, there can also be no linguistic ones. If there are no objects, for example, one cannot think of individual labels for them. No grammar would be possible, either, because it represents a result of generalizations, as well. Moreover, even if the person settles on some very fine-grained categories and low-level generalizations, it will be impossible to coordinate the labels for these categories with other people because the categories are based on highly individual experience.

Can the existence of frequent and simple vs. rare and compositional causatives be explained by the fact that an average speaker does not have a memory like the protagonist of Borges' story? This does not sound very plausible. We store huge amounts of information in our constructicons, including very rare form-meaning pairings. Adding a few thousand verbs would not be a problem.

Let us consider a different point of view. As was already discussed in Chapter 1, Hawkins subsumes lexicalization and non-compositionality under the principle "Minimize Forms":

> The more frequently selected properties are conventionalized in single lexemes or unique categories and constructions (…) Less frequently used properties must then be expressed through word and phrase combinations and their meanings must be derived

[27] It is a question whether one should take the absolute frequency of the base form or of the derived form in order to explain suppletion. There is evidence that it is the absolute frequency of the derived form that makes it resistant to analogical change (Bybee and Thompson 1997; Corbett et al. 2001).

> by semantic composition. This makes the expression of more frequently used meaning shorter, that of less frequently used meaning longer… (Hawkins 2014: 17).

This account does not explain, unfortunately, how this formal difference emerges in the first place.

I believe there are two main factors at play. The first one is the Principle of Communicative Efficiency. According to the Low-Cost Heuristic, the simple and short forms will gravitate towards expressing direct and intentional causation, which belongs to the core of human experience. We change the position and state of surrounding objects all the time. Following Quine's (1969) famous thought experiment, let us imagine that a hunter says *bara gavagai* and proudly shows a dead rabbit. Imagine we know from our previous encounters with the language that *gavagai* stands for 'rabbit', so *bara* should represent some causal action. It is likely, given the cues, that the speaker means 'I killed the rabbit'. It is less likely that he means some indirect causation, e.g. 'I caused the rabbit to die, e.g. by making it fall from a cliff'. This is the Low-Cost Heuristic in action. Imagine this word becomes spread in the community. Typically, *bara* will mean 'kill (intentionally and directly)', but occasionally one can also refer to less direct or non-intentional ways of killing. Usually, it is clear from the context which type of killing is meant. Imagine now that due to some reasons it becomes important to express this distinction. A possible reason is easy to think of. For instance, if one is charged with killing a rabbit that belonged to someone else, the distinction between different types of causation (especially intentional vs. accidental) may be crucial for resolving the conflict and deciding on the form of punishment.

Which kind of expression would the speaker use then, in order to say that the rabbit was killed accidentally? One cannot use the default *bara* because of the implicatures of the default, intentional killing, and there is no special expression for accidental causation yet. There are two options left for the speaker. One can invent a new simple word or recycle old ones in a compositional periphrastic expression, e.g. "I shoot, it dies". The choice depends on the second crucial factor – the desire of being understood (cf. the maxim "Talk in such a way that the other understands you" in Keller 1994: 98). The chances of being understood are higher when one recycles already existing form-meaning pairings than when one coins brand new labels. Recycling also allows one to spare the efforts involved in negotiating the meaning of new expressions. Therefore, the compositional expression will be preferred.

Since there is a causative *bara* ‘kill’, this new expression will be interpreted, based on the High-Cost Heuristic, as signifying something different from intentional and direct killing. Later this expression may become conventionalized and reduced, becoming a typical syntactic or morphological causative.

To summarize, the Principle of Communicative Efficiency plays an important role in the emergence of the formal asymmetries like *kill* vs. *cause to die*. The Low-Cost Heuristic makes the shorter and simpler expression gravitate towards the most typical meaning, while the High-Cost Heuristic is responsible for the association between the non-stereotypical interpretation and the longer (compositional) expression. The compositionality of the longer causative is not explained by memory limitations, but rather by the fact that it is advantageous to combine already existing conventional signs instead of negotiating the meaning of novel ones. In a way, this is also a manifestation of communicative efficiency.

Note that in more recent work on causative constructions, Haspelmath et al. (2014) do not focus on the relationship between usage frequency and the entrenchment of suppletive forms, and propose instead the following causal link, which is similar to the account developed in this chapter:

(24) factor X → usage frequency → predictability → short form

where factor X stands for any causal factor that can explain usage frequency. This causal chain explains why one observes a robust cross-linguistic correlation between the degree of spontaneity of events and the participation of the corresponding verbs in the causative or anticausative alternations. Compare the expressions for the causal and non-causal ‘break’ and ‘freeze’ in Turkish:

(25) (Turkish: Uralic, from Haspelmath 1993: 119):

a. *don-mak* ‘freeze intr.’ > *don-dur-mak* ‘freeze tr.’

b. *kir-ıl-mak* ‘break intr.’ < *kir-mak* ‘break tr.’

The more spontaneous events, which do not involve typically an external agent, like 'freeze', more frequently participate in the causative alternation, as in (24a). That is, they have basic non-causal forms and derived causative forms. An even more radical case is agentive events, such as 'dance'. All languages express the corresponding causative event as 'make/cause X dance'. In contrast, verbs that are associated with an external agent, e.g. *break*, typically participate in the anticausative alternation, as in (25b). They tend to have basic causal forms, and derived non-causal forms.

It is difficult to understand why this case should be any different from the opposition between lexical (or suppletive) and compositional causatives. Haspelmath et al. (2014) argue that frequently used, or predictable meanings are expressed by shorter forms than rare meanings, due to the considerations of efficiency. They claim that human languages have recurrent diachronic mechanisms which create such efficient form-meaning patterns. Yet, it is easy to see that the shorter forms in the causal and non-causal pairs, as in (25), are often suppletive and need to be learned as a whole, similar to the lexical causatives in the pairs like *kill – cause to die*. From this follows that the phenomena should have the same underlying causal mechanism – that is, the one based on the Principle of Communicative Efficiency and the desire of being understood.

3.6. Summary

This chapter has introduced some popular accounts of formal and semantic variation of causative constructions, focusing on the relationships between their form and meaning. Special attention has been paid to the iconic relationship between formal integration, or compactness, and (in)directness of causation. The following claims have been made:

1) Cross-linguistic differences in the semantics of more and less compact forms of causative constructions in languages of the world cannot be reduced to the (in)directness distinction. Other features, such as forcefulness of causation, intentional actions of the Causer, etc. are observed, as well. Dixon's (2000) findings were supported by the data from a sample of 59 typologically diverse languages, and some amendments to his account have been made.

2) The data from English, Lao and Russian informal spontaneous spoken dialogues demonstrate that the features of causative situations expressed by the less compact forms are less frequent than the features represented by the more compact forms.

3) From this follows that the iconic relationships between the form and the meaning of causatives, which was proposed by Haiman (1983), can be best explained by the frequency asymmetries. The latter can also explain the other differences between more and less compact causatives, such as natural vs. forceful causation and other types. The more predictable meanings are normally expressed by short forms due to the Low-Cost Heuristic, whereas language users tend to use costlier forms to express less probable meanings based on the High-Cost Heuristic. This explains why lexical causatives tend to express the default, i.e. direct and intentional causation, and derived forms express less stereotypical events.

Arguing for efficiency, I do not want to exclude iconicity completely. It may be that iconicity matters at the early stages, when a new causative expression is coined. For example, the temporal or spatial distance between the causing and caused events may be emphasized by presenting the cause and effect by two independent clauses, as in the following example:

(26) *He pressed the button and down in the control room all the screens suddenly came to life.*[28]

However, when a novel expression goes into the grammaticalization grinder, efficiency becomes more important.

The next chapter will zoom in on the distinction between direct and indirect causation, which plays a central role in the iconicity-based explanations, to demonstrate that efficiency provides a better explanation of the form-meaning mapping than iconicity. I will also discuss another formal parameter, namely, productivity, which has been proposed as a factor correlated with (in)directness.

[28] https://books.google.de/books?isbn=1476712964, p. 199 (last access 25.11.2018).

Chapter 4. Direct and indirect causation: Finding the best explanation

4.1. Aims of this chapter

This chapter continues the discussion of causative constructions.[29] The previous chapter introduced the iconicity-based and efficiency-based accounts of cross-linguistic variation of causatives. According to the iconicity-based account, the correlation between the formal integration and (in)directness of causation is explained by iconicity of cohesion (see Section 3.2 in Chapter 3 for more information). The degree of formal integration can be interpreted in terms of autonomy or formal distance between X and Y, where X represents the causing event, and Y stands for the caused event, e.g. *cause* and *to die*. Autonomy stands for the independence of X and Y, while distance reflects the number of intervening elements between X and Y. Lexical causatives like *kill* have both zero autonomy and zero distance, because X and Y are perfectly fused in the root *kill*. Compare that with an example of a syntactic causative. The causing and caused event in the sentence *John caused Bill to die* are represented by two relatively autonomous units, i.e. the words *cause* and *die*, and are separated by the past tense morpheme *-ed*, the proper name Bill and the particle *to*.

The other view explains this correlation, as well as the associations between forms and other functions, as a result of efficiency: shorter forms, such as *kill*, usually express more probable causative types, including direct causation, while longer forms, such as *cause to die*, tend to represent less probable meanings, including indirect causation. This was supported by the relative corpus frequencies of different causation types presented in the previous chapter.

There exists yet another explanation, which was proposed by Shibatani and Pardeshi (2002: Section 5), who argue that the semantic distinction between direct and indirect causation closely corresponds to the degree of productivity of constructions. Moreover, (in)directness correlates with productivity stronger than with such well-established formal classes as lexical, morphological and syntactic causatives:

[29] The findings presented in this paper have been published in Levshina, Natalia. 2018. Finding the best fit for direct and indirect causation: a typological study. *Lingua Posnaniensis* 58(2) (2016): 65–83.

> Cross-linguistically, productive forms align (whether they are morphological or periphrastic) in expressing indirect causation, and lexically restricted forms align (whether they are morphologically unanalyzable or morphologically complex) in expressing direct causation. (Ibid.: 112)

One piece of evidence comes from Japanese morphological causatives. Consider the verb *oros-* 'bring down' from *ori-* 'come down', which represents a non-productive construction, in contrast with *ori-sase-* 'cause to come down', which is formed with a productive suffix *-(s)ase.* The non-productive form expresses direct causation, similar to lexical causatives, whereas the productive form conveys indirect causation (Comrie 1981: 163).

As another illustration, one can take the causative prefixes *a-* and *as-* in Amharic (Amberber 2000). The former only applies to intransitive unaccusative verbs (e.g. verbs of state, change of state and motion, e.g. 'exist', 'melt', 'grow', 'enter') and transitive verbs of ingestion (e.g. 'eat' and 'drink'), but not to unergative verbs (e.g. 'dance' or 'laugh'). It expresses situations when the Causer is directly involved in the causation (Ibid.: 317-320), as in (2b), where the Causer transports the Causee. The causative prefix *as-* applies to both transitive and intransitive verbs of all classes and expresses causation types when the Causer is not directly involved and can simply issue an order or permission, as in (2c). Again, the (in)directness distinction correlates with productivity. Since both causatives are morphological, the traditional three-way classification does not correlate with this semantic distinction.

(2) (Amharic: Afro-Asiatic, Amberber 2000: 320)

a. *aster* *wət'ə-čč*

A. exit+PERF-3F

'Aster exited.'

b. *ləmma* *aster-ɨn* ***a-****wət'ə-t*

L. A.-ACC CAUS-exit+PERF+3M-3F.OBJ

'Lemma took Aster out (as in 'out of the house').'

c. *ləmma* *aster-ɨn* ***as***-*wət't'a-t*

L. A.-ACC CAUS-exit+PERF+3M-3F.OBJ

'Lemma made/let Aster exit.'

These and other examples (cf. Shibatani and Pardeshi 2002: Section 5) demonstrate that more productive morphological causatives often express indirect causation, similar to syntactic causatives, and less productive ones express direct causation, similar to lexical causatives. However, they belong to the same traditional class of morphological causatives. The degree of grammatical autonomy and the formal distance between the root and the affix also remain the same, as follows from the examples. Thus, productivity may be more directly aligned with the (in)directness distinction than autonomy and formal distance. According to Shibatani and Pardeshi, this alignment is a result of grammatical change. However, iconicity plays a role, as well, in that the strength of connection between the non-causal root and the causativizing element correlates with the strength of the conceptual integration of the causing and caused events.

Thus, the previous research suggests that direct and indirect causation can be correlated with formal distance, autonomy, productivity and length, which represents here the amount of effort. However, to the best of my knowledge, these hypotheses have not been tested and compared systematically on typologically diverse data. The aim of this study is to fill this gap and to investigate whether these formal parameters correlate with (in)directness of causation in different languages, and to what extent. To answer these questions, I will use the same sample of languages from different language families that was introduced in the previous chapter.

The rest of the chapter is organized as follows. First, I describe the typological data and variables in Section 4.2. Next, Section 4.3 presents the answers to the research questions based on quantitative analyses of the typological data. Finally, a summary and a discussion of the results are offered in Section 4.4.

4.2. Typological data

4.2.1. Language sample

For the purposes of this study, I used the same data set as the one presented in the previous chapter. The list of languages is available in Appendix 1. In 46 languages, (in)directness or similar semantic distinctions were mentioned as a distinctive semantic parameter of two or more different causative constructions. Each language represents one language family, based on the genetic classification from the World Atlas of Language Structures online (Dryer and Haspelmath 2013). The languages represent six major geographical and linguistic areas (Africa, Australia, Eurasia, North America, Papua/Austronesia and South America). The geographical distribution of the languages is shown in Figure 4.1. The semantic and formal properties of the causative constructions were taken from reference grammars and research articles. These differences are described in Sections 4.2.2 and 4.2.3.

Figure 4.1. Geographical distribution of the languages in the sample, where (in)directness or closely related semantic distinctions were found

4.2.3. Semantic parameters

The data points in my data set were contrasting pairs of causative constructions. If the data are available for two causative constructions in a language, then the language has one contrast between Cx1 and Cx2. If three constructions are described, there are three possible contrasts between Cx1 and Cx2, Cx2 and Cx3 and Cx1 and Cx3. For four constructions, six contrasts are possible, and so on.

From all possible semantic distinctions between the constructions in a language, I selected only those that can be interpreted as direct vs. indirect causation and the degree of integration of the causing and caused events or closely related functions. The full list of the distinctions used in the grammars and research papers, which I interpreted as direct vs. indirect, is as follows (see Table 3.2 in the previous chapter for the examples of languages):

- direct vs. indirect causation;
- strong vs. weak integration of the causing and caused events, separability of events;
- manipulative vs. directive causation;
- contact vs. distant causation;
- direct vs. mediated causation;
- the Causee as non-controlling undergoer vs. controlling agent (and therefore the main source of energy);
- default vs. ballistic causation;
- factitive vs. permissive causation;
- caused state (or change of state) vs. caused activity;
- default causation vs. causation with human intermediary;
- default vs. curative or 'indefinite' causation;
- general vs. 'mild' or 'weak' causation;
- default vs. caused by ordering X to do Y;
- implicative vs. non-implicative causal relationships.

As an illustration, consider two causative constructions in the Amur dialect of Nivkh (a Paleosiberian isolate) in (3). One of the constructions consists of a non-productive causative suffix *-u* and expresses contact factitive causation, as in (3b), while the other contains a

productive suffix *-ku/-ɣu/-gu/-xu* and usually expresses distant factitive or permissive causation (Nedjalkov and Otaina 2013: 133), as in (3c).

(3) (Nivkh: isolate, Nedjalkov and Otaina 2013: 234)

a. *Lep ʈ'e-ḍ.*

bread be.dry-IND

'The bread dried up.'

b. *If lep+se-**u**-ḍ.*

s/he bread+be.dry-CAUS-IND

'He dried up the bread' (for dried crusts).

c. *If lep+ətu-doχ q'au-r* ʈ'e-***gu***-*ḍ.*

s/he bread+cover-SUP not.be-CONV:NAR:3SG be.dry-CAUS-IND

'Not covering the bread, he let (it) dry up.'

A less common semantic contrast is found in Ainu, as shown in (4), where (4b) is an example of the causative construction which expresses 'normal' causation, whereas (4c) is an illustration of so-called indefinite causation, when the Causee is not known or is omitted due to politeness reasons. The latter type can be perceived as less direct because the Causee is not affected (cf. Kemmer and Verhagen 1994). This type is also known as curative causation. Similar constructions exist in English, namely, *have/get* smth. *done* (*by* smb.), e.g. *I've had my hair cut*.

(4) (Ainu: isolate, Tamura 2000: 214)

a. *e* 'to eat'

b. *é-re* 'to have (someone) eat, feed'

c. *e-yar* 'to have something eaten'

Often, a contrast involves more than two fine-grained distinctions. This is not surprising, since the semantic parameters of causatives are often strongly correlated (Levshina 2016). Example

(5) from a Cariban language Waimiri-Atroarí illustrates a combination of the factitive/permissive distinction and implicativity. The causative suffix *-py* in (5a) expresses factitive causation, whereas the periphrastic construction with *injaky* 'let/permit' and particle *tre'me* shown in (5b) expresses permission. In addition, causation in (5b) is non-implicative, which means that the causative does not express whether the caused event actually happens or not. This serves as an indication of weaker event integration (Givón 1980).

(5) (Waimiri-Atroarí: Cariban, Bruno 2003: 100, 103)

a. *Ka k-yeepitxah-**py**-pia.*
3PRO 1+2OBJ-laugh-CAUS-IM.P
'She/he made us laugh.'

b. *A ka m-**injaky**-piany wyty ipy-na **tre'me**.*
1PRO ? 2OBJ-permit/let-REC.P meat look for-? PART
'I permitted you to/let you leave to hunt.'[30]

As was already discussed in the previous chapter, the authors often do not specify the nature of causation, or call it 'general' or 'default', treating the corresponding construction as a grammatical tool for increasing valency. For example, one can find such distinctions as default vs. permissive, or default vs. curative causation. As was argued in the previous chapter, if the non-default construction is frequent, then the Q-implicature will be derived, which precludes the non-default interpretation of the default construction. If the non-default construction is infrequent, and the Q-implicature is weak, the default construction will be more frequently associated with the stereotypical causative situations, based on the Low-Cost Heuristic. Even if two constructions are interchangeable to some extent, the speaker will be less likely to use the 'default' construction if there is a more specific alternative that expresses the intended meaning. This is why such constructional pairs are also counted as instances of the (in)directness distinction.

In total, I found 74 contrasts related to (in)directness in 46 languages. Other semantic distinctions, which are not considered in the present study, involve forceful, unintentional, distributive, iterative causation and other types. They were discussed in the previous chapter.

[30] The author of the cited grammar used question marks to gloss some units, as in this example.

4.2.3. Formal parameters

All constructional pairs that have the semantic distinctions listed in the previous section were coded for the four formal parameters mentioned in Section 4.1: distance, autonomy, productivity and length. The coding was relative, rather than absolute. The construction that expresses (more) direct causation (including the features from the list in Section 4.2.2) is referred to as DIR-Causative, and the construction that represents (more) indirect causation, is represented here as INDIR-Causative. This means, for example, that I did not code the length of constructions DIR-Causative and INDIR-Causative separately, but only coded whether DIR-Causative was shorter than INDIR-Causative, longer than INDIR-Causative or equally long as INDIR-Causative. The contrasts between DIR-Causative and INDIR-Causative were coded for the four formal variables using the guidelines described below.

1. **Relative distance**. This variable reflects the differences in the linguistic distance between the elements representing the causing and caused events in DIR-Causative and INDIR-Causative. The formal criterion is the number of phonological segments (i.e. phones or phonemes) in the in-between elements, including affixes and clitics and autonomous words, which are obligatorily used between the elements representing the cause and effect. According to the iconicity-based account, we can expect the elements in DIR-Causative to be less distant than those in INDIR-Causative (i.e. DIR-Causative < INDIR-Causative).

2. **Relative autonomy**, which reflects the differences in the degree of autonomy of the elements representing the causing and caused events in DIR-Causative and INDIR-Causative. It is similar to bondedness of a sign, or "the degree to which it depends on, or attaches to, (…) other signs" (Lehmann 2015: 131). To determine the level of autonomy, I used the following cline:

(6) one morpheme < morphemes in a word < clitic + host < parts of one verbal phrase (monoclausal) < clauses in a sentence (biclausal)[31]

[31] I'm aware of the fact that the categories in (6) are problematic as comparative concepts for language comparison. However, the comparisons made in this study were only done within one specific language. I relied on the language-specific categories used by the authors of the grammars in the hope that those categories represent the differences in the degree of autonomy adequately.

The minimal autonomy is observed in lexical causatives, such as *kill*, *break* or *raise*. The causing and caused events are merged in one word or morpheme. In morphological causatives, the non-causal verb and causative morpheme are more autonomous. In syntactic causatives, such as *cause X to die*, the degree of autonomy is even higher because the words or clauses expressing the causing and caused events are more autonomous units. Again, the iconicity-based account predicts that DIR-Causative will tend to be less autonomous than INDIR-Causative (i.e. DIR-Causative < INDIR-Causative).

3. **Relative productivity**, which represents the differences in productivity between DIR-Causative and INDIR-Causative. Productivity is the ability of a unit or pattern to freely combine with other units. For example, lexical causatives display zero productivity, whereas periphrastic expressions like *John caused Bill/Jane/Mary to run/die/go…* display high productivity. Some morphological causative patterns can be more productive than others, as the examples from Amharic and Japanese in Section 4.3.1 demonstrate. A common instance of productivity asymmetries is when some causatives can be formed with all verbs, and some only with intransitives (cf. Dixon 2000). Following Shibatani and Pardeshi (2002), one can expect DIR-Causative to be less productive than INDIR-Causative (i.e. DIR-Causative < INDIR-Causative).

4. **Relative length**, which tells about the difference in the length of DIR-Causative compared to INDIR-Causative. This parameter represents the amount of effort and is directly related to efficiency. Many different operationalizations of length are possible, including very sophisticated ones (cf. Bybee et al. 1994: Ch. 4, who consider vowels to be longer than consonants, also distinguishing between long and short consonants and vowels). When available, relative length was based on the number of segments in grammatically equivalent forms of the same verb. For example, Filomeno (Totonacan) has a DIR-Causative with the prefix *maa-*, e.g. *maa-xiksw-í* 'I asphyxiate you' (directly), and an INDIR-Causative with the prefix *maqa-*, e.g. *maqa-xikswá* 'I make you asphyxiate' (indirectly, e.g. by making you laugh while you eat) (McFarland 2009: 155). In this situation, one can directly compare the number of segments in the words. When such immediate contrasts were not available, I compared the number of segments in the causative morphemes or auxiliaries and other obligatory elements, such as complementizers or finite markers on the auxiliary in the examples. This simple approach, however, leads to the same results as when I took into account the difference between

vowels and consonants and their length, following the more sophisticated method employed by Bybee et al. (1994). The reason is that the differences in length are rather conspicuous. When there were different allomorphs, I took the average length. Following the Principle of Communicative Efficiency, we can expect direct causation forms DIR-Causative to be shorter than indirect causation forms INDIR-Causative (i.e. DIR-Causative < INDIR-Causative).

To illustrate the coding procedure, let us compare two morphological causatives from Urarina, a language isolate spoken in Peru. One of them typically expresses direct causation and is formed with the help of the suffix *-a* (cf. 7a). The other usually expresses indirect causation and contains the suffix *-erate* (cf. 7b). The first causative is shorter and less productive than the second. In addition, it can be attached only to intransitives (Olawsky 2006: 609–621). The causatives, however, do not seem to differ in terms of distance from the non-causal root and autonomy.

(7) (Urarina: isolate, Olawsky 2006: 610–611, 616)

a. *eno-a* 'enter' > *eno-**a**-a* 'make enter'

nalɯ-a 'fall' > *nalɯ-**a**-a* 'drop'

b. *saɯ-a* 'cut' > *sa-**eratia*** 'make cut'

hjani-a 'leave' > *hjane-**ratia*** 'make leave'

4.3. Quantitative analyses: which parameter makes the best match with (in)directness?

4.3.1. Correlations between formal parameters and (in)directness

This section presents the quantitative analyses based on the data described in Section 4.2. Table 4.1 displays the counts for the individual contrasts (74 in total) where the formal parameters behave in accordance with the predictions (DIR-Causative < INDIR-Causative), against them (DIR-Causative > INDIR-Causative), or display no association with (in)directness (DIR-Causative = INDIR-Causative). Table 4.2 shows the number of languages. The counts are lower because some languages have more than one contrasting pair. As an example, compare the

bottom rows in the column 'DIR-Causative < INDIR-Causative'. The number 59 (Table 4.1) means that 59 contrasts were found where DIR-Causative < INDIR-Causative with respect to length. The number 39 (Table 4.2) means that these 59 contrasts occurred in 39 languages. Note that the numbers in each row of Table 2 do not add up to the total number of languages (46) because some languages have contrasting pairs that behave differently.

Table 4.1. Formal parameters associated with (in)directness of causation: number of contrasting pairs

Parameter	**DIR-Causative < INDIR-Causative**	**DIR-Causative = INDIR-Causative**	**DIR-Causative > INDIR-Causative**
Distance	44	30	0
Autonomy	41	33	0
Productivity	40	33	1
Length	59	13	2

Table 4.2. Formal parameters associated with (in)directness of causation: number of languages

Parameter	**DIR-Causative < INDIR-Causative**	**DIR-Causative = INDIR-Causative**	**DIR-Causative > INDIR-Causative**
Distance	27	26	0
Autonomy	24	28	0
Productivity	24	28	1
Length	39	10	2

Both tables reveal that the formal parameter most commonly associated with (in)directness is formal length. DIR-Causative is shorter than INDIR-Causative in 39 languages and in 59 contrasts. DIR-Causative is as long as INDIR-Causative in only 10 languages and 13 contrasts. There are two exceptions, when DIR-Causative is longer than INDIR-Causative, which are discussed below. Length is followed by distance: DIR-Causative is less distant than INDIR-

Causative in 27 languages and 44 contrasts, whereas in 26 languages and 30 contrasts there is no difference. Next follows autonomy, with 24 languages and 41 contrasts, where the predictions are met. The parameter the least strongly associated with (in)directness is productivity (24 languages and 40 contrasts, plus one exception from the predicted direction of association).

One of the exceptions related to length is found in Kayardild, where the causative suffix expressing direct causation is actually longer than the one expressing indirect causation, as shown in (8). However, the indirect causative suffix {-lu-tha} is also used in the factitive function, which means 'cause to be in a state' (Evans 1995: 355). This functional overlap makes it difficult to say which of the constructions in general is more direct and which is less direct, since causing a state is usually associated with less agentive Causees.

(8) (Kayardild: Tangkic, Evans 1995: 355)

a. direct causation: suffix *-THarrma-tha*

thulatha 'descend' > *thulatharrmatha* 'take down'

dalija 'come' > *dalijarrmatha* 'bring'

b. indirect causation: suffix {-lu-tha}

dulbatha 'sink (intr)' > *dulbalutha* 'cause to sink, drown' (e.g. by shooting and not allowing to get out of water)

Another length-related exception is found in Mutsun (Penutian), where the mediopassive-causative suffix *-mpi* (causing a change of state) is actually longer than the active causative *-si* (making someone do something). An example is provided in (9), where (9a) illustrates the causative with *-mpi* and (9b) the causative with *-si*.

(9) (Mutsun: Penutian, Okrand 1977: 216, 219)

a. *mala-n* 'to get wet' > *mala-**mpi**-* 'to cause (someone) to get wet'

b. *ka·n-was* *lolle-**si**-Ø* *sinnise*

I-him babble-CAUS-NPST baby.OBJ

'I made the baby babble.'

This exception can be explained historically: the suffix *-mpi* in fact represents a fusion of the mediopassive suffix *-n* and the suffix *-pi*, which no longer occurs autonomously (Okrand 1977: 215–216).

The third exception, which is related to productivity, is found in Filomeno (Totonacan), where the construction expressing indirect causation *maq(a)-*, which can only be combined with verbs of emotion and physical sensation, e.g. 'make cry by scolding', is less productive than the prefix *maa-* expressing direct causation, which is extremely productive (with the exception of postural verbs) (McFarland 2009: Section 5.4.1). Examples of these constructions were provided in Section 4.2.2. However, one can explain this exception by the fact that the (in)directness contrast is only present in the verbs of emotion and physical sensation. For all other verbs, the default prefix *maa-* seems to express causative situations that can be interpreted as direct and indirect. Consider an example:

(10) (Filomeno: Totonacan, McFarland 2009: 153)

kinkaatiimaamaqtaqáłni

kin-kaa-tii-**maa**-maqtaqal-nii-l̨i

1OBJ-OBJ.PL-PASS-CAUS-care.for-DAT-PERF

'He made us pass by to care for him.'

Moreover, in two languages, Finnish (Uralic) and Lahu (Sino-Tibetan), there were contrasts without any formal differences. For example, Lahu contains an analytic causative with benefactive auxiliary *pî* 'give', e.g. *šɨ* 'die' > *šɨ pî* 'make (him) die', which expresses less direct causation than the default analytic causative with *cɨ* 'make, let; originally send on errand',

e.g. *ši ci* 'make him die' (Matisoff 1976: 430). However, I did not manage to detect any differences between the constructions with regard to the formal parameters in the literature. [32]

One might wonder whether these biases towards the predicted differences between DIR-Causative and INDIR-Causative (or at least, against the negative differences) are statistically significant. To answer this question, I performed the binomial exact test, using the frequencies of individual contrasts presented in Table 4.3. The null hypothesis is that there is no difference with regard to the direction of the asymmetry. In other words, we can have either DIR-Causative > INDIR-Causative or DIR-Causative < INDIR-Causative. There is no preference. The alternative hypothesis is that there is a cross-linguistic bias towards DIR-Causative < INDIR-Causative. The *p*-values of the one-tailed binomial exact test (with the Holm correction for multiple comparisons) are all below 10^{-10}, which is a very small number. Unfortunately, due to the non-independence of contrasts (many represent one and the same language), these results may be unreliable. To solve this problem, I performed a randomization test. For each formal variable, I constructed 100 samples, with only one contrast per language randomly selected, and ran the binomial test again with the help of an R script. After that I took the maximal *p*-values for each formal variable that were generated by the randomization procedure. All adjusted *p*-values were less than 0.0001, which means that the biases in the positive direction are highly statistically significant.

4.3.2. Analysis of formal types

It would also be interesting to see whether the effects observed in the previous section vary across the traditional and well-known types of causatives: lexical, morphological and syntactic. This subsection compares the formal parameters in three different situations:

a) when both DIR-Causative and INDIR-Causative are morphological;

b) when DIR-Causative is morphological and INDIR-Causative is syntactic;

c) when both DIR-Causative and INDIR-Causative are syntactic.

[32] The construction with *pî* 'give' can be used only with intransitives as a valency increasing device. With transitives, it retains its benefactive-directional meaning. It is also available as a causativizer only when the Causee is in the third person (Matisoff 1976: 430).

No lexical causatives were considered because they are not described in grammars systematically.

It goes without saying that the language-internal descriptive categories used for identification of these constructional types may not always be adequate for cross-linguistic comparisons (Haspelmath 2010), but I will use these categories as a proxy. As morphological causatives I considered affixal derivations, root changes, augmentations, reduplications and tonal changes (cf. Dixon 2000). Under the label of syntactic causatives, I conflated monoclausal verbal compounds, serial verbs, light verbs and biclausal periphrastic causatives. A finer-grained classification would be very problematic, due to the relative scarcity of these constructions in the data. Labile verb patterns and clitics were excluded as intermediate types, as well as a few pairs where the status of one of the causatives was dubious. In particular, this was the case with the Basque causative verb or suffix *-(e)raz* (Hualde and Ortiz de Urbina 2003: 593).

Morphological DIR-Causative and INDIR-Causative with the (in)directness distinction are observed in 21 contrasting pairs and in 20 languages. The counts for the pairs are shown in Table 4.3.

Table 4.3. Formal parameters associated with (in)directness of causation: number of contrasting pairs when both DIR-Causative and INDIR-Causative are morphological

Parameter	**DIR-Causative < INDIR-Causative**	**DIR-Causative = INDIR-Causative**	**DIR-Causative > INDIR-Causative**
Distance	3	18	0
Autonomy	0	21	0
Productivity	10	10	1
Length	15	4	2

One can see that length is again the most strongly associated with (in)directness, but this time it is followed by productivity. The fact that all pairs of DIR-Causative and INDIR-Causative have the same level of autonomy is an artefact of this analysis, since autonomy was measured on the basis of the same morphosyntactic cline in (6), which was also used for determining

whether a construction is lexical, morphological or syntactic. The exceptions (DIR-Causative > INDIR-Causative) were already explained in the previous section. The only positive bias that is statistically significant, however, is displayed by length (maximal adjusted $p = 0.01$), according to the binomial test based on resampling (see the description of the procedure in Section 4.3.1). Counting the languages instead of the constructional pairs yields an identical picture.

Let us now examine the pairs with a morphological DIR-Causative and syntactic INDIR-Causative. The total number of contrasting pairs is 33, and the number of languages is 21. The counts for the contrasting pairs are displayed in Table 4.4. This time, autonomy asymmetries are present in all pairs, but this is again an artefact of the definitions of morphological and syntactic causatives, as mentioned above. If we ignore that, we will see that length is again in the leading role, followed by distance. This time, productivity is the least strongly associated parameter. All positive biases are statistically significant, based on the binomial test (all maximal adjusted $p < 0.001$). Counting the languages gives the same picture as counting the contrasting pairs of constructions. This is why the counts of languages are not reported.

Table 4.4. Formal parameters associated with (in)directness of causation: number of contrasting pairs where the DIR-Causative is morphological and INDIR-Causative is syntactic

Parameter	**DIR-Causative < INDIR-Causative**	**DIR-Causative = INDIR-Causative**	**DIR-Causative > INDIR-Causative**
Distance	30	3	0
Autonomy	33	0	0
Productivity	24	9	0
Length	31	2	0

Finally, I have found twelve contrasts and eight languages when both the Dir-Causative and Indir-Causative are syntactic. The counts for the contrasts are presented in Table 4.5.

Table 4.5. Formal parameters associated with (in)directness of causation: number of contrasting pairs when both the DIR-Causative and INDIR-Causative are syntactic

Parameter	**DIR-Causative < INDIR-Causative**	**DIR-Causative = INDIR-Causative**	**DIR-Causative > INDIR-Causative**
Distance	4	8	0
Autonomy	2	10	0
Productivity	2	10	0
Length	5	7	0

Again, length is the most prominent parameter. It is closely followed by distance. Autonomy and productivity have the same counts. The biases are observed again, but they do not reach statistical significance, probably due to the small sample size. The counts of languages instead of contrasts reveal the same picture.

4.4. Summary and discussion

The results of the analyses presented in Section 4.3 show that in general, all factors seem to be associated with (in)directness in the predicted direction at a statistically significant level. The constructions that express indirect causation are either more distant/ autonomous/productive/longer than the constructions that express direct causation (DIR-Causative < INDIR-Causative), or are as distant/autonomous/productive/long as those (DIR-Causative = INDIR-Causative). The exceptions (DIR-Causative > INDIR-Causative) are very scarce. This shows that all previous accounts contain some grain of truth.

However, one can see that relative length is the parameter which is the most strongly associated with the distinction, both in the whole data set and in each constructional type. The results thus favour the explanation based on the Principle of communicative Efficiency. Indirect causation forms are longer than direct causation forms because the indirect causation is less probable than direct causation, which results in efficient formal asymmetries. As was argued in Chapter 3, these asymmetries emerge because of the Low-Cost and High-Cost

Heuristics, which associate a short form with a more probable meaning, and a long form with a less probable meaning. In addition, formal reduction and loss of coding material happen as a result of grammaticalization, which, as was argued in Section 2.2, is also based on the Low-Cost Heuristic.

In addition, we also observed variation in the strength of the other formal parameters depending on the type of causatives compared. If one compares morphological causatives, a direct causation construction will also tend to be more productive than an indirect causation construction, but we observe no or hardly any differences with regard to autonomy and distance. However, if one compares morphological and syntactic causatives, productivity becomes the least strongly associated parameter, although all biases are still statistically significant. A similar tendency is observed when one compares syntactic causatives, but the results do not reach statistical significance, most likely due to the small sample size.

It is time to revisit Shibatani and Pardeshi's (2002) claim that productivity is the formal parameter that is more associated with (in)directness than the traditional formal types, i.e. lexical, morphological and syntactic causatives. Following this claim, one would expect productivity differences to appear more systematically in the data than autonomy, since the latter forms the basis for the classification. However, my data show that productivity is more relevant than autonomy only when we compare two morphological causatives, one being less productive than the other. As for the other causatives examined in this study, productivity fails to play a prominent role. This finding is not surprising because Shibatani and Pardeshi had a bias towards morphological causatives in their study. This parameter seems to gain in importance for those causatives that are at a late stage of grammaticalization, but does not display a reliable correlation with (in)directness for all types of causative constructions.

Distance is also particularly prominent when the direct causative is morphological and the indirect causative is syntactic. It is almost as important as length. This can be explained again by the difference in grammaticalization, since morphemes, unlike auxiliaries, tend to have a fixed position very close to their hosts. As a result, their potential for variability in distance is rather limited.

All this shows that the formal differences depend on the type of constructions and the level of grammaticalization of the causativizing elements. Yet, the pervasiveness of length asymmetries suggests that efficiency-related effects may be more universal than the others. I hope that future research will shed light on the historical stages in the development of causative

constructions and the principles that function at the different stages of grammaticalization, so that the causal relationships between the formal and semantic parameters could be described with greater precision. Another task for the future is the investigation of variation in the marking of the core arguments (the Causee and the object), since in many languages, this marking reflects (in)directness, representing the role of the Causee in the caused event (cf. Comrie 1976; Kemmer and Verhagen 1994).

Chapter 5. Evolution of efficient formal asymmetries: Evidence from artificial language learning of causative constructions

5.1. Aims of this chapter

This chapter concludes the part dedicated to the efficient formal asymmetries in causative constructions.[33] The aim of this chapter is to demonstrate that language users have a bias towards efficiency. More exactly, they tend to spontaneously create efficient form-meaning mappings of semantically similar expressions, whereby the more frequent and probable meaning is expressed by a shorter construction, and the less typical one is expressed by a longer form. This bias is demonstrated with the help an artificial language learning experiment.

The chapter is organized as follows. First, I present the artificial language learning paradigm, focusing on its use for testing of universal biases of language users. Next, I describe the artificial language learning experiment with causative constructions. The final section concludes this chapter and the second part of the present study.

5.2. The artificial language learning paradigm

5.2.1. Aims and types of artificial language learning

Finding and testing functional explanations of linguistic behaviour is not an easy task. Ideally, we would need data from genealogically and geographically diverse languages over a large time span. Needless to say, this is unrealistic: as a rule, such data are not available. The time depth and typological breadth of available diachronic data are very limited. Moreover, even in an ideal world where any kind of linguistic data is obtainable at the click of a button, this might

[33] The findings presented in this chapter are to appear in the following publication: Levshina, Natalia. In press. Linguistic Frankenstein, or How to test universal constraints without real languages. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne M. Michaelis & Ilja Seržants (eds.), *Explanation in linguistic typology: Diachronic sources, functional motivations and the nature of the evidence*. Berlin: Language Science Press.

still be insufficient. First, real language data are observational, which makes a causal interpretation of the correlational results rather difficult (this does not mean there are no attempts, e.g. Moscoso del Prado 2014). Second, real language is a battleground of various forces, many of which can be mutually exclusive, e.g. over- and underspecification, iconicity and economy-driven arbitrariness, and so on. Disentangling these factors in real 'messy' language data is not a trivial task. Moreover, as pointed out by Smith et al. (2017) in their discussion of the universal bias against free variation, transmission of language in populations can mask the biases of learners: the language in a population might retain variability even though every learner is biased against acquiring such variation. Unless the data contain meta-information about the speakers, these effects may go undetected.

These problems can be solved with the help of the artificial language learning paradigm, which has gained popularity recently. One can observe in real time how linguistic systems undergo change, revealing the cognitive and communicative biases of language users. One can control for some factors while testing those of interest, and study the behaviour of each individual speaker within a population.

There are several popular types of artificial language learning experiments (see Figure 5.1). First of all, learning can be iterated and non-iterated. In non-iterated learning, one can only study the individual process of acquisition. There is no further language transmission. In iterated learning, a subject learns a certain linguistic behaviour by observing the behaviour of one or more subjects who learnt it the same way, i.e. in the process of implicit induction and production (Kirby et al. 2014). The output of one generation of speakers serves as the input for the next one, similar to the transmission of real language and culture in general.

Some communicative and learning biases may be strong enough to be detected in non-iterated learning. Sometimes even one generation is enough to radically change the language (Hudson Kam and Newport 2009). Weaker biases may require several generations in order to manifest themselves (e.g. Reali and Griffiths 2009; Smith and Wonnacott 2010).

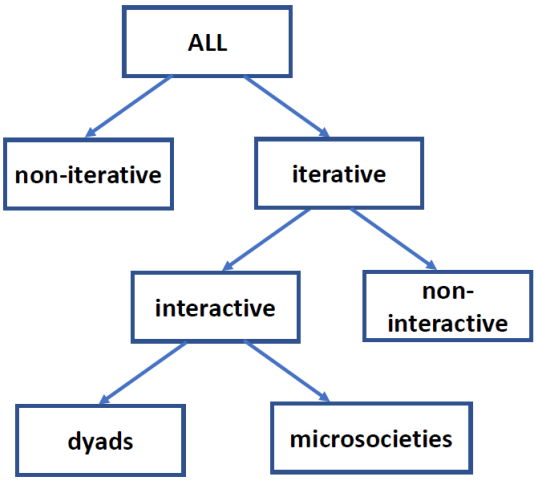

Figure 5.1. Main types of artificial language learning experiments

Iterated learning can be further subdivided into interactive and non-interactive (cf. Tamariz 2016). In non-interactive designs, one creates transmission chains where one subject's output is another subject's input. There is no actual interaction between the subjects. No common ground is created, and no feedback is given. Interactive experiments involve dyads of interacting users or even microsocieties, where everyone interacts with everyone else (Tamariz 2016). Language is transferred from one dyad to the following one, or from old members of a microsociety to the new ones. By using this approach, one can preserve common ground and feedback, which are crucial in everyday communication (Caldwell and Smith 2012).

The artificial language learning paradigm is very flexible, allowing for investigation of diverse forms: non-existent words, whistles, graphical scribbles. A language can also be fully artificial or semi-artificial. For instance, Smith and Wonnacott (2010) use some lexical items (nouns) from English, but novel verbs and plural noun markers. Usually, it is assumed that the results based on various media are comparable, although some recent studies suggest that the role of universal constraints (e.g. iconicity and compositionality) varies across different media (e.g. manual signs vs. sounds in Little et al. 2017).

Crucially, the studies based on artificial language learning share one fundamental assumption. Namely, those linguistic features that are easier to learn and use in communication will spread at the expense of less "fit" alternatives (Smith et al. 2017). By adjusting the linguistic input in a similar way, language users reveal their communicative and

learning biases, which are so strikingly similar that one can speak about universal preferences.

5.2.2. Evidence of universal constraints from artificial language learning

The main results of recent studies can be concisely and non-exhaustively presented in a list of the following universal biases:

(1) a bias towards arbitrariness (as opposed to iconicity), conventionalization and simplification of signs in interaction (e.g. Caldwell and Smith 2012). Simplified arbitrary signs are easier to select, produce and replicate than more complex iconic signs. At the same time, symbolic signs are more difficult to learn at first encounter, while iconicity seems to enhance the learnability of signs for new group members, as shown by Fay and Ellison (2013). They also found that the semiotic systems of larger populations reach a kind of a compromise: they favour simple iconic signs, i.e. those that are minimally complex and maximally informative;

(2) a bias towards combinatorial structure, when meaningless elements (which serve as basic building blocks) are combined in higher-order units. This is also known as duality of patterning (Verhoef 2012);

(3) a bias towards compositional structure of syntax (Kirby et al. 2008). During the process of iterative learning, language becomes more structured;

(4) a bias towards discrete structure as opposed to holistic signals. For example, in an iterated language learning experiment with a language based on whistles, participants come up with categorical distinctions, rather than paying attention to the precise acoustic realizations, e.g. in terms of pitch (Verhoef 2012);

(5) a bias towards regularity. Languages exhibiting free variation become increasingly regular, revealing a strong bias towards regularity in adult learners (Smith and Wonnacott 2010). This bias may be obscured by so-called probability matching: in a language in which two forms are in free variation, adult learners have also been found to produce each variant in accordance with its relative frequency in the input (Hudson Kam and Newport 2009). The interplay between regularization and probability matching depends on the frequency distribution. The more forms with lower frequencies are used as free variants of the main form, the more scattered the pattern and the stronger the bias towards production of the main form (Hudson Kam and Newport 2009);

(6) a bias towards economy and communicative efficiency, when more predictable information gets less formal coding, and less predictable information gets more formal coding. This bias has been observed in a study of differential case marking (Fedzechkina et al. 2012). The hypothesis is that a referential expression should be more likely to receive overt case marking when its intended grammatical function is less expected. The experiment shows that learners deviate from the initial input to make the language more communicatively efficient;
(7) a bias towards underspecification of irrelevant conceptual dimensions. Silvey et al. (2015) have found that their artificial language, which was originally fully specified in the sense that it had a unique label for each object, became underspecified by losing contrasts across irrelevant dimensions, i.e. those that are not important for discriminating between the stimuli. In contrast, Tinits et al. (2017) found a bias towards overspecification and redundancy in the contexts when the relevant dimensions were difficult to discern.

To the best of my knowledge, a bias towards efficiency has not been tested yet. In the remaining part of the chapter, I will focus on the bias towards communicatively efficient form-meaning pairings in causatives, using a non-iterative online experiment.

5.3. Frequency effects in causative constructions

5.3.1. Hypothesis: Efficiency and formal length

As was shown above, there is ample evidence that language learners are generally sensitive to frequency information. In this case study, I focus on the claim that more frequent situations are expressed by means of less coding material than less frequent ones. Such differences are predicted by the Principle of Communicative Efficiency. According to these principles, more probable information needs less coding material than less probable information. The experiment in Fedzechkina et al. (2012), which was mentioned above, demonstrated the effect of predictability based on semantic categories. In my own study, I want to focus on predictability based on frequency information. To the best of my knowledge, these effects have not been tested previously in artificial language learning experiments.

Causatives serve as convenient and well-studied material for testing the bias in question. A lot of information about their variation was provided in Chapters 3 and 4. There is a cross-linguistic correlation between form and meaning: more formally integrated causatives, such as lexical causatives *kill* or *break*$_{TR}$, tend to denote more integrated causing and caused events than less integrated forms, such as *cause to die* or *make break*$_{INTR}$. This correlation has been often explained on the basis of the principle of iconicity. In the previous chapters, I developed an alternative account, based on the Principle of Communicative Efficiency: more expected causative situations are usually expressed by shorter causatives, whereas less typical ones are expressed by longer constructions. Thus, these parameters (conceptual integration, formal compactness and relative frequency) are intercorrelated: more compact causatives represent both more frequent situations and more integrated events, whereas less compact causatives represent less frequent situations and also less integrated events. This creates a situation in which it is difficult to decide based on observational data alone which of the functional principles actually explains the cross-linguistic correlation between formal and conceptual integration, i.e. iconicity or efficiency. Although an attempt to disentangle different parameters was made in the previous chapter, the interpretation would be more convincing if we could take the iconic correspondence out of the equation and demonstrate that efficient asymmetries can arise due to efficiency alone. This is the purpose of the case study presented below.

5.3.2. Design and procedure

The participants of the experiment were asked to learn an alien language. At the beginning, they read an introduction:

"In this experiment you will learn the lingua franca of a highly developed civilization that exists on a planet in a galaxy far, far away... The planet is called Atruur. Its only vegetation form is called 'grok'. It is similar to a cactus and is used by the Atruurians for food, as fuel for their flying vehicles and for entertainment. Because the Atruurians traditionally detest any form of physical activity, they have developed a technology for teleportation and telekinesis."

The introduction also mentioned that the word order is SV (for intransitives) or SOV (for transitives). To explain that to non-linguists, examples were provided, which are shown below for illustration:

(1) *Grok* *babum.*

cactus grow

'A grok (cactus) grows.'

(2) *Sia* *grok* *hum.*

Atruurian cactus see

'An Atruurian sees a grok (cactus).'

The subjects were first asked to learn the language by copying the sentences in Atruurian that describe situations shown in video clips. At first, they saw four situations: a cactus-like plant appears, disappears, grows and shrinks in size. The goal of that task was to introduce the basic vocabulary.

Next, the participants saw 32 causal situations, which represented a causal version of the same situations. In each of these causal situations, there was a flying saucer (sometimes with an alien inside) which hovered above the plant and flashed a yellow or blue light three times in a row. As a result, the plant either appeared, disappeared, grew or shrunk. Varying types of saucers were shown.

Crucially, the subjects saw two types of causing events. The first of them involved the saucer flashing a yellow light above the plant. The other one was when the saucer flashed a blue light from the left of the plant. The yellow-light causing event was three times as frequent as the blue-light causing event (i.e. 75% vs. 25%). The distribution of the four caused events was the same for each of the causing events. There were no reasons to assume that one type of causation is more or less direct than the other. The colour and the position of the Causer with regard to the Causee are not mentioned in the semantic parameters that are distinctive of different causative constructions in the languages of the world (see Chapter 3).

As for the artificial language, the most important thing is that each causing event is represented by two allomorphs. One of the causing events was associated with the forms *tere-* or *te-*, as in (3), and the other one was described by using the forms *gara-*/*ga-*.

(3) a. *Sia* *grok* *te-babum.*

Atruurian plant CAUS-grow

'The Atruurian caused the plant to grow (by flashing with yellow light from above).'

b. *Sia* *grok* *tere-babum.*

Atruurian plant CAUS-grow

'The Atruurian caused the plant to grow (by flashing with yellow light from above).'

Note that the form-meaning mapping varied across the subjects. That is, for some of them, *te-/tere-* denoted the causing event with yellow light flashed from above, whereas the *ga-/gara-* forms were used for the causing event with blue light flashed from the left of the plant. For the others, this was the other way round. The prefixes were evenly distributed among the stimuli, so that there was truly free variation. There was no condition in the experimental design that could explain the preference for the longer or the shorter form.

One should mention here that free variation is less exotic than it seems. It occurs in the language of late learners of a second language, e.g. hearing parents of a deaf child who learn to sign, or during the emergence of a new language, e.g. Tok Pisin and other pidgins and creoles (see an overview in Hudson Kam and Newport 2009). This is why the input language is not completely outlandish from a functional point of view. However, since language users have a bias towards regularization and against free variation, I expected that the subjects would regularize the free variation in the input, preferring the short allomorphs to convey the frequent causing events, and using the long allomorph to express the rare causing events.

After the training session, the subjects were asked to describe in Atruurian what is going on in videoclips. The stimuli represented a selection from the previous stimuli: each of the caused events was presented with causing event A and causing event B. In total, there were eight test situations.

The experiment was performed online, using Google Forms with built-in YouTube videos. The latter were created with the help of Adobe Animate CC software by myself. Figure 5.2 demonstrates four fragments from one of the video clips, with the causing event A and the caused event of disappearance.

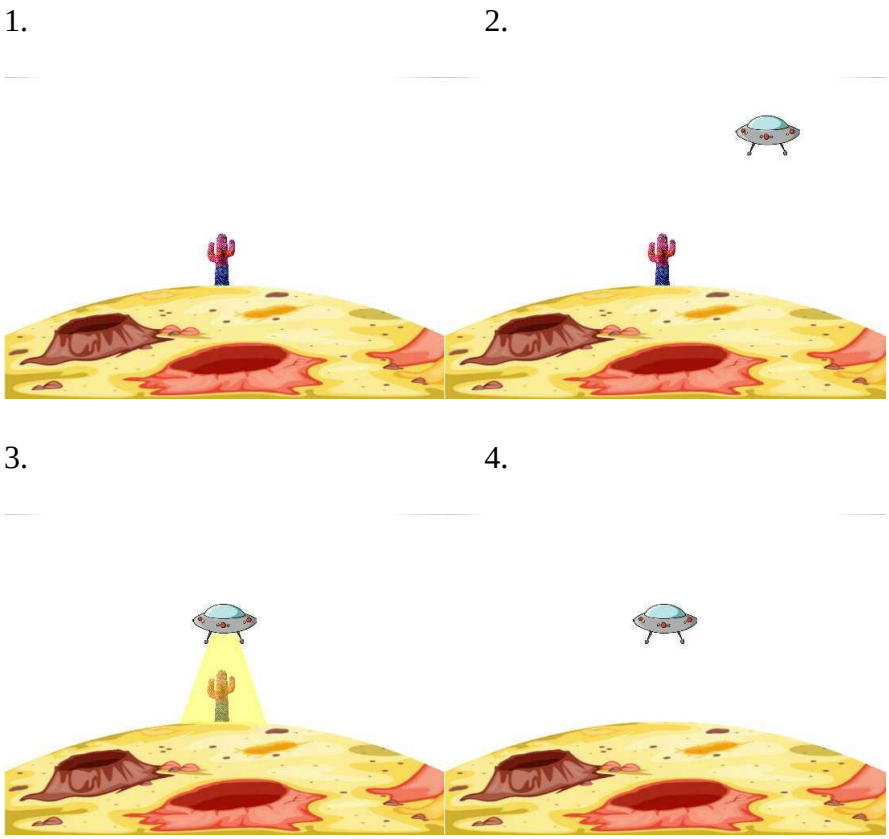

Figure 5.2. Fragments from one of the video clips.

5.3.3. Participants

The participants were recruited via my personal network and LinguistList. Most of them had a background in linguistics or languages. After the experiment, they were asked about the aims

of the experiment. None of them guessed the true purpose. Overall, I obtained responses from 84 participants. Some of the responses were removed. This was the case if the participants did not follow the training procedure instructions (e.g. a participant did not type in the training sentences), or if the output was unanalysable. As a result, I had 554 valid data points from 70 participants.

The participants with valid responses had different L1s, but mostly had a Slavic and Germanic linguistic background. There were 40 native Czech speakers, 12 native German speakers, 7 native English speakers, 2 Dutch speakers, 2 Italian speakers, as well as native speakers of Brazilian Portuguese, Croatian, Danish, Polish, Russian, Slovak and Turkish. None of these languages has productive causative prefixes.

5.3.4. Results of the experiment

The counts aggregated across all participants are presented in Table 5.1. Lexical and spelling errors were ignored. Figure 5.3, which visualizes these counts, shows that there is a difference between the proportions of short and long forms expressing the frequent and rare causing events. The short forms are overall more preferred than the long ones, but the situations with the more frequent causing event are more frequently expressed by the short forms in comparison with the situations that involve the rare causing event, where the proportions of the short and long forms are almost equal.

Table 5.1. The number of forms selected and their marginal sums

Form	**Frequent**	**Rare**	***Total***
Short	168	137	*305*
Long	109	140	*249*
Total	*277*	*277*	*554*

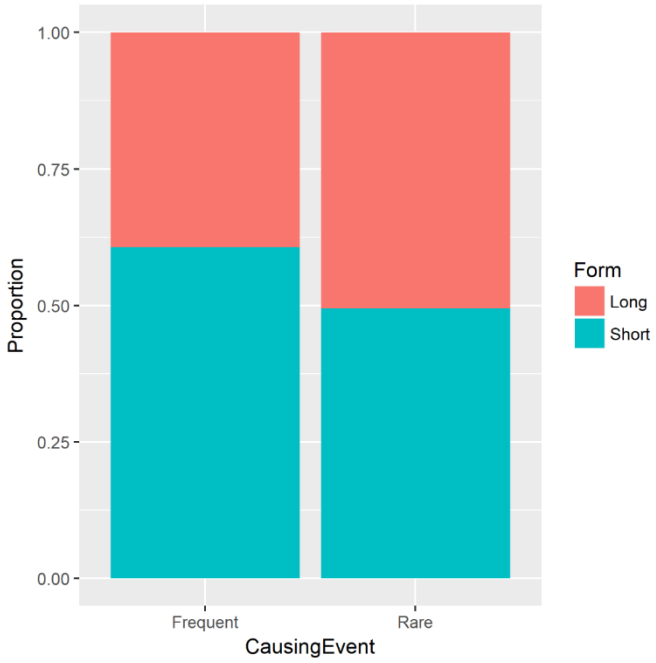

Figure 5.3. Proportions of short and long forms in the subject's responses

A closer look at the individual subjects' preferences reveals that most of them use both long and short forms. Seven subjects produced only the short forms. There were no subjects who always preferred the long forms. The distribution is shown in Figure 5.4.

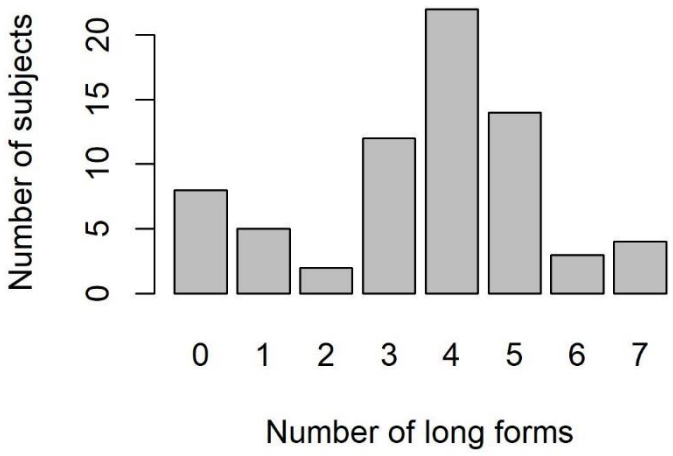

Figure 5.4. Individual preferences for the long and short forms

The main question, however, is whether the choice of forms is influenced by the type of causing event. In order to test this, I fit a generalized linear mixed-effects model with logit as the link function (R package *lme4*, function *glmer*, Bates et al. 2015). The type of prefix – long or short – was the response variable. The individual participants were treated as random effects (intercepts). There is a significant effect of the type of causing event: if the event is rare, the odds of the longer form to be chosen are 1.66 times greater than when the event is frequent (log-odds ratio $b = 0.501$, $p = 0.006$). This result supports the hypothesis that speakers have a bias towards the use of shorter forms to represent more frequent situations, and longer forms to represent less frequent situations. Random slopes, which represented individual differences in the effect of the predictor on the response, were tested as well, but they did not improve the explanatory power of the model.

The likelihood ratio test, a standard tool for variable selection and model comparison in regression analysis, demonstrates that the caused event does not have a significant effect on the choice of form ($p = 0.84$), and does not interact with the type of causing event ($p = 0.6$). This means that lexical conditioning can be excluded (cf. Smith and Wonnacott 2010).

5.4. Summary and discussion

This case study has tested the Principle of Communicative Efficiency in an online experiment. The results demonstrate that frequent causative situations become more commonly expressed by shorter forms, whereas the subjects are more tolerant of longer forms when expressing rarer causative situations. As a result of this experiment, a more efficient system emerges. The fact that the effect was detected in a non-iterative experiment with only one "generation" of language learners, suggests that the bias is very strong. This provides evidence in favour of the efficiency-based account of functionally similar expressions. It is possible that both the Low-Cost and High-Cost Heuristics are involved: the speakers tend to use the shorter form to express the more probable meanings, but one cannot exclude that the longer form is used in some cases to mark the less typical variant. Since both forms were originally available to the learners, we witness competition between the existing forms, and how they become specialized in different

causative situations. It would be interesting to see if language users can also innovate in experimental settings, creating new reduced or enhanced forms.

Thus, the evidence provided by typological data, corpora and the experimental approach converge. The form-meaning correspondences in causative constructions, which have been explained in terms of iconicity or productivity, are best explained by the Principle of Communicative Efficiency. In addition, this account explains other form-meaning correspondences, as well as the emergence of new asymmetries, which do not involve iconicity.

Part III. Coding asymmetries and splits

Chapter 6. Differential case marking of A and P: Reverse engineering and recycling of corpus data

6.1. Aims of this chapter

This chapter discusses differential marking of A and P arguments. Differential marking is observed when the same argument has different coding depending on its semantic or pragmatic properties. Case markers include both case affixes and adpositions. The main aim of this chapter is to compare different explanations of differential case marking in the literature and see which of them fits the cross-linguistic tendencies the best. Three possible strategies of marking choices are compared and operationalized in the form of different types of conditional probabilities of the A and P features that one can find in discourse. These strategies are based on the previous functional-adaptive accounts that involve the notions of disambiguation and iconicity of markedness. The cross-linguistic data are collected from the typological database AUTOTYP (version 0.1.0) created by Bickel et al. (2017). Using these data, I perform 'reverse engineering', trying to predict from the observed cross-linguistic patterns how the features of A and P should be distributed in discourse for the observed patterns to arise, according to each of the strategies.[34] In order to compare the predictions with the discourse data, I perform a descriptive analysis of several previous corpus-based studies of the Preferred Argument Structure in different languages and registers of communication.

[34] This approach is similar to the process of abductive reasoning. One begins with some RESULT (or conclusion), applies a general RULE (major premise), and comes to a so-called CASE (minor premise). For example, one observes that there are no lights in the neighbours' home at 9 pm (RESULT), applies some background knowledge (e.g. they always go to bed early – RULE), and comes to the explanation, or CASE (they went to bed early) (Andersen 2016). The term 'abduction' was introduced by Charles Pierce. I'm grateful to Doris Schönefeld for making me aware of this parallel. At the same time, in the large context of this study, this case study is deductive because we apply a general rule (i.e. that predictability determines the amount of coding) to a specific case (i.e. differential case marking) in order to obtain the result (i.e. that predictability of role-feature relationships determines the amount of coding in differential case marking).

The results of this reverse engineering and data recycling show that the conditional probabilities of roles (A and P) given relevant features (animacy, pronominality, etc.) in discourse provide the best fit for the cross-linguistic distribution of the coding splits. More exactly, the cross-linguistically marked A and P are the arguments with the features that exhibit low cue validity with regard to their syntactic role, or, in other words, low conditional probabilities of A or P given the feature, while the unmarked arguments are those with high cue validity. This supports the efficiency account based on the Low-Cost and High-Cost Heuristics.

This chapter provides a novel contribution to the well-explored topic of differential argument marking because the intuitions about the causes that lead to the emergence of cross-linguistic patterns are captured in the form of measurable probabilities, which can be found in corpus data.

The next section gives an overview of the main ideas about differential case marking. In Section 6.3, I provide data from AUTOTYP, which reveal the most common coding splits in the differential marking of A and P. Section 6.4 presents explanations of differential case marking, which have been formulated in the literature. It also provides testable predictions for corpus data. In Section 6.5, these predictions are checked in a descriptive analysis of frequencies from 19 different corpora of different registers and in diverse languages. Section 6.6 draws the conclusions.

6.2. Differential case marking: an overview

This chapter discusses differential case marking of A and P arguments. These roles are defined as the syntactic functions that include the prototypical Agent and Patient, respectively, as in the sentence *John broke the window*. Differential case marking of A is referred here to as differential agent marking (DAM) with a broader interpretation – i.e. marking of subjects of transitive clauses. An example of DAM from Quiang, where the inanimate A is marked, and the animate A is unmarked, is given below:

(1) Qiang (Sino-Tibetan, LaPolla and Huang 2003: 79–80)

a. Animate A: unmarked

The: *qa* *dʑete.*
3SG 1SG hit
'He is hitting me.'

b. Inanimate A: marked

*Moʁu-**wu*** *qa* *da-tuə-ʐ.*
wind-**AGT** 1SG DIR-fall.over-CAUS
'The wind knocked me over.'

Differential marking of P is known as differential object marking (DOM). An example in (2) comes from Spanish, where animate objects tend to be formally marked, while inanimate objects are unmarked, although definiteness and individual lexical verbs also play a role (von Heusinger and Kaiser 2007):

(2) (Spanish, García García 2018: 211)

a. Inanimate P: unmarked

Pepe *ve* *la* *película.*
Pepe see.3SG DEF film
'Pepe sees the film.'

b. Animate P: marked

Pepe *ve* ***a*** *la* *actriz.*
Pepe see.3SG **OBJ** DEF actress
'Pepe sees the actress.'

Differential case marking of A and P has been widely discussed in typological literature (e.g. Silverstein 1976; Comrie 1978; Dixon 1979; Bossong 1985; Dixon 1994; Aissen 1999, 2003; de Hoop and de Swart 2008; see also a comprehensive overview in Witzlack-Makarevich

and Seržant 2018). In particular, there is a widespread opinion that the use or absence of markers is constrained by universal referential scales, such as the ones presented by Croft (2003: 130–132) and Haspelmath (Forthcoming-b):

(3) a. Person: 1 and 2 > 3

b. Nominality: pronoun > noun[35]

c. Animacy: human (> animal) > inanimate

d. Definiteness/specificity: definite > specific > non-specific

e. Givenness: discourse-given > discourse-new

f. Focus: background (topic) > focus

Such hierarchies are to some extent logically independent (not perfectly; for example, 1st and 2nd person referents are also pronominal, human, definite and given). This is important because each of the scales may have more or less influence on the use of differential marking in an individual language (Haude and Witzlack-Makarevich 2016). There is a remarkable typological tendency, which has fascinated linguists for many years. If a language has a coding split in A or P, more prominent A arguments (i.e. the ones on the left) are usually unmarked, while less prominent ones (i.e. the ones on the right) are marked. For P arguments, the reverse holds. The examples (1) and (2) provided above illustrate this symmetry: A is marked when it is inanimate, as in Qiang (1), while P is marked when it is animate, as in Spanish (2).

At the same time, it has been observed that DOM and DAM exhibit different 'preferences' for the individual scales. In general, DOM is more frequent typologically (also see below), and the most relevant features in DOM are animacy and definiteness (Sinnemäki 2014). For example, the definiteness hierarchy plays a role in DOM in Biblical Hebrew, where the marker *ʔet* is only used with definite nouns:

[35] This scale sometimes includes kinship terms and proper names. As for the former, their position is different depending on the usage of these terms in the language. In some languages, kinship terms are used very broadly, having a pronominal-like status, while in others they are closer to common nouns. As for proper names, Helmbrecht et al. (2018) have recently demonstrated that proper nouns do not play any important role in typological generalizations based on the animacy hierarchy. The typological data in Section 6.3 support this claim.

(4) a. (Gen. 1:26)

naase adam be-tzalme-nu

create man in-image-our

'Let's create a man in our image…'

b. (Gen. 1:27)

va-yivra	*Elohim*	***ʔet***	*ha-adam*	*b'-tzalm-o*
and-created	Almighty	**OBJ**	def-man	in-image-his

'So God created the man in his image…'

Animacy of P is important in Sinhala, an Indo-Aryan language. Inanimate objects are never marked, whereas animate objects may receive case-marking optionally:

(5) *mamə miniha(-**wə**) daekka*

I man(-**OBJ**) saw

'I saw the man.'

In Hindi, animacy and definiteness of P exhibit a complex interplay. Case marking is obligatory with human objects, definite and specific inanimates. It is optional with non-specific human-referring objects. Finally, it is not used with indefinite inanimate objects (Aissen 2003). Similarly, many scales are involved in the DOM in Romanian: the marker *pe*, mostly combined with clitic doubling, only applies to human objects if they are definite, specific or topicalized (von Heusinger and Onea Gáspár 2008).

Similar to differential object indexing (Haig 2018), differential case marking is an attractor state for objects (Schmidtke-Bode and Levshina 2018: Appendix 2). This notion from dynamic systems theory can be thought of as a region within the space of variation of human language where the trajectories of individual languages tend to settle over time. If a language reaches the attractor state, it is unlikely to leave that state.

DAM is less common typologically. Moreover, DAM systems that involve the animacy hierarchy, as in Qiang example (1), are rare (Fauconnier 2011). There seems also to be little consistent evidence (due to many exceptions) that indefinite or non-specific A's are marked.

Instead, languages use alternative strategies, e.g. passives or presentative constructions similar to the English construction *there is…* (Comrie 1981:123; Malchukov 2008).

Although Aissen (2003) says that the scale effects in DOM represent one of the most robust typological generalizations (cf. also Bossong 1985), there have been critical voices, claiming that the scale effects are not supported enough by cross-linguistic data. In particular, personal pronouns are the greatest offenders (Filimonova 2005). For example, some Indo-Aryan and Iranian languages undergo a change from ergative to tripartite and then to accusative case marking. Due to their resistance to change, personal pronouns retain the old ergative forms in the subject position, while the nouns, which have undergone a sweeping change, have lost their marking. This explains why the pronominal agents are marked, and the nominal ones are unmarked, contrary to what one would expect. The resistance of pronouns can be explained by the conserving effect of frequency (Bybee and Thompson 1997), since pronouns are normally very frequent in discourse.

In addition, Bickel et al. (2015) claim that there are no universal scale effects in differential case marking, which would hold in all parts of the world. As argued in a recent paper (Schmidtke-Bode and Levshina 2018), however, the scale effects can still be considered valid if one uses an alternative statistical method (mixed-effects regression analysis instead of the Family Bias method) and considers alternative possibilities of answering the fundamental question, what it means, for a scale to be universal. In the present study, I will argue that the directionality of the effects is universal (e.g. animate, and not inanimate P's tend to be marked), but which of these individual scales are manifest in different languages varies a lot. In other words, the directionality is universal, but there is no single universal scale that all languages conform to. In fact, this situation is similar to what we saw in Part II where we discussed the semantics of the less compact causatives: the directionality of the formal asymmetries is universal (e.g. indirect causation constructions are less compact than direct causation constructions), but the causation types involved in the semantic distinctions in a given language, are individual. This combination of universal and individual patterns can be explained by the Low- and High-Cost Heuristics. There is a variety of highly correlated untypical features of A and P that can become marked in a given language as a result of conventionalization, but all of them are expressed by costlier forms than the typical features.

6.3. Cross-linguistic distribution of differential case marking

In this section I present frequencies from the AUTOTYP database (Bickel et al. 2017), version 0.1.0, which show which of the scales are more and less relevant for A and P cross-linguistically. Languages have many possible ways of combining the scales. Imagine Language X, which marks animate nominal and pronominal P arguments, and Language Y, which marks only animate nominal ones. In this case, Language X has a split on the animacy scale, whereas Language Y has a split on both the nominality scale (Noun vs. Pronoun), and on the animacy scale within nouns (Inanimate vs. Animate).

One feature of AUTOTYP is that it uses the labels 'high' and 'low' to indicate high or low prominence in discourse (DP). High DP means that an argument is either given, definite, animate, topical, etc., or a combination of such features. Low DP indicates the absence of these features (or some of them). This is why it is impossible to separate definiteness from animacy precisely. Instead, we have a very syncretic discourse prominence (DP) hierarchy:

(6) Discourse prominence hierarchy:

High DP (human/animate, definite/specific, given, topical) > Low DP (non-human/inanimate, indefinite/non-specific, non-topical, new)

The frequencies of the splits found in AUTOTYP are presented in Tables 6.1 and 6.2. I excluded the ones that also involve splits between different language-specific lexical classes of nouns, the first and second person, singular and plural forms, or inclusive and exclusive pronouns. The splits in number and person are particularly frequent in the pronouns (see more information in Schmidtke-Bode and Levshina 2018), which is probably explained by the conserving frequency effects of the type described by Filimonova (2005). I also excluded the cases when one could interpret the data as both a fit and a violation. For example, Menya (an Angan language spoken in Papua New Guinea) has a DOM system, where pronouns and nouns with low discourse prominence are unmarked, while nouns with high discourse prominence are marked. This means that the system fits the DP prominence scale within nouns (high DP nouns are marked, while low DP nouns are unmarked), but violates the nominal scale, according to which pronouns should be marked.

Let us consider the results for A presented in Table 6.1. The double vertical bar '||' represents a split. The features on the left of the sign are unmarked, and the features that are on the right are marked. The marking is understood by the authors in the Silversteinian sense (see Bickel et al. 2015): an argument is unmarked if it has the same expression as S (intransitive subject). In practice it usually means zero expression. The parentheses contain the language family and the geographic macroarea.

The predominant category for A marking is clearly nominality, followed by person. Animacy and discourse prominence only rarely play a role on their own. This supports the previous observations. There is a violation of the nominality scale (see the shaded row): in Gitksan (a Tsimshianic language spoken in Canada) A's expressed by common nouns are unmarked, but A's expressed by proper nouns and pronouns are marked.

The splits in P are shown in Table 6.2. The results partly support the previous observations. Inanimate, indefinite and low-DP nouns are overwhelmingly unmarked. However, the most frequent is the distinction between nouns and pronouns. Person seems to produce the lowest number of splits. Again, there are a few exceptions (see shaded cells), but they are not numerous.

To summarize, the A arguments tend to be marked in DAM if they are nominal and, less frequently, if they are 3rd person. The evidence for the other scales is rather weak. The P arguments tend to be marked in DOM if they are pronominal, high-DP (including animacy and definiteness) and 1st or 2nd person. The evidence for the person scale is the weakest.

Table 6.1. Cross-linguistic distribution of DAM in AUTOTYP 0.1

Distinction	Scale(s)	No. languages	No. families	No. areas	Examples
Pronouns || Nouns	Nominality	20	7	3	Djapu (Pama-Nyungan, Pacific), Kryz (Nakh-Dagestanian, Eurasia), Cashinahua (Pano-Tacanan, Americas)
1st & 2nd Person || 3rd Person (incl. Nouns)	Person	11	4	2	Kham (Sino-Tibetan, Eurasia), Yidiny (Pama-Nyungan, Pacific)
Animates || Inanimates (both pronouns and nouns)	Animacy	2	2	2	Mayali (Gunwingguan, Pacific), Northern Qiang (Sino-Tibetan, Eurasia)
Pronouns, Animate Nouns || Inanimate Nouns	Nominality, Animacy	1	1	1	Hittite (Indo-European, Eurasia)
Pronouns, Personal Proper Names || Other Nouns	Nominality, Animacy	1	1	1	Djinang (Pama-Nyungan, Pacific)
Pronouns, DP Animate Nouns || Other Nouns	Nominality, Animacy, DP	1	1	1	Mangarayi (Mangarayan, Pacific)
Pronouns, Proper Nouns, Kinship terms || Other Nouns	Nominality	1	1	1	Central Pomo (Pomoan, Americas)
The most DP Pronouns and Nouns || Other	DP	1	1	1	Tukang Besi (Austronesian, Pacific)
Pronouns, Proper nouns, Kinship terms || Non-kin common nouns	Nominality	1	1	1	Eastern Pomo (Pomoan, Americas)
Common Nouns || Pronouns, Proper Nouns	Nominality	1	1	1	Gitksan (Tsimshianic, Americas)

Distinction	Scale(s)	No. languages	No. families	No. areas	Examples
Nouns || Pronouns	Nominality	65	33	4	Logba (Kwa, Africa), Khanty (Uralic, Eurasia), Garrwa (Garwwan, Pacific), Rama (Chibchan, Americas)
Low-DP nouns || Pronouns, high-DP nouns	Nominality, DP	59	21	4	Dizi (Omotic, Africa), Awa Pit (Barbacoan, Americas), Tamil (Dravidian, Eurasia), Akoye (Angan, Pacific)
Indefinite nouns || Pronouns, animate nouns	Definiteness, Nominality	14	8	3	Amharic (Semitic, Africa), Chuvash (Turkic, Eurasia), Barasano (Tucanoan, Americas)
Nouns, 3rd person pronouns || 1st & 2nd person pronouns	Person, Nominality	7	6	4	Dyirbal (Pama-Nyungan, Pacific), Kutenai (isolate, Americas), Waskia (Madang, Pacific), Tsova-Tush (Nakh-Daghestanian, Eurasia)
Inanimate nouns || Pronouns, animate nouns	Animacy, Nominality	7	4	2	Hittite (Indo-European, Eurasia), Anamuxra (Madang, Pacific)
Inanimate and low-DP animate nouns || Pronouns and high-DP animate nouns	Nominality, DP	6	3	2	Hup (Nadahup, Americas), Djapu (Pama-Nyungan, Pacific)
Low-DP pronouns and nouns || High-DP pronouns and nouns	DP	5	5	3	Tariana (Arawakan, Americas), Kharia (Austroausiatic, Eurasia), Tainae (Angan, Pacific)
Inanimate nouns and pronouns || Animate nouns and pronouns	Animacy	2	2	1	Imonda (Border, Pacific)
Common, non-kin nouns || All other	Nominality	2	2	2	Eastern Pomo (Pomoan, Americas), Sardinian (Indo-European, Eurasia)
Non-specific nouns || All other	Definiteness, Nominality	2	2	1	Persian (Indo-European, Eurasia)
Noun and low-DP 3rd person pronoun || All other pronouns	Nominality, DP	1	1	1	Yidiny (Pama-Nyungan, Pacific)
Non-kin noun, low-DP 3rd person pronoun || All other	Nominality, DP	1	1	1	Central Pomo (Pomoan, Americas)
Low-DP noun, inanimate pronoun || All other	DP, animacy	1	1	1	Afrikaans (Indo-European, Africa)
Pronoun, high-DP noun || low-DP noun	Nominality, DP	1	1	1	Maithili only in dependent clauses with converbs (Indo-European, Eurasia)
1st & 2nd person pronoun || 3rd person pronoun, noun	Person	1	1	1	Osage (Siouan, Americas)
Highest-DP pronoun and noun || All other	DP	1	1	1	Tukang Besi (Austronesian, Pacific)

Table 6.2. Cross-linguistic distribution of DOM in AUTOTYP 0.1

6.4. Previous explanations of differential case marking and predictions for reverse engineering

6.4.1. Disambiguation and economy

The most popular explanation of the scale effects in differential case marking is that of disambiguation, or discrimination. It is often argued that two nominal phrases, when they are simultaneously present in the transitive clause, should be distinguished (Comrie 1978: 379—380; 1989: 124—127, Dixon 1979, Givón 1984: 184, Aissen 2003, among others). This is supported by the existence of systems in which the marking of A depends on the properties of P, and the other way round. For example, Malayalam (a Dravidian language) marks animate objects and usually does not mark inanimate objects. However, when the sentence is potentially ambiguous, the object marker can also be used on inanimate objects, as in (7).

(7) Malayalam (Dravidian, Asher and Kumari 1997:204, cited from de Swart 2007: 88)

a. *Kappal* *tiramaalakaí-e* *bheediccu.*
ship.NOM waves-ACC split.PST
'The ship broke through the waves.'

b. *Tiramaalakaí* *kappal-ine* *bheediccu.*
waves. NOM ship-ACC split.PST
'The waves split the ship.'

More examples can be found in de Swart (2007: Section 3.2). Yet, such languages are infrequent synchronically. In the majority of languages with differential marking, the marking of A or P does not depend on the properties of the other argument. One can still say that the function of differential marking in such languages is to some extent disambiguating because its existence leads to a decrease of cases with potential syntactic ambiguity (Witzlack-Makarevich and Seržant 2018).

There are at least two strategies that the language users may employ. First, they may want to mark the A's that look like P's, and the P's that look like A's. As Aissen puts it (2003: 437), "it is those direct objects which most resemble typical subjects that get overtly case-marked". In other words, those A's that have properties of typical P's get extra marking, and

the other way round. I will call this strategy **"Mark Confusable Arguments, Don't Mark Non-Confusable Arguments"**.

Formally, this can be expressed with the help of conditional probabilities. Let ℙ (Feature|A) stand for the conditional probability of a certain Feature (i.e. animacy, definiteness, etc.) given the A role. The symbol ℙ stands for probability and should not be confused with P, which denotes the P argument. This probability can be approximated by the proportion of A's with this feature in the total number of A's in a reasonably large sample of discourse. The predictions are then as follows:

(8) Reverse engineering predictions based on the disambiguation principle and the strategy "Mark Confusable Arguments, Don't Mark Non-Confusable Arguments"

Prediction for A:
A with Feature_i tends to be formally marked in languages of the world when the probability $\mathbb{P}(\text{Feature}_i|\text{P})$ in discourse is high, and not marked when $\mathbb{P}(\text{Feature}_i|\text{P})$ is low.[36]

Prediction for P:
P with Feature_j tends to be formally marked in languages of the world when the probability $\mathbb{P}(\text{Feature}_j|\text{A})$ in discourse is high, and not marked when $\mathbb{P}(\text{Feature}_i|\text{A})$ is low.

If this approach is correct, one would expect that a typical P is nominal and 3rd person, since A's with these features are marked cross-linguistically, as was shown in Section 6.3. An untypical P is then pronominal and 1st or 2nd person, because these features are not marked on A. As for A's, their most common properties should be pronominality, high discourse prominence (givenness and definiteness, including humanness/animacy), and the 1st or 2nd person, because P's with these features are usually marked in the typological data. From this follows that A's should be rarely nominal, low-DP, including inanimate, and 3rd person.

Disambiguation, however, can also be understood in a different way, as the predictability of a role (A or P) given some features. As Comrie writes (1978: 385–386),

[36] Here and below, 'high' stands for greater than 0.5, or 50%, and 'low' stands for less than 0.5.

> There seems to be a general supposition in human discourse that certain entities are inherently more agentive than others, and as such inherently more likely to appear as A of a transitive verb and less likely to appear as P of a transitive verb.

Already Royen (1929: 590) observed, "Eine Person ist vor allem agens, ein Impersonale vor allem patiens". If this is true, then additional marking on a human P or on non-human A would help the hearer to identify the role easier. In fact, this idea was already proposed by Bishop Robert Caldwell (1856: 271, cited from Filimonova 2005: 78):

> [. . .] the principle that it is more natural for rational beings to act than to be acted upon; and hence when they do happen to be acted upon – when the nouns by which they are denoted are to be taken objectively – it becomes necessary, in order to avoid misapprehension, to suffix to them the objective case-sign.

If this explanation is valid, a core argument with a particular feature (animate, definite, 3rd person, etc.) is marked if referents with this feature are infrequently used in this role in discourse. Conversely, they are not marked if they are frequently used in a certain role. These associations can be measured as conditional probabilities $\mathbb{P}$ (Role|Feature$_i$), which represent cue validity (see Chapter 1). The referential features are here cues, and the syntactic role (A or P) is the intended interpretation. This disambiguation strategy can be labelled as **"Mark Weak Cues, Don't Mark Strong Cues"**.

(9) Reverse engineering predictions based on the disambiguation principle and the strategy "Mark the Weak Cues, Don't Mark Strong Cues"

Prediction for A:
A with Feature$_i$ tends to be formally marked when $\mathbb{P}$ (A| Feature$_i$) is low, and unmarked when $\mathbb{P}$ (A| Feature$_i$) is high.

Prediction for P:

> P with Feature$_j$ tends to be formally marked when $\mathbb{P}$ (P| Feature$_j$) is low, and unmarked when $\mathbb{P}$ (A| Feature$_i$) is high.

The cross-linguistic distribution of the features (see Section 6.3) suggests that one could expect to find few A's among nominal core arguments (i.e. A and P taken together) and possibly among 3rd person referents because A's with these features tend to be marked cross-linguistically. We can also expect low proportions of P's in pronominal, high-DP, animate or definite and possibly 1st and 2nd person arguments, because P's with these features tend to be marked.

It is assumed explicitly or implicitly that disambiguation interacts with economy. If a context is not ambiguous, or the argument in question has typical properties, it does not require formal marking. Only 'problematic' arguments, which may lead to misunderstanding, require marking.

We should also say a few words about the notion of ambiguity, since the proposed function is that of disambiguation. On the neural level, ambiguity occurs when two nodes in the same domain simultaneously receive similar levels of priming. (MacKay 1987: 133). I will treat here ambiguity as a gradual phenomenon whose maximum is achieved when two intepretations of a linguistic cue are equally probable given all available information from the context, the experience with discourse, etc. In this case, the entropy is 1. When no other interpretation is possible, the entropy is 0. Nearly always, the entropy would be between 0 and 1 for binary outcomes. True ambiguity is rare in language because usually there is enough contextual information:

> …what is surprising (…) is not that people *sometimes* [emphasis by the author – NL] experience difficulty with ambiguity, but that they experience difficulty so rarely (MacKay 1987: 133).

From this follows that true disambiguation is very rare, as well.

6.4.2. Iconicity of markedness and economy

Iconicity of markedness means that semantically marked members of grammatical categories are usually marked formally, while semantically unmarked ones are also formally unmarked. The relationship between semantic and formal marking is iconic (cf. Section 1.2.3 in Chapter

1). This idea has also been used for explanation of differential marking (cf. Dalrymple and Nikolaeva 2011: 3). As noted by Aissen (2003), iconicity of markedness interacts with the principle of economy because semantically unmarked arguments are left formally unmarked. Semantically and formally unmarked subjects are prominent (animate or human, definite and specific), while marked subjects are non-prominent. For objects, the reverse is claimed to be the case (Aissen 2003). This principle can be called **"Mark Untypical Features, Don't Mark Typical Features"**. Since markedness is associated with the relative frequency of two or more members within one category, e.g. singular and plural within the category of number (Greenberg 1966), we need to compute the proportions of particular referents (animate, definite, etc.) in the total number of A or P, and take the least probable ones. One can formulate the predictions as follows:

(10) Reverse engineering predictions based on markedness principle and strategy "Mark Untypical Features, Don't Mark Typical Features"

Prediction for A:
A with Feature_i tends to be formally marked when $\mathbb{P}$ (Feature_i |A) is low, and marked when $\mathbb{P}$ (Feature_i |A) is high.

Prediction for P:
P with Feature_j tends to be formally marked when $\mathbb{P}$ (Feature_j |P) is low, and marked when $\mathbb{P}$ (Feature_i |P) is high.

In fact, this strategy can also be interpreted in terms of disambiguation: the less typical A or P may need additional marking because they are more difficult to identify as such.

If this principle is relevant, A should have a low proportion of nouns and possibly 3rd person referents, while P should have a particularly low number of pronouns, high-DP, animate and possibly 1st and 2nd person referents, since all these are the formally marked features of A and P in the cross-linguistic data (see Section 6.3). The predictions for the three marking strategies are summarized in Table 6.3.

Table 6.3. Reverse engineering predictions for the distribution of features of A and P in discourse

Marking strategy	Prediction for A distribution	Prediction for P distribution	Conditional probability
Mark Confusable Arguments **Don't Mark Non-confusable Arguments**	A is usually pronominal, high DP/animate (1&2 person). A is rarely nominal, low DP/animate (3rd person)	P is usually nominal and 3rd person. P is rarely pronominal and 1st & 2nd person.	ℙ (Feature|Role), the other role
Mark Weak Cues **Don't Mark Strong Cues**	Nominal and 3rd person arguments are rarely A. Pronominal and 1st & 2nd person arguments are usually A.	High-DP/animate, pronominal arguments (and 1st & 2nd person) are rarely P. Low-DP/inanimate, nominal arguments (and 3rd person) are usually P.	ℙ (Role|Feature), the same role
Mark Untypical Features **Don't Mark Typical Features**	A is rarely nominal and rarely 3rd person. A is usually pronominal and 1st and 2nd person.	P is rarely pronominal, high-DP/animate (and 1st & 2nd person). P is usually nominal, low-DP/inanimate (and 3rd person)	ℙ (Feature|Role), the same role

6.4.3. Indexing function and high transitivity

Another explanatory principle suggested in the literature is the identifying, or indexing function of differential marking. As de Hoop and Malchukov (2008) argue, both strong A's and strong P's will be marked because they make for better, more individuated and prominent participants. Animate subjects, are, for instance, volitional and agentive, which is why they make good A's. Animate objects are considered to be more affected than others and therefore make for better P's. This is why, it is argued, DOM is more common than DAM. In DOM, distinguishing and identifying functions overlap: animate or definite P's are both prominent and similar to A (and

therefore need to be distinguished from the latter). In DAM, these principles are in conflict because animate subjects are prominent, but they do not require disambiguation, unlike inanimate subjects (Malchukov 2008). In principle, one could expect all A's to be marked – animate and definite because of their prominence, and inanimate and indefinite because of the disambiguation needs, while the opposite seems to be the case in languages of the world.

This approach goes back to the influential work by Hopper and Thompson (1980), who say that the situations where DOM is observed indicate high transitivity. Highly transitive events involve highly individuated subjects and object, and the object is highly affected. Presumably, definite and animate subject and objects satisfy these conditions, and animate beings are more affected than inanimate ones.

In my opinion, there are a few problematic issues with this account. First, there is disagreement about what constitutes typical transitivity. For instance, Næss (2007) claims that the prototypical transitive clause is the one where the subject and the object are maximally semantically distinct. Second, the use of the notion of affectedness to explain DOM by some researchers has not brought conclusive results so far (see García García 2018). Finally, it is not quite clear what kind of cognitive or pragmatic mechanism is responsible for marking the more salient, more representative, etc. participants, while leaving the other ones unmarked.

6.4.4. Source-based explanations

Cristofaro (In press) argues that universal functional pressures, such as economy-based disambiguation, are not supported by diachronic data. In particular, she says that the distributional restrictions on the use of an additional subject or object marker in DAM and DOM (e.g. only with animate objects or nominal subjects) originate from the reinterpretation of an element of a pre-existing construction with similar distributional restrictions. For instance, some object markers come from topical markers. Since topics are usually definite, animate and pronominal (e.g. *As for me, …*), the marking has spread to the objects with all or some of these features, as in Kanuri:

(11) Kanuri (Nilo-Saharan: Cyffer 1998: 52, cited from Cristofaro, In press)

(a) *Músa shí-ga cúro.*
Musa 3SG-ACC saw
'Musa saw him.'

(b) *wú-ga*
1SG-as.for
'as for me'

The idea of a tight connection between DOM and topic markers has been developed in detail by Dalrymple and Nikolaeva (2011), who claim that objects often serve as secondary topic. Consider an illustration:

(12) - *What has happened to John?*
- *He married Jane. But he doesn't love her.*

The pronoun *her* serves as the secondary topic here because the referent has already been mentioned (*Jane*). A good illustration of what the secondary topic may look like can be found in Ostyak, a Uralic language, where object agreement is required for topical objects, as in (13).

(13) Ostyak (Uralic, Dalrymple and Nikolaeva 2011: 14): topical object

A. What did he do to this reindeer?

B.	*tam*	*kalaŋ*	*we:l-s-əlli*	*/*we:l-əs.*
	this	reindeer	kill-PST-OBJ.3SGSUBJ	kill-PST-3SGSUBJ

'He killed this reindeer.'

However, focal objects disallow agreement, as shown below:

(14) Ostyak (Uralic, Dalrymple and Nikolaeva 2011: 14): focal object

A. Which reindeer did he kill?

B.	tam	kalaŋ	we:l-əs	/* we:l-s-əlli.
	this	reindeer	kill-PST-3SGSUBJ	kill-PST-OBJ.3SGSUBJ

'He killed this reindeer.'

Note that Dalrymple and Nikolaeva's account does not exclude the functional explanations, although it would make the causation less direct. In fact, subjects tend to be the

topic par excellence, while objects can be both topics and foci. In particular, Maslova (2003) reports that transitive sentences in Kolyma Yukaghir (Yukaghir language spoken in the Russian Far East) have mostly topical subjects and topical objects (65%), while topical subjects and focal objects are responsible only for approximately 35%, and less than 1% of all clauses have focal subjects and topical subjects. These numbers suggest that $\mathbb{P}$ (P|Focal) is much higher than $\mathbb{P}$ (A|Focal). Focalness therefore serves as a perfect cue for objecthood, while topical objects need some extra 'help' to be distinguished from subjects, the primary topics. If topical arguments in general are expected to be subjects, we need additional markers for disambiguation of topical objects. That would support the explanations based on disambiguation. However, we need systematic historical evidence that the majority of DOM systems around the world come from topic markers. Such evidence is difficult to provide not only due to the lack of data, but also because topicality is correlated with many other semantic features, as was mentioned above.

Other sources for object markers include various oblique markers, such as the dative marker *a* in Spanish, the genitive case marker *-a* in Russian, the comitative marker *saṇī (haṇī)* 'with' in Garhwali and Kumaoni, Indo-Aryan languages spoken in India (Montaut 2018: 308), or the possessive marker in Wayana, a Carib language spoken in South America. In Wayana, the object marker is used with nominal inanimates, since it has developed from a possessive construction:

(15) Wayana (Carib: Gildea 1998: 201, cited from Cristofaro, In press)

1-pakoro-n *iri* *pək* *wai*

1-house-ACC make occupied.with 1.be

'I'm (occupied with) making my house.' (originally 'I am occupied with my house's making')

Verbs can also be sources for object markers, e.g. *bă* 'hold, take' in Chinese (Yang 1995: 165), the verbal root *lag* 'touch, be stuck to' > *lā* in Marathi (Montaut 2018: 307) or a copula or presentative verb *(-)(ʔ)à* in the Khoe languages (McGregor 2018).

As for DAM, many instances can be described as the absence of the ergative marker on the 1st and 2nd person pronouns. In some cases, the ergative markers developed from instrument markers, which imply inanimate nouns (Cristofaro, In press).

As was already discussed in Section 2.3 of Chapter 4, reanalysis of other source constructions and markers does not constitute evidence against efficiency. Languages create new, more efficient constructions from the available, semantically compatible material. What is important is that all these different scenarios lead to similar outcomes: a highly grammaticalized marker for subjects or objects, used in a similar fashion in diverse languages of the world. It is highly unlikely that such convergence can be explained by the lasting impact of the source constructions alone. One should also ask why such markers do not emerge on the other end of the scales and, if they do, why they do not survive. Moreover, as argued by Seržant (In press), there are clear cases when an additional DOM marker helps to disambiguate between the participants. He provides historical evidence from Old Russian, where the marker appeared exclusively in those contexts where the old morphological distinctions were lost and there was functional pressure for the emergence of DOM.

6.5. Descriptive analysis of corpus data

6.5.1. Data and method

If any of the functional-adaptive accounts presented here in Sections 6.2.1 and 6.2.2 is correct, we will expect to find a correspondence between corpus-based $\mathbb{P}$ (Role|Feature) or $\mathbb{P}$ (Feature|Role) and the cross-linguistic distribution. A and P arguments will be treated separately without considering the properties of the other argument, since the synchronic evidence for co-argument sensitivity (the Malayalam type) is rare.

In fact, some statistical evidence was provided already by Thomson (1909), who looked at transitive verbs in contemporary Russian and found that almost ¾ of them have exclusively a person as the subject. Only 14% of transitive verbs have a human being as the object (e.g. the Russian verbs denoting ‘undress’ and ‘hug’), and more than a half can be used with a human object (e.g. ‘see’ and ‘steal a child’). In contrast, inanimate things are normally subjects for 10% of all transitive verbs. About 45% of the verbs always have an inanimate thing as a patient, and ¾ of the verbs can take an inanimate object. Thomson (1909) does not provide the exact frequencies, unfortunately, nor described his sampling method.

After a long period when language in use was basically neglected, the end of the 20th century witnessed a strong interest in discourse-based explanations of grammatical patterns. There was a lively discussion of the Preferred Argument Structure as the basis for ergativity, as well as quite a few studies of animacy effects in grammar (e.g. Du Bois 1987; Dahl and Fraurud 1996; Dahl 2000; Du Bois et al. 2003; Haig and Schnell 2016). As a very useful by-product of these studies and debates, one can find a considerable amount of available data showing the distributional properties of A and P in different types of corpora and languages. This section examines these data as evidence for or against the predictions formulated in Section 6.4. I chose those studies in which it was possible to find the original frequencies for the parameters of interest and a description of data sources. Learner corpora and language impairment data were not considered. Mild discrepancies in the number of subjects and objects (which were often left unexplained) were tolerated.

The definitions of transitivity varied from one study to another. For example, Sutherland-Smith (1996) uses, in addition to syntactic transitivity, the semantic criterion. If a verb has a highly affected and individuated object, it is considered transitive, even if the marking is dative. Dahl and Fraurud (1996) only include those clauses in which both the subjects and the objects are overt noun phrases. However, since Swedish is not a subject pro-drop language, the results can still be comparable with the other studies, where non-overt arguments are also counted (e.g. the Multi-CAST data). When pro-drop languages were considered, I only included those studies where both overt and covert A's and P's were counted. This was done for the sake of comparability of the results. Moreover, there were some discrepancies in the features. For example, most studies coded humanness, but some describe animacy of A and P. This difference was tolerated, since in my own experience, animals are very infrequently mentioned in contemporary corpora in comparison with humans. Also, most studies count the 1st and the 2nd person reference based on the grammatical properties of the arguments, but Dahl uses the notion 'egophoric', which includes reference to the speech act participants, generic and logophoric reference, as *he* in *Peter says that he is sick.* Due to all these differences, it was problematic to perform a full-fledged confirmatory analysis with statistical inference.

The results that are directly relevant for our research question are summarized in Tables 6.4–6.6. The tables display the percentages which represent different conditional probabilities of the features of A and P. The original corpus counts are provided in Appendix 2. The features

presented there were chosen because they were available in numerous studies. They are the following:

- Lexical, i.e. full nominal phrase vs. pronouns, affixes or zero;
- Human or animate (when specified) vs. all other semantic types;
- '1st and 2nd 'person (or egophoric reference in Dahl's approach) vs. the 3rd person;
- New vs. given and accessible.

Note that I did not aggregate the frequencies of pronouns because there was a lot of variation with regard to the acceptability of zero anaphora in a language. For instance, English allows pronouns to be omitted only in a few cases (most typically, the subject in the imperative), while in Lao zero anaphora is a very common reference-management device for both subjects and objects (see Chapter 1, Section 1.1). This is why I make a distinction between lexical and non-lexical expressions here. Note that definiteness is absent from this list because this feature is very rarely reported. However, preliminary studies based on spoken corpora suggest that the frequencies of definite A's and P's are very similar to those of given arguments (Levshina and Witzlack-Makarevich 2018).

The total number of corpora is nineteen, which represent fourteen languages from seven families from all over the world: the Indo-European (English, French, Northern Kurdish, Persian, Portuguese, Spanish, Swedish), Austronesian (Teop and Vera'a), Afro-Asiatic (Hebrew), Araucanian (Mapudungun), Eskimo-Aleut (Inuktitut), Mayan (Sakapultek) and Sino-Tibetan (Chinese). Different registers and types of discourse are included: spontaneous conversations, transcribed talk shows, narratives retelling films, autobiographical and traditional narratives, stimulus-based monologues in sociolinguistic interviews, child language use and miscellaneous written texts.

6.5.2. Distribution of features within the A role

Table 6.4 shows the distribution of the above-mentioned features within the A role. In other words, the numbers show the proportion of lexical, human, 1st or 2nd person or new A's in the total number of A's. The distributions of these features and their opposites (i.e. nonlexical, non-human, 3rd person and non-new) are also shown in Figure 6.1. The percentages are ordered

from the lowest to highest according to the medians, or the values that separate the first 50% of the distribution from the other half (see the thick black lines inside the boxes). A median is a measure of central tendency that is less prone to the influence of outliers (extremely low or high values) than the mean. Each box contains 50% of the distribution. The whiskers of the plots show 1.5 times the range between 25% and 75% of the distribution. The dots are the minimum or maximum values that lie outside this range. Note that the boxes with opposite features (e.g. New vs. NonNew) are symmetric around the middle of the plot (50%) because, for example, $\mathbb{P}$ (New|A) = 1 – $\mathbb{P}$ (NonNew|A).

The data show that typical A's are human/animate, not new and not lexical. Therefore, untypical A's are new, non-human/inanimate and lexical. The person varies a lot. The highest proportion of the 1st and 2nd person arguments is observed in Inuktitut child language (97.4%), followed by the English autobiographical narratives and Swedish spontaneous conversations. The lowest is observed in the Persian stimulus-based narratives (3.6%).

Do these results support any of the predictions based on the conditional probabilities of a feature given A? Recall that these probabilities are relevant for two marking strategies, namely, "Mark Confusable Arguments, Don't Mark Non-Confusable Arguments" and "Mark Typical Features, Don't Mark Untypical Features". As for the strategy "Mark Confusable Arguments, Don't Mark Non-Confusable Arguments", we expected A's to be predominantly pronominal and high-DP/animate (and 1st and 2nd person). The A arguments in the corpora are overwhelmingly non-lexical, non-new and human/animate. There is no preference for the 1st and 2nd person, but this feature does not play a very important role in the differential case marking of P. Therefore, we can say that the predictions are mostly supported by the corpus data. Since the proportions of different values of the same feature (e.g. human and non-human) are mirror images of each other, the conclusion for the second part of the strategy, "Don't Mark Non-Confusable Arguments" is the same.

Table 6.4. Distribution of features of A (transitive subjects) within the role

Study	**Corpus**	**Lexical**	**Human**	**1st & 2nd**	**New**
Du Bois 1987	*Pear Story* narratives in Sakapultek	6.1%	100%	NA	3.2%
Chui 1992	*Ghost* narratives in Chinese	38.1%	NA	NA	2.9%
Ashby & Bentivoglio 1993	interviews (monologues) in French	6.7%	NA	NA	0%
	interviews (monologues) in Spanish	6.1%	NA	NA	0.4%
Dahl & Fraurud 1996	written Swedish	NA	56%	NA	NA
Sutherland-Smith 1996	Several oral narratives in modern Hebrew	6.7%	NA	NA	2.2%
Dahl 2000	spontaneous conversations in Swedish	NA	93.2% (animate)	60.7% (ego)	NA
Allen & Schröder 2003	Inuktitut child language	1.1%	99% (animate)	97.4%	0.7%
Arnold 2003	Mapudungun narrative texts	14.9%	NA	NA	1.2%
Everett 2009	English talk shows	9.7%	91.8%	NA	NA
	Portuguese talk shows	17.1%	87.1%	NA	NA
Lin 2009	Chinese conversations	20%	NA	NA	15%
	Chinese narratives	18.8%	NA	NA	12.5%
	Chinese written texts	15.9%	NA	NA	20.5%
Schiborr 2016	English autobiographical narratives	8.2%	92.8%	59%	NA
Haig & Thiele 2016	Northern Kurdish traditional narratives	13%	96.7%	32.5%	NA
Abidifar 2016	Persian stimulus-based narratives	13.6%	96.2%	3.6%	NA
Mosel & Schnell 2016	Teop traditional narratives	9.7%	95.4%	25.6%	NA
Schnell 2016	Vera'a traditional narratives	7.4%	94.7%	15.4%	NA

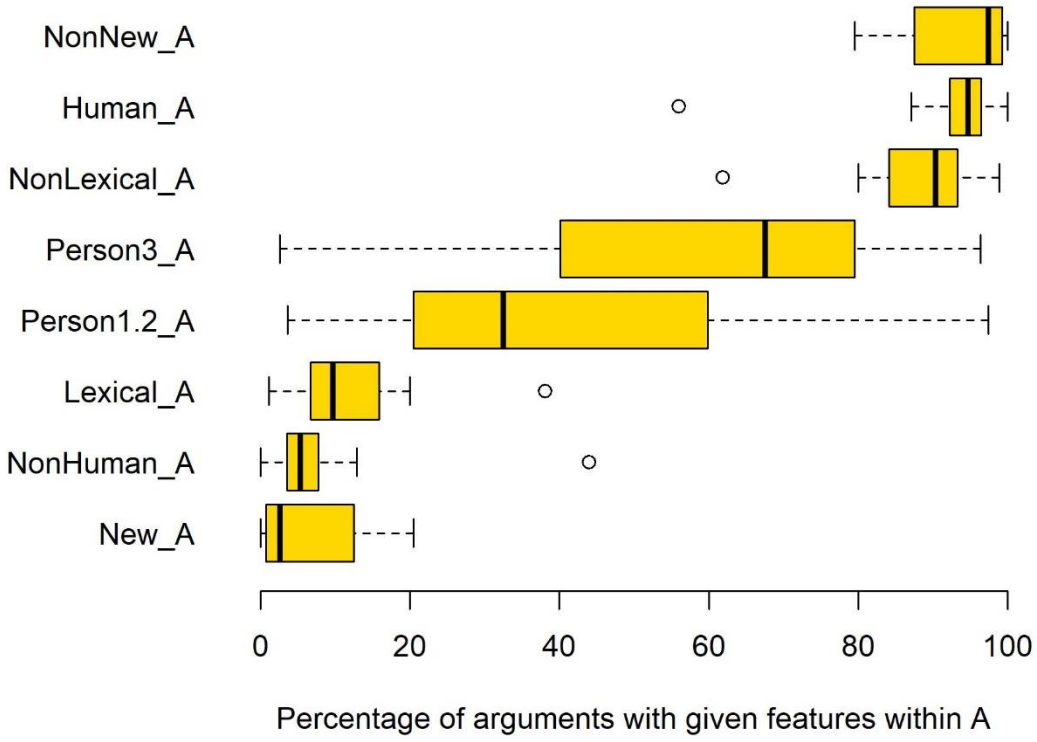

Figure 6.1. Distribution of features within the A role

As for the strategy "Mark Untypical Features, Don't Mark Typical Features", we expected A to be infrequently nominal and frequently pronominal, and also rarely have the 3rd person reference, and frequently 1st and 2nd person reference. Indeed, we find relatively few lexical A's. However, the 3rd person predominates in the data. Overall, the distribution of the person values is very scattered. The 3rd person seems to be frequent in traditional narratives, and the 1st and the 2nd person in autobiographic narratives, child speech and spontaneous dialogues. We can conclude that the predictions are met only partly. In addition, non-human/inanimate and new A's are also rare, but these features are rarely marked formally by languages, as was shown above. Judging from the very low probabilities of non-human and new A's, it is surprising that the languages do not mark them systematically.

6.5.3. Distribution of features within the P role

Table 6.5. Distribution of features of P within the role

Study	**Corpus**	**Lexical**	**Human**	**1st & 2nd**	**New**
Du Bois 1987	*Pear Story* narratives in Sakapultek	45.8%	10%	NA	24.7%
Chui 1992	*Ghost* narratives in Chinese	84.3%	NA	NA	33.6%
Ashby & Bentivoglio 1993	interviews (monologues) in French	67.4%	NA	NA	29.7%
	interviews (monologues) in Spanish	59.7%	NA	NA	24.9%
Dahl & Fraurud 1996	written Swedish	NA	13%	NA	NA
Sutherland-Smith 1996	Several narratives in modern Hebrew	56.3%	NA	NA	23.9%
Dahl 2000	spontaneous conversations in Swedish	NA	16.4% (animate)	4.3%	NA
Allen & Schröder 2003	Inuktitut child language	6%	21.1% (animate)	14.3%	27%
Arnold 2003	Mapudungun narrative texts	85.1%	NA	NA	47.8%
Everett 2009	English talk shows	59.7%	12.6%	NA	NA
	Portuguese talk shows	84.7%	6.1%	NA	NA
Lin 2009	Chinese conversations	80%	NA	NA	55.6%
	Chinese narratives	94%	NA	NA	72.5%
	Chinese written texts	81.8%	NA	NA	70.1%
Schiborr 2016	English autobiographical narratives	47.8%	12.4%	4.8%	NA
Haig & Thiele 2016	Northern Kurdish traditional narratives	54.7%	25.9%	6.8%	NA
Abidifar 2016	Persian stimulus-based narratives	52.7%	18%	0%	NA
Mosel & Schnell 2016	Teop traditional narratives	43%	43.3%	4.7%	NA
Schnell 2016	Vera'a traditional narratives	56.1%	35.3%	8.6%	NA

Table 6.5 shows the proportions of the features within the P role. The distributions are also displayed in Figure 6.2. The numbers represent the proportion of lexical, human, 1st and 2nd person, or new arguments in all occurrences of P in the samples. The most typical properties of P are the 3rd person reference and non-humanness/inanimacy. The other features display substantial variation. The proportion of lexical P's varies from 6% in Inuktitut child language to 94% in Chinese narratives. The fraction of new arguments also fluctuates from 24.7% in Sakapultek to 72.5% in Chinese narratives.

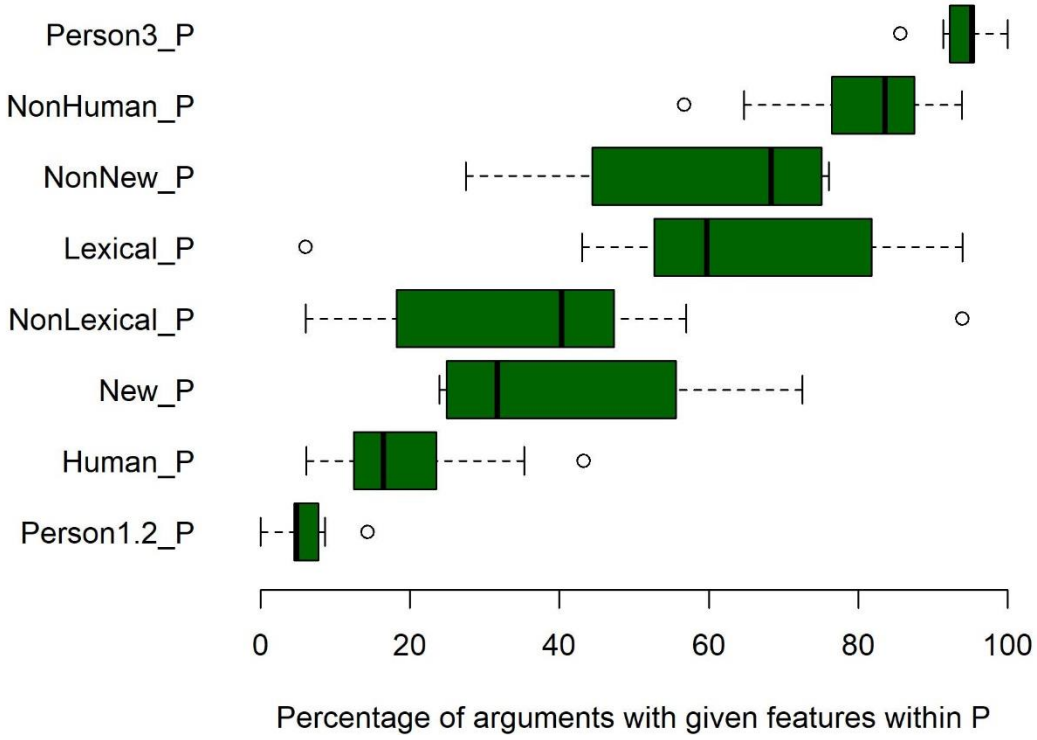

Figure 6.2. Distribution of features within the P role

Are our predictions borne out? As for the strategy "Mark Confusable Arguments, Don't Mark Non-Confusable Arguments", we expected to find predominantly nominal, or lexical P's, and the 3rd person referents, and rarely non-lexical and 1st and 2nd person referents. Most of the P's are indeed lexical, although there are a few exceptions. The 3rd person P's are indeed predominant. Thus, we can say that this strategy is mostly supported.

The other relevant strategy, "Mark Untypical Features, Don't Mark Typical Features", is partly supported by the data. We expected P's to be rarely pronominal, animate/ high-DP, and 1st and 2nd person, and frequently nominal, inanimate, low-DP and 3rd person. These features should be infrequent in discourse. P's in discourse are indeed rarely animate/human and 1st and 2nd person, but they are neither overwhelmingly lexical, nor new. Therefore, these predictions are only partly supported.

Table 6.6. Distribution of the roles within the features (only A shown)

Study	Corpus	Lexical A	Non-lexical A	Human A	Non-Human A	1st & 2nd A	3rd A	New A	Non-new A
Du Bois 1987	*Pear Story* narratives in Sakapultek	11.9%	63.8%	91.7%	0%	NA	NA	12.5%	58.6%
Chui 1992	*Ghost* narratives in Chinese	33.6%	81.6%	NA	NA	NA	NA	9%	62.1%
Ashby & Bentivoglio 1993	interviews (monologues) in French	9%	74.1%	NA	NA	NA	NA	0%	58.7%
	interviews (monologues) in Spanish	9.3%	70%	NA	NA	NA	NA	1.4%	57%
Dahl & Fraurud 1996	written Swedish	NA	NA	75.3%	25.9%	NA	NA	NA	NA
Sutherland-Smith 1996	Several narratives in modern Hebrew	14%	74.6%	NA	NA	NA	NA	11.3%	63.8%
Dahl 2000	spontaneous conversations in Swedish	NA	NA	88.8% (animate)	6.9%	92.9% (ego)	27.5%	NA	NA
Allen & Schröder 2003	Inuktitut child language	15.9%	51.3%	82.7% (animate)	1.4%	87.2%	2.9%	2.4%	58.1%
Arnold 2003	Mapudungun narrative texts	14.9%	85.1%	NA	NA	NA	NA	2.5%	65.4%
Everett 2009	English talk shows	13.8%	68.9%	87.8%	8.4%	NA	NA	NA	NA
	Portuguese talk shows	16.4%	83.7%	93.1%	11.6%	NA	NA	NA	NA
Lin 2009	Chinese conversations	20%	80%	NA	NA	NA	NA	21.7%	66.4%
	Chinese narratives	16.1%	92.9%	NA	NA	NA	NA	14.7%	76.1%
	Chinese written texts	16.3%	82.2%	NA	NA	NA	NA	22.8%	72.9%
Schiborr 2016	English autobiographical narratives	13.9%	62.3%	87.6%	7.1%	92%	28.8%	NA	NA
Haig & Thiele 2016	Northern Kurdish traditional narratives	19%	65.4%	78.6%	4.2%	82.5%	41.7%	NA	NA
Abidifar 2016	Persian stimulus-based narratives	19.9%	63.7%	83.7%	4.3%	100%	48.1%	NA	NA
Mosel & Schnell 2016	Teop traditional narratives	22.5%	67.2%	74%	9.6%	87.6%	50.3%	NA	NA
Schnell 2016	Vera'a traditional narratives	16.4%	75.8%	79.9%	10.8%	72.7%	57.8%	NA	NA

6.5.4. Distribution of the roles within the features

Table 6.6 displays the distribution of the roles (only A) within each feature. The percentages stand for the proportions of A's and P's among all arguments in the sample with a given feature. To obtain the corresponding proportions for P, one can simply subtract these numbers from 100%. Figure 6.3 displays the results for all combinations of roles and features.

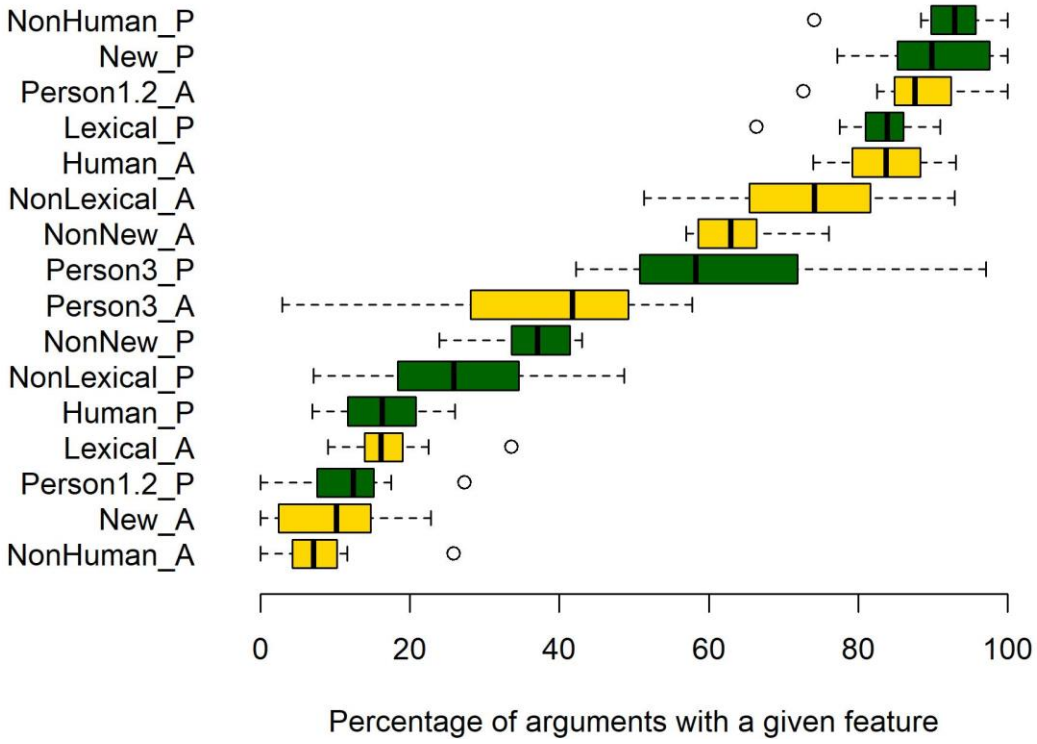

Figure 6.3. Distributions of A or P within a feature

With the exception of the 3rd person (see the middle of the plot), the distributions are quite compact and mostly contain values less than 50% or greater than 50%. Overall, if an argument (A or P) is non-lexical, human, 1st or 2nd person, and not new, it is more likely to be A; if it is lexical, non-human and new, it is more likely to be P. As for the 3rd person arguments, they tend to be P rather than A, but there are exceptions in the data. In general, the results support the idealized scales (see Section 6.2).

Do the results support the predictions of the strategy "Mark Weak Cues, Don't Mark Strong Cues"? Let us first have a look at the weak cues and the predictions for the first part, "Mark Weak Cues". The relevant information is in the left-hand part of the plot with low probabilities ℙ (Role|Feature). We expected nominal (lexical), and 3rd person arguments to be A only infrequently. This is supported by the corpus data mostly, although two out of seven data points of the 3rd person are slightly greater than 50%. Also, we expected pronominal, high-DP/ animate and 1st and 2nd person arguments to be P's only rarely. This is also what we find in the corpora: non-lexical, non-new and human/animate arguments are unlikely to be P's, as well as the 1st and 2nd person arguments. Therefore, the predictions are supported. The cues that are marked are indeed weak. Still, there are a few features with low probability of A (i.e. non-human A and new A) that do not lead to widely attested marking in the cross-linguistic data. Therefore, the strategy overgenerates predictions.

If we focus on the second part of the strategy, "Don't Mark Strong Cues" and examine the features with particularly high median probabilities ℙ (Role|Feature), we will see that the predictions are met both for A (non-lexical and 1st & 2nd person) and for P (lexical, non-human, new, animate and predominantly 3rd person).

6.6. Summary and discussion

This study tested reverse-engineering predictions based on the cross-linguistic distributions of differential case marking. These predictions were formulated for three marking strategies, which are based on functional-adaptive explanations of the typological data involving disambiguation pressure and iconicity of markedness. I expected to find relative frequencies of different A and P that would account for the typological data. A summary of the results for the three strategies is shown in Table 6.7.

The weakest support is found for the strategy based on iconicity of markedness, "Mark Usual Features, Don't Mark Unusual Features". Its predictions are supported only partly. The first strategy based on disambiguation "Mark Confusable Arguments, Don't Mark Non-Confusable Arguments" fares better. Most of its predictions are supported. The exceptions concern either the less important features (i.e. 1st and 2nd person in P marking) or a minor fraction of data points that are outside the predicted range. The strategy "Mark Weak Cues, Don't Mark Strong Cues" provides the best account of the observed cross-linguistic tendencies,

especially the second part “Don’t Mark Strong Cues”. This strategy is also based on the notion of disambiguation, although, as was argued in Section 6.4.1, true ambiguity is very rare in language. From the efficiency perspective, this strategy should be regarded as an example of the Low- and High-Cost Heuristics at work. For example, in the case of DOM with an animate – inanimate split, the Low-Cost Heuristic guides the hearer, who processes an inanimate argument, that the most probable interpretation (i.e. the P role) should be taken. The High-Cost Heuristic tells the hearer, who processes an animate argument with additional marking, that the speaker uses the costlier expression because the P-role is less probable for this kind of arguments. This reasoning is similar to Haspelmath’s hypothesis about differential object marking and animacy:

> Since more inanimate nominals have P-function than animate nominals, hearers are less surprised when they encounter an inanimate P-argument and have less need for special coding (Haspelmath Forthcoming-a: 11).

Table 6.7. Reverse-engineering redictions and the data

Marking strategy	Prediction for A distribution	Prediction for P distribution
Mark Confusable Arguments	Mostly	Mostly
Don’t Mark Non-Confusable Arguments	Mostly	Mostly
Mark Weak Cues	Mostly	Yes
Don’t Mark Strong Cues	Yes	Yes
Mark Untypical Features	Partly	Partly
Don’t Mark Typical Features	Partly	Partly

These principles explain why the marker spreads on the arguments with low cue validity, and why the arguments with high cue validity remain unmarked in diachrony. Consider an example. The Spanish DOM has undergone little change for inanimate P's since Old Spanish, while P's with low cue validity – i.e. definite human nouns – have gradually acquired the obligatory marker *a*. Indefinite humans have variable marking (García García 2018: Section 3.4). A similar development is observed in Persian. The object marker *râ* first appeared on the first and second person pronouns, then spread to definite animate objects, then definite objects, but was still optional. In Modern Persian, it became obligatory on all definite objects and topical indefinite objects (Dalrymple and Nikolaeva 2011: 203). Focal indefinite objects thus remain the last bastion of zero marking. This persistence is based on the Low-Cost Heuristic.

Moreover, the conditional probabilities of A and P given specific features of the arguments, i.e. $\mathbb{P}$ (Role|Feature), which are associated with the winning strategy, are distributed quite uniformly across different registers and languages. It is logical to assume that cross-linguistic universals should correspond to similar distributions across and within languages. Therefore, the homogeneous distributions of $\mathbb{P}$ (Role|Feature) are good candidates for explaining the universal scale effects.

Some of the strategies overgenerate predictions. This means that there are discourse tendencies that do not have correspondences in the cross-linguistic distribution of differential case marking. Yet, overgeneration is not very dangerous because efficiency is a relatively weak factor, which is often overridden by analogical pressure (see Section 2.6 in Chapter 2), which would lead to uniform marking of an argument in question. In particular, this may be the case for A, which displays much fewer cases of differential marking than P.

There remain some questions and tasks for future research. First, one needs to provide more data from dialogical corpora, in particular, additional data for the grammatical person and new data for definiteness. This has been attempted by Levshina and Witzlack-Makerevich (2018), whose preliminary results corroborate the ones reported here. Second, it may be useful to try more sophisticated measures of attraction and repulsion (see Chapter 1, Section 1.1), which have been proposed in the psychological and information-theoretic literature. Third, it will be useful to perform multifactorial analyses in order to see how other factors, such as agreement and word order, may influence the marking in the languages which display variation in differential case marking. These tasks are left for the future.

Part IV. Efficiency and slot-filler predictability in English constructions

Chapter 7. The use of *help* with bare or *to*-infinitive in Present-Day English

7.1. Aims of this chapter

The fourth part of this study describes several English constructions:

- *help* + (*to*) Infinitive, as in *I helped him (to) install the software*.
- Stative verb + (at) home, as in *More and more young fathers stay (at) home.*
- *go (and)* Infinitive, as in *Go (and) bring me a beer.*
- *want to* + Infinitive vs. *wanna* + Inifinitive, as in *Girls just wanna/want to have fun*.

This part focuses on the use and omission of function words (the particle *to*, proposition *at* and conjunction *and*), as well as phonological and orthographic reduction (*want to* vs. *wanna*). As in the previous case studies, the relevant source of information that determines this variation is previous experience with discourse. More exactly, we are going to speak about the occurrence of the lexical slot fillers in the slots of the constructions in question, e.g. *help + (to) install, make, understand, get*, etc., and in other constructions. This knowledge can be represented in the form of conditional probabilities of a lexical slot filler given the construction and the other way round. Based on this information and the Principle of Communicative Efficiency, one can formulate the following hypothesis:

(1) Hypothesis of Slot-Filler Predictability and Formal Length

The less probable a slot filler given the constructional slot or the slot given the filler, the greater the chances of the longer constructional alternative; conversely, the more

probable the slot given a filler or the other way round, the greater the chances of the shorter variant.

In a series of quantitative analyses I will present statistical evidence that these probabilities allow us to predict the choice between the alternating variants, such that higher probabilities are associated with the shorter variants, and lower probabilities are associated with the longer variants. Interestingly, different directions of predictability are important in different cases, as will be shown below. Such evidence has never been presented before.

The present chapter focuses on the construction *help* + (*to*) Infinitive in Present-Day English. Section 7.2 discusses the previous research, which shows that this variation depends on a multitude of factors. Section 7.3 presents the corpus data from Google Books Ngrams. In Section 7.4, the results of statistical analyses based on Generalized Additive Models will be presented. Section 7.5 summarizes the results and provides a discussion.

7.2. Multifactorial probabilistic variation of *help*

The present paper investigates the English construction with *help* followed by the infinitive with or without *to*, as in (2):

(2) a. *If this book does not **help** you **to survive** the Zombie Apocalypse, a full refund may be obtained from the author.*[37]

b. *Just to be on the safe side you might want to start doing these 8 exercises that will **help** you **survive** the zombie apocalypse.*[38]

The construction *help* + (*to*) Infinitive is a rare case where the choice between the bare and *to*-infinitive is possible in Present-Day English. Different factors have been proposed to explain when one or the other variant is preferred. For instance, it has been argued that the variant with the bare infinitive designates a more active involvement of the Helper in carrying

[37] https://www.amazon.co.uk/Z-Day-UK-surviving-Apocalypse-Britain/dp/1490389873 (last access 1.08.2018).
[38] http://steadystrength.com/8-exercises-that-will-help-you-survive-the-zombie-apocalypse/ (last access 1.08.2018).

out the event expressed by the infinitival complement (Dixon 1991: 199). Consider the following examples, which were already mentioned in Section 1.2.4:

(3) (Dixon 1991: 199)

a. *John **helped** Mary **eat** the pudding (he ate half).*

b. *John **helped** Mary **to eat** the pudding (by guiding the spoon to her mouth, since she was still an invalid).*

When *to* is omitted, as in (3a), the sentence is likely to describe a cooperative effort where Mary and John ate the pudding together; when *to* is included, as in (3b), the sentence means that John acted as a facilitator for Mary, who actually ate the pudding herself (Dixon 1991: 199; 230). Similarly, Duffley (1992: Section 2.3) suggests that the use of the *to*-infinitive evokes help as a condition that enables the Helpee to bring about the event denoted by the infinitive. It has also been argued that animate Helpers have a potentially greater involvement in the event (Lind 1983). Indeed, Lohmann (2011) finds that animate Helpers have higher odds of the bare infinitive than inanimate Helpers, although the effect is not very strong.

Yet, many researchers have questioned the relevance of this semantic distinction. For example, Huddleston and Pullum (2002: 1244) argue that there are numerous contexts and examples where this distinction cannot be traced. Similar claims were made by McEnery and Xiao (2005).

Another relevant factor is the principle of avoidance of identity, or *horror aequi***.** *Horror aequi* is a widespread tendency to avoid repetition of identical elements (Rohdenburg 2003). This idea is also known as the Obligatory Contour Principle, which has been first formulated for phonology (Leben 1973), but has been used to explain different phenomena at all linguistic levels since then (e.g. omission of optional *that* in Walter and Jaeger 2008). Rohdenburg uses *horror aequi* to explain why the *to*-infinitive tends to be avoided immediately after a governing *to*-infinitive (e. g. *to try to do*). When the verb *help* is itself preceded by *to*, the following infinitive is usually without *to* (Biber et al. 1999: 737). See an example in (4):

(4) *Sorry, but how is this supposed **to help** answer the question?* (GloWbE, GB, general, 303502) [39]

Next, one should mention the principle of minimization of cognitive complexity (Rohdenburg 1996). It was already discussed in Chapter 1 (Sections 1.2.4 and 1.2.5). The more words there are between *help* and the infinitive, the more difficult it is to recognize the latter as an element of the construction. Consider an example of a complex environment in (5), where the distance between *help* and the infinitive is six words.

(5) *…it's a way for me to make a contribution, to **help** the country in a small way **to get back** on its feet.* (GloWbE, GB, Blogs, 3069710)

The longer the distance, the more likely it is that the infinitive will be marked by the particle *to*. This effect was already explained from the efficiency perspective in Section 1.2.4. Moreover, there is an interaction between this factor and complexity: The more words there are between *help* and the infinitive, the weaker the influence of *horror aequi* (Lohmann 2011).

It has also been shown that the inflectional forms of the verb *help* have different "preferences" with regard to the bare or *to*-infinitive. In particular, Lohmann (2011) observes that the form *helping* tends to be more frequently used with the *to*-infinitive in British English than the other inflectional forms of *help*. According to Rohdenburg (2009: 317), the effect of *helping* has an analogy with *daring* and *needing*, which differ from all forms of *dare* and *need* by being virtually always associated with marked infinitives. In addition to that, there is a weakly significant preference of the third person singular form *helps* for the *to*-infinitive in comparison with the base form (Lohmann 2011).

The presence or absence of the Helpee also plays a role. Biber et al. (1999: 735) show that the bare infinitive is particularly dominant in the pattern *help* + NP + infinitive clause. This observation is also supported by Lohmann (2011). Similarly, it matters whether the infinitive is passive or active: According to McEnery and Xiao (2005), the passive infinitive should

[39] This annotation means that the sentence is taken from the GloWbE corpus, general subcorpus from Great Britain (not blogs only), website ID 3039502.

always be marked with *to*. However, one can find examples of both the bare and *to*-forms of passive infinitives in English corpora, as shown in (6).

(6) a. *If rural voices are important – the bread basket, our farmers, our miners – then an electoral approach, not a pure popular vote,* ***helps*** *them* ***to be heard***. (GloWbE, USA, general, 288902)

b. *Thank you so much for sharing and* ***helping*** *our Vets* ***be heard****!* (GloWbE, USA, blog, 3177307)

Moreover, the shorter variant with the bare infinitive is considered to be less formal than the one with the marked infinitive (e.g. Rohdenburg 1996: 159; see also Biber et al. 1999: 736–737).

Finally, one should also mention phonological factors. There is some evidence that the use of *to* in different constructions depends on prosody. Wasow et al. (2015), in particular, found an effect of prosody on the use of the bare or *to*-infinitive in their investigation of the DO-BE construction, e.g. *All we want to do is (to) celebrate*. Namely, they discovered that *to* was used to eliminate stress clash when both the copula and the first syllable of the infinitive after *be* were stressed. I'm not aware of any studies of *help* that focused directly on the effect of stress clash. However, Lohmann (2011) tested two other phonetic variables, namely, if the infinitive begins with the vowel, and whether the first syllable of the infinitive is stressed. Neither of the variables had a significant effect on the choice between the forms of the infinitive.

7.3. Measures of slot-filler predictability

This case study tests the hypothesis that slot-filler predictability of the *help*-construction and the verbs that fill in the infinitival slot and the use of the longer or shorter variant are correlated. There is plenty of evidence that language users learn and store information about the probabilistic relationships between constructions and their slot fillers (e.g. Goldberg et al. 2004; Goldberg et al. 2005; Gries et al. 2005; Ellis and Ferreira-Junior 2009; Taylor 2012). These relationships are directional (Schmid 2000; Schmid and Küchenhoff 2013). In order to

represent these relationships, we will compute two measures. The first measure is the probability of a slot filler given the constructional slot. More exactly, we will use the informative content, a derived measure, which represents a log-transformed inverse of the corresponding conditional probability, as shown in (7):

(7) $$I_{Filler|Slot} = -log_2 \mathbb{P}(Filler|Slot) = -log_2 \frac{\mathbb{P}(F,S)}{\mathbb{P}(S)}$$

where $\mathbb{P}$ (F, S), which stands for the probability of the filler in the slot, and $\mathbb{P}$ (S) represents the probability of the slot. In practice, $\mathbb{P}$ (F, S) is measured as the token frequency of the filler in the slot, whereas $\mathbb{P}$ (S) in represented by the token frequency of the slot (or the frequency of the construction, which is identical) in a corpus. Due to the negative log-transformation, high probabilities correspond to low information content, whereas low probabilities correspond to high information content. Log-transformations of frequency data are ubiquitous in quantitative linguistics. They allow us to zoom in on the small differences between the low-frequency units and soften the effect of outliers (very large frequencies). In addition, the measure can be interpreted from an information-theoretic perspective (cf. Section 1.1). The logarithm base 2 is commonly used in information theory to measure information in bits.

The second probabilistic measure is the information content of the slot given a particular slot filler. It is shown in (8):

(8) $$I_{Slot|Filler} = -log_2 \mathbb{P}(Slot|Filler) = -log_2 \frac{\mathbb{P}(F,S)}{\mathbb{P}(F)}$$

where $\mathbb{P}$ (F, S), again, can be obtained from the token frequency of the filler in the slot, and $\mathbb{P}$ (F) is represented by the token frequency of the filler in a corpus.

A cognitive and pragmatic interpretation of the measures in (7) and (8) can be as follows. Information content of a filler given a slot in (7), e.g. $\mathbb{P}$ (*understand*|Inf-slot$_{help}$), where Inf-slot$_{help}$ stands for the infinitival slot in the construction, represents how expected the verb is if we encounter the construction. One can see it as category validity, where the construction is the category and the verb is the cue (see a similar approach in Goldberg et al. 2005).

As for information content of a slot given a filler in (8), e.g. $\mathbb{P}$ (Inf-slot$_{help}$|*understand*), it corresponds to the validity of the verb as a cue for the infinitival slot of the construction. In other words, it is similar to $\mathbb{P}$ (Function|Form), where the form of the verb as a lexeme is a cue, and its syntactic function is the category. The question is, if one encounters a verb, how likely it is to be a part of the construction with *help*? As was shown in the previous chapter about DCM, this type of probabilistic information is essential for efficiency because it allows the speaker to reduce the amount of effort based on the Low-Cost Heuristic.

These two measures have analogues in usage-based construction linguistics, which are known as *Attraction*, i.e. the conditional probability of a word given a construction, and *Reliance*, i.e. the conditional probability of a construction given a word (Schmid 2000), or Faith (Gries et al. 2005). Due to the negative log-transformation, the greater information content, the smaller Attraction or Reliance (Faith). Although many corpus linguists find it useful to compute one bidirectional measure that represents the association between a construction and one of its collexemes (e.g. Stefanowitsch and Gries 2003 and later work), Schmid has been arguing that *Attraction* and *Reliance* represent two different types of information, each valuable on its own (e.g. Schmid and Küchenhoff 2013). To the best of my knowledge, the effect of these two types of predictability – predictability of a lexeme given the construction and the other way round – on the formal length has not been investigated in the previous studies of morphosyntactic alternations.

7.4. Quantitative analysis of slot-filler predictability

7.4.1. Data and method

Since the construction with *help* is not very frequent, and the distribution is skewed in favour of the bare infinitive, especially in American English and in informal registers, I used the British dataset of Google Books Ngrams, which is based on books published in Great Britain.[40] I used Version 2 (marked as 20120701), which provides part-of-speech tags and is based on improved OCR methods and more accurate metadata than Version 1. For my analyses, I used 1-grams, 2-grams, 3-grams and 4-grams with POS tags. For modern English, Google promises

[40] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html (last access 30.07.2018).

the accuracy of the part-of-speech tags to be around 95%, and likely above 90% for older English texts.[41] The data represent mostly formal registers (e.g. academic publications).

Recall that the previous case studies in Chapter 3 and 6 provided the frequencies from spoken spontaneous corpora. Written data have been used in this case study. In fact, predictability effects have been observed in written data, as well, including the construction with *help* (Levshina 2018). As Wasow et al. (2015) hypothesize, this may happen because the speech habits are carried over to writing, or because of temporal pressures on readers. Regardless of the explanation, the predictability effects found in writing are robust enough to test the main hypothesis of the present study on data from a written corpus.

Only a fraction of these huge dataset was used, representing the years from 2001 to 2009. Instances of the construction were extracted from the datasets with 2-grams, 3-grams and 4-grams with the help of a Python script. The script searched for *help*, *helped*, *helping* or *helps* tagged as a verb and followed immediately a) by another verb in the base form with or without *to* or b) by a personal pronoun and a base verb form with or without to. More exactly, I extracted the following patterns, where X stands for any string, Y denotes object personal pronoun *me*, *you*, *him*, *her*, *it*, *us* and *them*, and * represents any ending, including zero:

- 2-grams: help*_VERB X_VERB, e.g. *helps_VERB make_VERB*
- 3-grams: help*_VERB Y_PRON X_VERB (3-grams), e.g. *helped_VERB me_PRON build_VERB*
- 3-grams: help*_VERB to_PRT X_VERB, e.g. *helping_VERB to_PRT achieve_VERB*
- 4-grams: help*_VERB Y_PRON to_PRT X_VERB, e.g. *help_VERB her_PRON to_PRT understand_VERB*

Both upper-case and lower-case characters were allowed. In addition, 1-grams were used for extraction of verb frequencies. The verbs in the open slot were later manually checked, and the finite forms, participles and misspellings were excluded.

By making this contextual restriction, it was possible to control for some of the relevant factors that influence the use of one or the other variant: linguistic distance (zero or one word) and the Helpee (explicit or implicit). The morphological form of *help* was also controlled for. The restrictions of the Helpees to personal pronouns are explained by the size of *n*-grams and by concerns about possible spurious hits. Previous work based on the GloWbE corpus

[41] https://books.google.com/ngrams/info, last access 18.06.2018.

(Levshina 2018) suggests that zero and pronominal Helpees account for approximately 80% of all uses of the construction in the data from both countries. I expect, therefore, no substantial data loss.

Table 7.1. Frequencies of different subschemata of the construction with *help*

Context	Total frequency	Frequency of the *to*-infinitive	Frequency of the bare infinitive	Number of verb types
help + Inf	897,120	328,329	568,791	1,329
	(100%)	(36.6%)	(63.4%)	
helped + Inf	459,042	273,218	185,824	1,354
	(100%)	(59.5%)	(40.5%)	
helps + Inf	295,028	193,759	101,269	873
	(100%)	(65.7%)	(34.3%)	
helping + Inf	120,815	106,905	13,910	750
	(100%)	(88.5%)	(11.5%)	
help + Helpee + Inf	497,241	155,565	341,676	688
	(100%)	(31.3%)	(68.7%)	
helped + Helpee + Inf	87,622	41,619	46,003	321
	(100%)	(47.5%)	(52.5%)	
helps + Helpee + Inf	73,982	41,687	32,295	236
	(100%)	(56.3%)	(43.7%)	
helping + Helpee + Inf	40,177	22,782	17,395	210
	(100%)	(56.7%)	(43.3%)	

Manual cleaning was performed in order to exclude inflected forms and misspellings. The verbs were normalized with regard to the spelling variant (e.g. *organise* and *organize* were treated as one lemma). The total number of occurrences of the construction was 2,471,027, and the total number of individual verbs was 1,672. The relative frequencies of the *to*-infinitive and bare infinitive were very similar: 47.1% and 52.9%, respectively.

The frequencies for separate combinations of the forms of *help* and the presence or absence of the Helpee are given in Table 7.1. Notably, we observe a correlation between the relative frequency of the bare form and the total frequency of the form of *help* followed by the (*to*) Infinitive with and without the Helpee. The higher the total frequency, the higher the relative frequency of the bare form. This can be regarded as a manifestation of efficiency. The higher the probability of a constructional schema in discourse, the shorter its form.

The frequencies of use with both variants were summed for each individual verb. The total frequencies of the verbs were obtained from the file with 1-gram frequencies in the entire British English dataset for the period from 2001 to 2009. Based on these frequencies, the two information content measures were computed as described in Section 7.2.

Consider an example. The verb *understand* occurs in the construction with *help* 85,815 times. The total frequency of the construction is 2,471,027. Therefore, the information content of *understand* given the constructional slot is $-log_2 (85{,}815/2{,}471{,}027) \approx 4.85$. The verb occurs 3,239,809 times in the entire data set. From this follows that the information content of the slot given the verb is $-log_2 (85{,}815/3{,}239{,}809) \approx 5.24$.

Figure 7.1 displays the distribution of the information content scores of individual verbs in the dataset. One can see that the measures are to some extent correlated (Spearman's $\rho = 0.59$, $p < 0.0001$). The bottom left corner contains the verbs that have low information content according to both measures. These are the verbs like *get, understand, keep, make* and *explain,* which occur very frequently in the construction, and the construction also represents a large proportion in the total uses of these verbs. One can say that there is mutual attraction between the construction and these verbs.

In the top left quadrant one can find very frequent verbs like *be, have, do, see, say* and *go*. They have high information content of the slot given the verb, due to their very high frequency of occurrence elsewhere. At the same time, they occur in the construction quite frequently.

The top right quadrant represents the verbs with high information content according to both variables. These are high-frequency verbs that are not very well compatible with the semantics of the construction, e.g. *like, seem, wish, suppose, worry, hate* and *depend*. Many of them are mental and stative verbs.

Finally, the bottom right quadrant contains low-frequency verbs, which apparently occur mostly in this construction, which explains why they have low information content of the slot given the verb. These are dialectal variants, e.g. *nuss* (a dialect variant of *nurse*), *bigg* (a dialectal variant of *build*), rare negative derivatives, such as *unshawl*, *unarm* and *unrobe*, and some very rare Latinate verbs, e.g. *tenderise* and *solubilise* (note that both spelling variants, with *-ise* and *-ize* were taken into account).

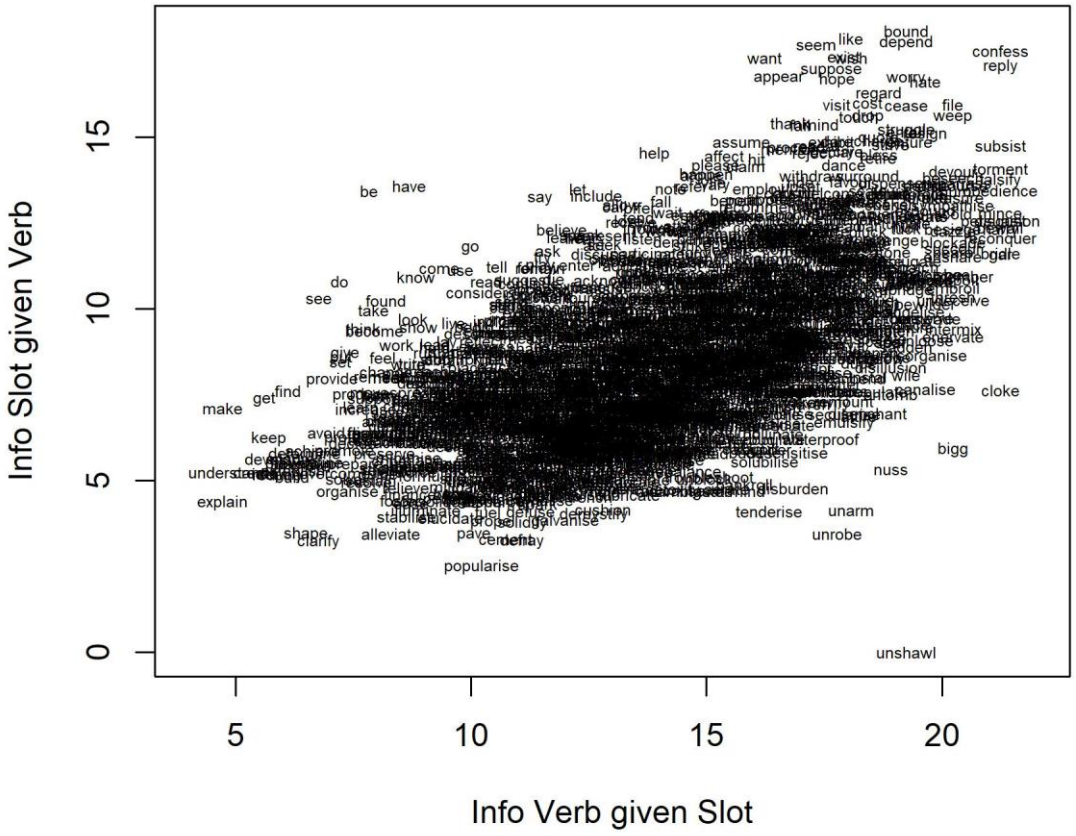

Figure 7.1. Information content values of individual verbs

The present study employs Generalized Additive Models (GAMs) to test the correlation between the length and information content. The response variable is binomial and represents the frequencies of the longer variant and the shorter variant for each verb. A log-transformed ratio of these frequencies represents the log-odds, or logit. The predictors were the information content variables computed for each slot filler as shown in Section 7.2. We expect a positive

correlation between the log-odds of the longer vs. shorter variant and the predictors. The models presented in this study were fitted with the help of the package *mgcv* (Wood 2006) in R, open-source statistical software (R Core Team 2017).

The main distinctive characteristic of Generalized Additive Models (GAMs) is that they allow for straightforward and convenient modelling of non-linear relationships between predictors and the response variable.[42] This is done with the help of smooth terms, which can be of various types and degrees of "wiggliness". In order not to oversmooth or undersmooth the data, various regression diagnostics were performed, based on the goodness-of-fit measures provided in the model summary, AIC (the measure that combines the goodness of fit with parsimony), visualization tools (different types of residual plots) and built-in tests, e.g. *gam.check()*.

GAMs allow one to test univariate or bivariate smooths. Similar to traditional regression modelling, the latter can be thought of as representing not only the effects of each individual predictor, but also their interaction. The choice for univariate or bivariate smooths was performed by comparing the deviance of the models fitted with Maximum Likelihood estimation. In all cases, the tensor product smooths turned out to provide a better fit than individual univariate smooths, since the former produced smaller deviance than the latter. For all other purposes, the default fast Restricted Maximum Likelihood estimation was used, which enables computationally efficient calculations.

In most cases, the models revealed some overdispersion due to excess zero proportions. To fix that, quasi-binomial models were fitted. More information about the basic concepts and modelling strategies of GAMs can be found in Wood (2006), Sóskythy (2017) and Wieling (2018).

Generalized Additive Models (GAMs) were fitted separately for each morphological form of *help*: *help*, *helped*, *helps* and *helping*, with and without the Helpee. The response variable was binomial. This means that it treated the frequencies of the bare and *to*-infinitive with each verb as a result of a series of Bernoulli trials with two possible disjoint outcomes, similar to heads or tails of a coin. The bare infinitive was treated as outcome '0', and the *to*-infinitive as '1'. This means that we try to predict the odds of the *to*-infinitive against the bare infinitive. Note that the order (what is 0 and what is 1) is only important for the interpretation

[42] Here, the logit, or log-odds of *to*-infinitives vs. bare infinitives.

of the coefficients. The information content variables modelled with the help of tensor product smooths were used as predictors. The reason for fitting the separate models instead of modelling different smooths as interactions in one generalized additive mixed model were of technical nature: there were problems with convergence of generalized additive mixed models in which the infinitives were treated as random effects. The large size of the data allowed fitting of several small models.

7.4.2. Results of quantitative analyses

The properties of the GAMs, including the estimated coefficients, *p*-values and goodness of fit statistics, are shown in Table 7.2. The settings of the models were the following: binomial (logit) family, tensor product smooths, thin plate regression splines as the basis, the wiggliness parameter gamma set at 1.4. All tensor smooths had highly significant estimates ($p < 0.0001$). The *** sign stands for $p < 0.0001$. Note that the intercepts represent the log-odds of the *to*-form, when controlling for the information content. The smaller the log-odds, the lower the chances of the *to*-infinitive and higher the chances of the bare form, and the other way round. Again, we see that the chances of the *to*-infinitive within the schemata with and without the Helpee negatively correlate with the total frequencies of the schemata, which meets the expectations based on the Principle of Communicative Efficiency.

Figure 7.2 shows the effects of the information content variables on the odds of the *to*-infinitive vs. bare infinitive for different forms of *help*. The yellow regions show the values which are associated with higher odds of the marked infinitive, whereas the red regions correspond to the values with the lower odds of the marked infinitive.[43] With the exception of *help*, the plots display higher chances of the *to*-infinitive in the areas associated with higher informativeness of the verb given the slot (the horizontal axis). Verbs with low information content, which are located on the left-hand side, tend to be used with the bare infinitive (for example, *explain*, *make*, *understand*, *create*, *get* and *keep*). On the right are mostly verbs that occur only once or twice in the construction. Most of them are low-frequency verbs, e.g. *bewail*, *reconquer* or *subsist*. They tend to occur with the *to*-infinitive.

[43] Here and in the other plots, the shading is based on the estimates (see the numbers on the contour lines), which represent the partial effect of the predictors. The effect is measured in log-odds ratios, which can be added to the intercepts in order to obtain the fitted log-odds (i.e. the logits) of the longer form vs. the shorter form for each verb.

Table 7.2. Important statistics and properties of the GAM models

Model	Intercept (log-odds of the *to*-form)	Effective degrees of the smooth term	*P* of smooth term	Adjusted R^2	Explained deviance	Scaling parameter
help + Inf	-0.4***	20.81	< 0.0001	0.184	18.3%	36.16
helped + Inf	1.42***	25.28	< 0.0001	0.111	22.7%	22.13
helps + Inf	1.91***	27.14	< 0.0001	0.379	42.7%	19.81
helping + Inf	4.41***	19.46	< 0.0001	0.387	45.5%	9.09
help + Helpee + Inf	-0.83***	17.55	< 0.0001	0.382	35.6%	28.28
helped + Helpee + Inf	0.095.	7.48	< 0.0001	0.222	23.2%	19.57
helps + Helpee + Inf	0.68***	11.36	< 0.0001	0.355	36.5%	15.28
helping + Helpee + Inf	0.92***	8.06	< 0.0001	0.25	30.6%	14.57

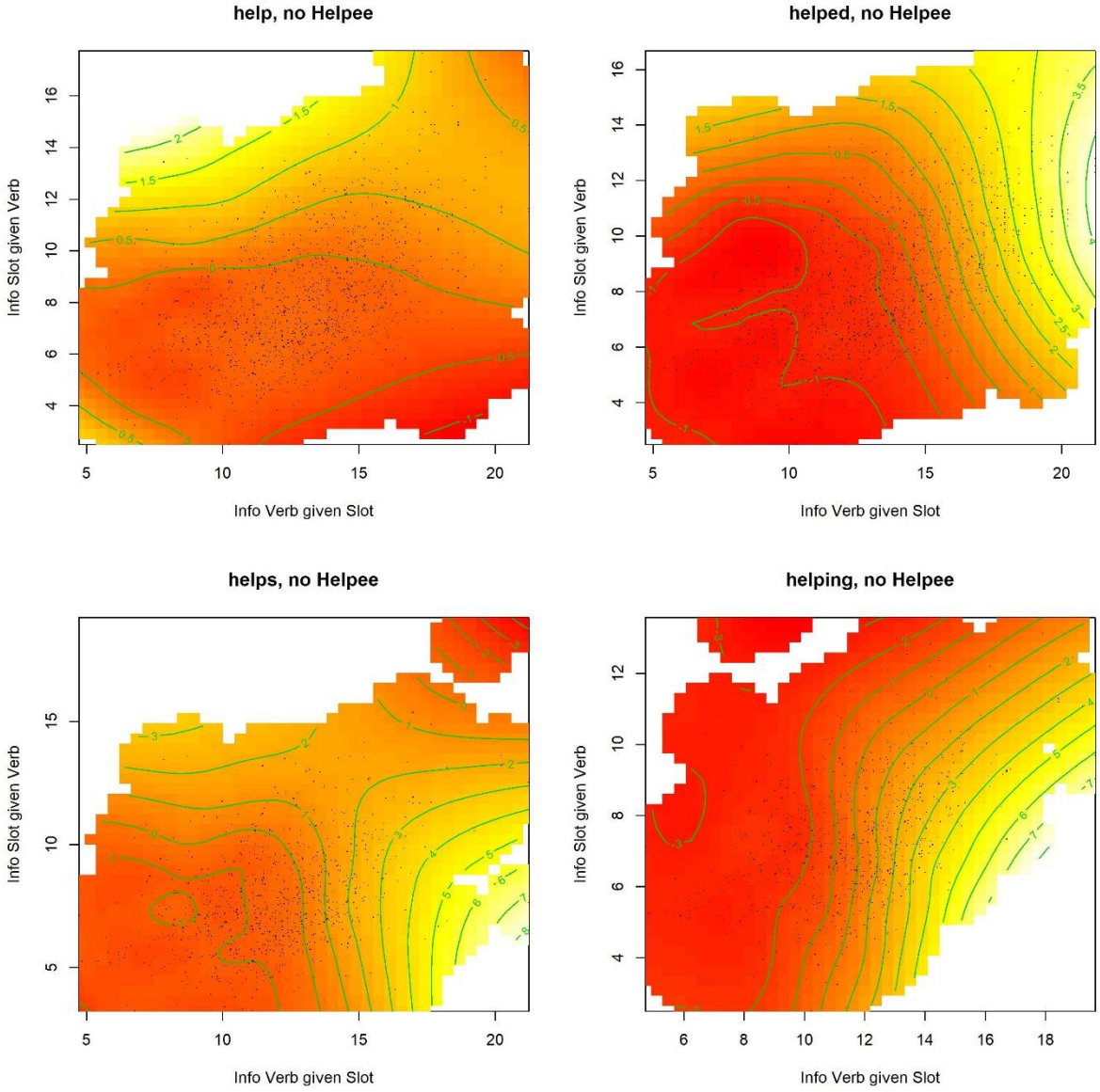

Figure 7.2. Effects of information content measures on the log-odds of the *to*-infinitive vs. bare infinitive immediately after different forms of *help* in the British dataset of Google Books

As for information content of the slot in the HELP-construction given a verb, which corresponds to the vertical axis, the relationship between amount of coding and information content is positive in the models with *help*, *helped* and *helps*. The model with *helping* does not suggest any clear relationship. The verbs with high information content of the slot given a verb are usually highly frequent verbs, such as *be*, *have*, *think* or *go*. Examples of verbs with low information content are mostly Latinate verbs, such as *alleviate*, *clarify*, *solidify* and *stabilize*,

which describe some desirable changes that some instrument or means helps to implement. See also Figure 7.1.

The explanatory power of the models, as one can see from the R^2-statistics and explained deviance, which can range from 0 to 1, or from 0% to 100%, was weak to moderate. This is not surprising, since informativeness effects tend to be quite subtle (e.g. Piantadosi et al. 2011).

We also observe that the effect of predictability of the constructional slot given a verb (the vertical dimension) relative to the effect of predictability of the verb given the constructional slot (the horizontal dimension) is the strongest for the form *help*, followed by *helped* and *helps*, and finally *helping*. In the latter, only the horizontal dimension is important. This variation is quite unexpected. However, it is possible to think of a possible explanation. If we compare the relative prominence of the vertical dimension of different forms, we can see that it corresponds to the total frequencies and the proportions of the bare infinitive in the constructional subschemata: the highest for *help*, and the lowest for *helping*. This observation will play an important role in the diachronic development of *help*, which is discussed in the next chapter. We will see that the effect of the vertical dimension increases with the degree of auxiliarization of *help*.

Now let us have a look at the models with the Helpee, which are displayed in Figure 7.3. Again, the relative prominence of the vertical dimension in comparison with the horizontal dimension, which is the highest for *help* and the lowest for *helps* and *helping*, closely corresponds to the ranking of the subschemata according to their total frequencies and the chances of the bare infinitive (see Table 7.1).

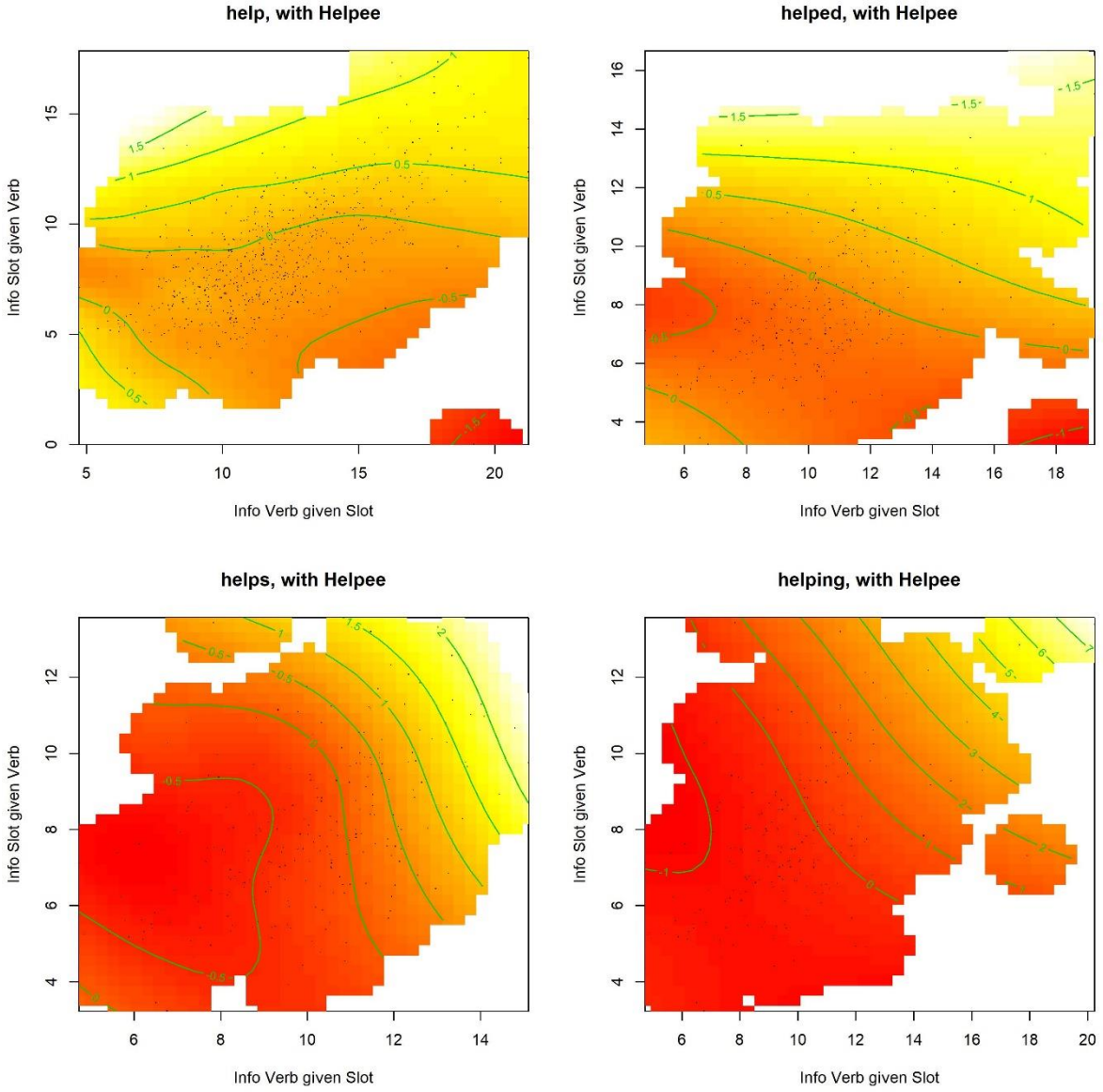

Figure 7.3. Effects of information content measures on the log-odds of the *to*-infinitive vs. bare infinitive after different forms of *help* and a pronominal Helpee in the British dataset of Google Books

7.5. Summary and discussion

As this case study demonstrates, at least one of the types of information content is correlated with the amount of coding in all forms of *help*. This effect is in the predicted direction: The higher the information content, the higher the chances of the *to*-infinitive. There is some interesting and unexpected variation, though. In the case of the form *help* without the Helpee,

one observes only a clear effect of information content of the slot given a filler, while the model with *helping* (also without the Helpee) tends to display the greatest prominence of the reverse measure, i.e. information content of a filler given the slot. As for the subschemata with the pronominal Helpee, only *helps* and *helping* display an effect of information content of a verb given the slot, in a combination with the other information content measure. Thus, the subschemata with individual inflectional forms display different behavioral properties, which is typical of so-called inflectional islands (Newman and Rice 2006).

This variation can be related to the level of auxiliarization of the individual forms of *help*, which seems to be the highest for *help* and the lowest for *helping* without the Helpee. The auxiliarization is determined by the total frequency of the subschemata and the proportion of the bare form. For the constructions with a pronominal Helpee, *help* + Helpee + (to) Infinitive is the most auxiliarized subschema, and the similar structures with *helps* and *helping* are the least auxiliarized ones. A diachronic investigation, which is presented in the next chapter, reveals an increasing prominence of information content of the slot given a filler as the verb *help* becomes more auxiliarized. The causes for such variation are discussed in the next chapter.

Moreover, the total frequency of the subschemata with different forms of *help* with and without the Helpee seems to correlate with the relative frequencies of the bare infinitive in these subschemata. This means that the more frequent subschemata tend to be shorter – yet another manifestation of efficiency. Although there are too few data points to test the significance of this correlation, this is an interesting finding.

Chapter 8. Slot-filler predictability and efficiency in diachrony: the case of *help* + (*to*) Infinitive

8.1. Aims of this chapter

This chapter continues to examine the alternation *help* + (*to*) Infinitive and traces its development in American and British English from 1901 to 2009. Its main aim is to investigate the diachronic development of slot-filler predictability effects, which were discussed in the previous chapter.

The previous studies show that the bare infinitive has been gradually replacing the *to*-infinitive after *help* over the last two centuries. A corpus study by Rohdenburg (2009: 318–319) shows that the infinitive marker *to* was dropped very rarely in British and American English with the authors born to the end of the 18th century, but there was a significant increase in the drop of the marker by the end of the 19th century. This tendency continued in American English also in the 20th century. British English speakers followed the suit with some delay. Similarly, McEnery and Xiao (2005) find that the bare infinitive is used more frequently in the British and American corpora representing 1991 than in the data collected in 1961, and that the American variant of *help* is more frequently used with the bare infinitive. These data support the idea about Americanization and auxiliarization of *help* (Mair 2002).

A synchronic study of this alternation, which was presented in the previous chapter, revealed that the variant with *to* in most subschemata of the construction has higher chances to occur with verbs that are infrequently used with *help*, in comparison to their total frequencies in the corpus. In some subschemata, we also observed the effect of predictability of a verb given the slot. In both cases, the marked form is preferred in contexts with low predictability of *help*, or, in the information-theoretic terms, high information content.

Importantly, we found that the relative prominence of the probability of the construction given a slot filler in comparison with the probability of the slot filler given the construction is greater in the subschemata with a higher total frequency in the corpus, and with a higher proportion of the bare form. These subschemata have different inflectional forms of *help*, i.e. *help/helped/helping/helps* + (Pronoun) + (*to*) Infinitive.

One may wonder if these effects change, as well, when the construction of interest undergoes change. This question has never been asked before. The main aim of this chapter is to find out how the grammatical changes in the construction with *help* over the last century are reflected in the strength and direction of the correlations between the slot-filler predictability and the use of the bare or *to*-infinitive.

To answer this question, more than 38 million instances of the construction were extracted from the American and British components of the Google Books Ngram dataset. The main advantage of this freely available data source is its huge size (more than 400 billion words from both countries for the period under investigation). The size is important when one needs to obtain reliable measures of information content for relatively infrequent constructions and when the distribution of constructional variants is heavily skewed. The greatest drawback of the dataset is the lack of full contexts where the construction with *help* is used. To mitigate this problem and to interpret the results of the analyses, one can find some illustrations in Google Books published in the corresponding time periods, as will be done below.

The rest of the chapter is organized as follows. Section 8.2 describes the data (Google Books ngrams). Section 8.3 reports the results of statistical analyses. Finally, Section 8.4 summarizes the findings and draws the conclusions.

8.2. Data and variables

8.2.1. The Google Books Ngram data set

As in the previous chapter, the data for this case study were extracted from the Google Books Ngram dataset.[44] These ngrams were collected from the books in English that were published in Great Britain and the USA. The main bulk of the data is rather formal and contains a large fraction of scientific literature, especially after 1900 (Pechenik et al. 2015). The main advantage of this data is its huge size, which makes it possible to estimate predictability effects (cf. Piantadosi et al. 2011). More information about the *n*-gram data is provided in Section 7.4.1 of Chapter 7.

[44] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html (last access 19.11.2018)

I used the ngrams from the British and American English datasets from 1901 to 2009, the latest year for which the ngrams are provided. The data before 1901 were not considered because the frequencies of *help* followed by the bare infinitive were extremely low. In addition, an examination of examples of *help* + bare Infinitive from the time-specific segments of Google Books revealed many spurious hits. Also, the OCR results may be less reliable for the data before 1800.[45]

The data were subdivided into five periods: 1901–1925, 1926–1950, 1951–1975, 1976–2000 and 2001–2009. The frequencies of the constructional variants and the total number of words (i.e. 1-grams) in the datasets are presented in Table 8.1. The normalized frequencies of the variants are also represented visually in Figure 8.1.[46]

Table 8.1. Total 1-gram counts and the frequencies of *help* + (*to*) Infinitive in Google Books Ngram datasets

Period	**Country**	**HELP with bare Infinitive**	**HELP with *to*-infinitive**	**Total frequency of HELP + (*to*) Infinitive, per million 1-grams**	**Total number of 1-grams**
1901-1925	Great Britain	15,934 (6.6%)	223835 (93.4%)	239,769 (23.5 per 1M)	10,220,101,900
	USA	259,578 (28.6%)	647,285 (71.4%)	906,863 (32.5 per 1M)	27,920,263,848
1926-1950	Great Britain	32,715 (12.5%)	228,066 (87.5%)	260,781 (33.3 per 1M)	7,837,760,300

[45] A notorious example is the high frequency of the word **uck* before the 19th century. This is explained by the fact that the elongated medial-s (ſ) used in pre-19th century English was often interpreted as an *f*. As a result, the word *suck* (and sometimes even *such*) was interpreted as the f-word. Although Google claims to have improved its OCR (see https://books.google.com/ngrams/info), the problem is still present, as one can see in the online Google Ngram Viewer.

[46] Note that low-frequency ngrams (occurring less than 40 times) were not included in the Google dataset.

	USA	515,927 (40.7%)	751,251 (59.3%)	1,267,178 (53.3 per 1M)	23,767,015,451
1951-1975	Great Britain	200,278 (25.1%)	598,224 (74.9%)	798,502 (49.3 per 1M)	16,192,705,836
	USA	2,004,627 (52%)	1,848,249 (48%)	3,852,876 (72.9 per 1M)	52,843,234,649
1976-2000	Great Britain	1,091,158 (45.5%)	1,309,257 (54.5%)	2,400,415 (77.0 per 1M)	31,189,357,201
	USA	9,198,832 (69.9%)	3,959,185 (30.1%)	13,158,017 (115.0 per 1M)	114,459,569,563
2001-2009	Great Britain	1,307,189 (52.9%)	1,163,864 (47.1%)	2,471,053 (96.5 per 1M)	25,594,987,968
	USA	9,704,071 (73.5%)	3,497,030 (26.5%)	13,201,101 (127.7 per 1M)	103,409,814,982

Figure 8.1. Normalized frequencies of the constructional variants in the American and British data in five periods

The numbers in the table demonstrate that the total frequency of *help* + (*to*) Infinitive increases with time in both varieties. This increase is mostly due to the growing popularity of the variant with the bare infinitive, as one can see in Figure 8.1. The most striking increase in the frequency of the bare infinitive is observed in both varieties in the second half of the 20th century. In addition, the American data display a higher relative frequency of the bare infinitive and a greater overall frequency of the construction in all periods. These results match the previous accounts (cf. Section 8.1).

As shown in Appendix 3, the form *help* is also the most frequent one in both countries in all periods, with the exception of the first half of the 20th century in Great Britain, when *helped* was more frequent in the contexts without the Helpee, e.g. *helped to create*. The form *help* is also followed by the bare infinitive more frequently than the other forms. The least frequent form is *helping*. When followed immediately by the infinitive, this form is disproportionately frequently used with the *to*-form, e.g. *helping to understand*. In the cases when the Helpee is present, the form *helps* is particularly often followed by a *to*-infinitive, e.g. *helps us to understand*. In accordance with these findings and the results reported in the previous chapter, we can say that the forms of *help* can be regarded as inflectional islands (Newman and Rice 2006), since different inflectional forms exhibit different syntactic behaviour. These differences will be relevant for the interpretation of the findings of this chapter.

One should also mention here that the choice of the infinitival form after *help* depends on several contextual factors, such as the linguistic distance between *help* and the infinitive, the presence of the Helpee, the form of *help*, *horror aequi* and the register, as was discussed in Chapter 7. In the present study, most of these variables are controlled for because the effects of the information content variables are tested separately in the subsets of data with the same form of the verb *help*, the same value for the Helpee (absent or present) and the same linguistic distance, which represents the number of words between *help* and the infinitive (with or without *to*). One can also expect the register – predominantly formal writing – to be quite homogeneous. One important factor that cannot be taken into account in this study is *horror aequi*, which represents the tendency to avoid the *to*-infinitive when the verb *help* is already preceded by the particle (e.g. *She wants* ***to help support*** *the local community*) (Rohdenburg 2003). It should be mentioned, however, that the synchronic analyses presented in other work (Levshina 2018)

based on manually annotated and fully contextualized data revealed a significant effect of the information-theoretic variables when *horror aequi* and the other factors were taken into account.

8.2.2. Measures of information content: distribution and linguistic interpretation

Using the frequencies of *help* with the specific infinitives, the total frequencies of individual verbs in the corpus, and the total frequencies of the construction in different periods, I computed the measures of information content, which were defined by the formulas in Chapter 7 (Section 7.3). Figure 8.2 shows the distribution of both information-theoretic measures in the USA in 2001–2009 as an illustration. The horizontal axis represents the information content of the infinitive given the slot in the construction with *help* (variable *InfoVerb*), whereas the vertical axis displays information content of the constructional slot given the verb (variable *InfoSlot*). Due to the high type frequency of verbs in the whole dataset, the plot is based on a random sample of 20,000 instances of the construction.[47]

The verbs like *get*, *make*, *create* and *understand*, which are found in the bottom left corner, have low values on both dimensions. They occur very frequently in the construction in all periods and in both countries. An example from Google Books published in the same time period is given in (1):

(1) *I use this word symbolically to* **help you understand** *why keeping things inside can harm you emotionally.* (URL https://books.google.de/books?isbn=1440626871)

The high-frequency verbs in the top left area like *be*, *do*, *see* and *think* have high *InfoSlot* and low *InfoVerb*, due to their relatively high frequency in the construction, but at the same time very high frequency elsewhere. An examination of the available examples from Google Books reveals that these verbs often co-occur with *help* in religious and pedagogical texts. Consider as an example a popular paraphrase of the famous saying by St. Augustine, *O Lord, help me to be pure, but not yet*. An example from Google Books is given below:

[47] Two methods of computing information content were used in this study: 1) based on the total frequencies of the construction in all forms, and 2) based on the frequencies of particular subschemata (e.g. *helping* + pronominal Helpee + (*to*) Infinitive). The results lead to the same conclusions. For the reasons of space, only the results based on the first method are reported.

(2) *She found an independent midwife who **helped her to have** a home birth in water.* (URL https://books.google.de/books?isbn=1134193033)

The verbs in the top right corner are generally frequent, but occur in the construction very rarely, e.g. *like* and *want*. This is the area where the parsing seems to be the least reliable. Many of these verbs are highly ambiguous (e.g. *last*, *like* and *please*), and one can imagine that spurious hits are possible, especially if the punctuation is not perfectly correct, as in the example below:

(3) *Trey looked over at Donald again as if to say, **help me please**.* (URL https://books.google.de/books?isbn=1465322213)

Finally, the bottom right area contains verbs that are generally rare and do not occur with *help* frequently, e.g. *vulgarize* and *reacquaint*. One finds many scientific Latinate terms in this area. Consider an example:

(4) *Reviewing the following diagrams will perhaps **help to visualize** this concept of transfer of action.* (URL https://books.google.de/books?isbn=0889771553)

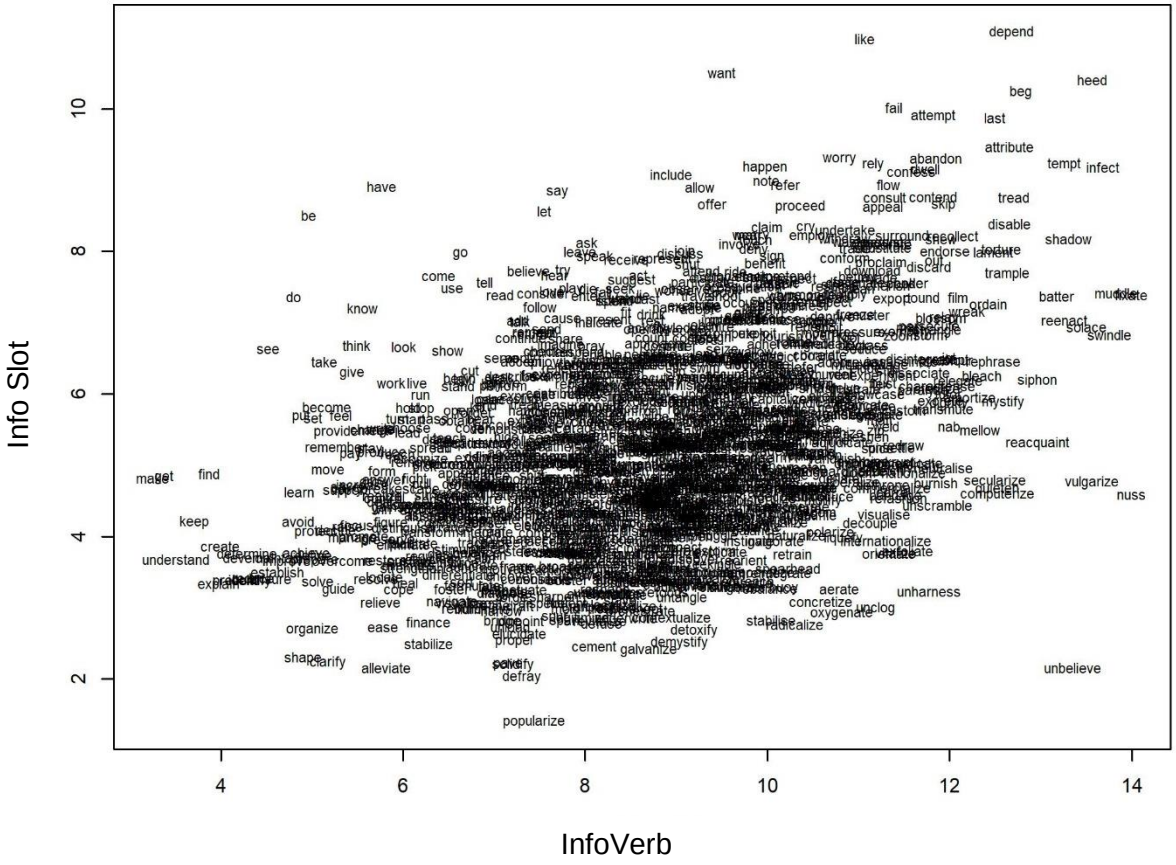

Figure 8.2. The distribution of the information-theoretic scores in the US data, 2001–2009

Overall, there is substantial similarity between the scores of individual verbs in different time periods, although the magnitudes of the scores change somewhat with time. In particular, the scores of information content of the slot decrease due to the increasing relative frequency of the construction.

8.3. Results of quantitative analyses

8.3.1. Partial correlations: method and design

This section reports the results of correlation analyses that investigate the relationships between the proportions of the bare infinitive for each individual verb and its information content measures. The method of partial correlations provides a convenient way of measuring the strength of relationships between two variables while controlling for the effect of a third variable (or several other variables). Partial correlations have been applied in some information-theoretic corpus-based studies, for example, for measuring the effects of frequency and average information content on word length (Piantadosi et al. 2011).

Here, the data units are specific verbs that fill in the infinitival slot after *help*. For every verb, I measured its information content given the slot, the information content of the slot given the verb and the proportion of the bare forms in the total uses of this verb with the *help*-construction. Next, the non-parametric rank-order partial correlation coefficients (Spearman's *rho*) were computed. for each of the five periods in both countries and for each of the following subschemata:

- *help* + no Helpee + (*to*) Infinitive,
- *helped* + no Helpee + (*to*) Infinitive,
- *helping* + no Helpee + (*to*) Infinitive,
- *helps* + no Helpee + (*to*) Infinitive,
- *help* + pronominal Helpee + (*to*) Infinitive,
- *helped* + pronominal Helpee + (*to*) Infinitive,
- *helping* + pronominal Helpee + (*to*) Infinitive,
- *helps* + pronominal Helpee + (*to*) Infinitive.

This was done because the correlations vary across the subschemata, as will be demonstrated below.

The analyses were performed with the help of an add-on package *ppcor* (Kim 2015) in R (R Core Team 2017).

8.3.2. Effect of information content of a verb given the slot

Figure 8.3 displays the correlation coefficients of the relationship between the proportion of the bare infinitive for a specific verb and its information content given the slot, while controlling for information content of the slot given the verb. The points represent the beginning of each period, e.g. the point at 1901 shows the correlation observed in the period from 1901 to 1925. A positive value means an increase in the proportion of the bare infinitive, as the information content increases. A negative value represents increasing chances of the *to*-infinitive. According to the Hypothesis of Slot-Filler Predictability and Formal Length, one can expect the correlations to be negative: the greater information content, the smaller the proportion of the bare infinitives.

The left panel displays the results for the cases when the infinitive immediately follows *help*, without a Helpee. All forms have significant negative correlations in the construction without the Helpee, with the exception of the form *help*. The latter develops weak positive correlations over time, which are also statistically significant ($p < 0.05$). The form *helping* has the strongest negative correlations in all periods. Overall, one can see greater divergence between the forms in the later periods than in the beginning.

The right panel shows the correlations in the presence of a pronominal Helpee. Here, all correlations stay or become significantly positive over time. We can also see that most of the subschemata gradually converge with time. The form *help* has the highest correlations, and the form *helps* the lowest ones (except for the last two periods, where its behaviour is similar to that of the form *helping*).

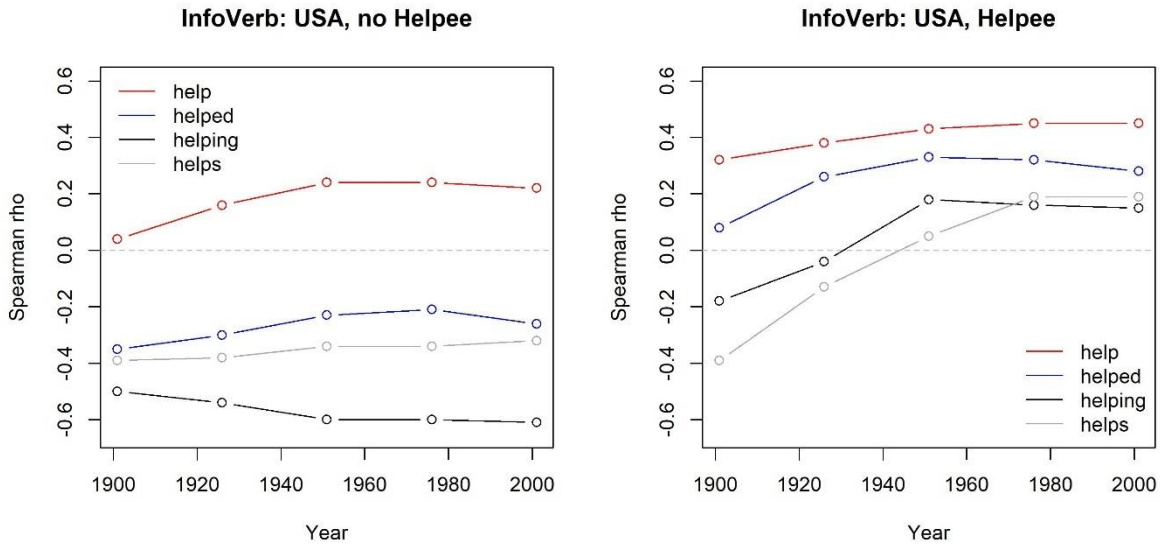

Figure 8.3. The USA data: correlations between the proportions of the bare infinitive and information content of a verb given the slot

The results for the British data are displayed in Figure 8.4. Again, the left-hand plot displays the correlations found for the subschemata without the Helpee, whereas the right-hand plot displays the correlations for the subschemata with a pronominal Helpee. Here, most subschemata display significant negative correlations, as predicted. An exception is again the verb form *help*, which loses its significant correlation. We also observe that the forms of *help* diverge over time.

To summarize, the relationships between the proportion of bare infinitives and information content of a verb given the slot exhibit a lot of variation. In the British data, the predicted negative correlation is observed in all subschemata in the beginning, but then it becomes weaker or disappears. In the subschemata with the form *help*, this correlation becomes the weakest earlier than in the other subschemata. As for the American data, the form *help* even displays a positive correlation. The other forms follow it in the subschemata with the pronominal Helpee.

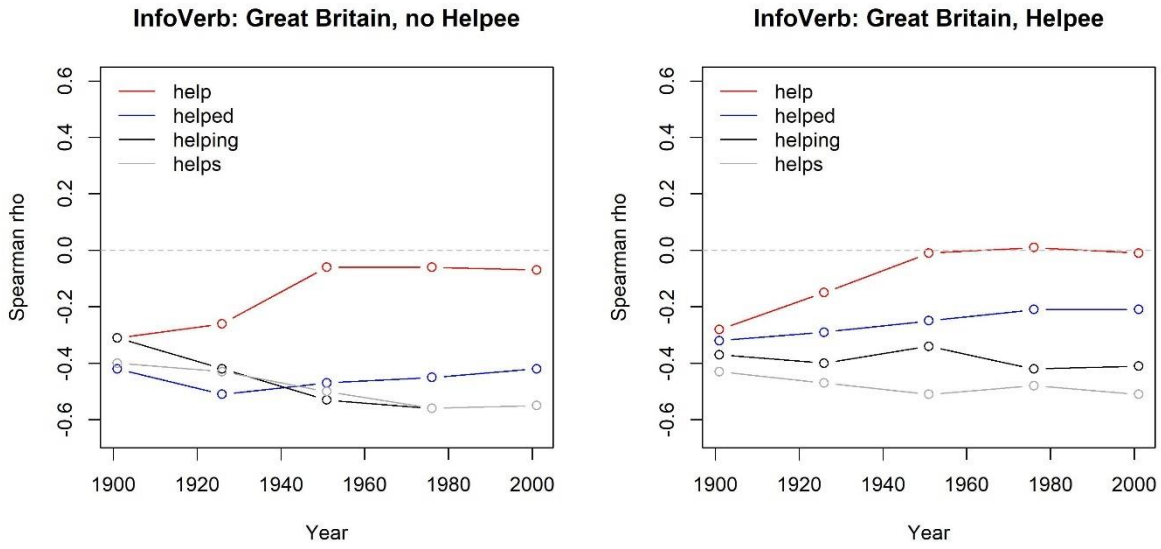

Figure 8.4. The British data: correlations between the proportions of the bare infinitive and information content of a verb given the slot

8.3.3. Effect of information content of the slot given a verb

This subsection discusses the partial correlations between the proportions of the bare infinitive and information content of the slot given a verb, while controlling for information content of the verb given the slot. Figure 8.5 displays the Spearman's correlation coefficients in the American data. In the data without the Helpee (see the left panel), the effects gradually become significantly negative, with the exception of the form *helping*, which remains slightly above zero ($p > 0.05$). The effect in the constructions with the Helpee (see the right panel) either remains or becomes weakly negative with time. The form *helps* now has the weakest negative effects, whereas the form *help* displays the strongest correlations.

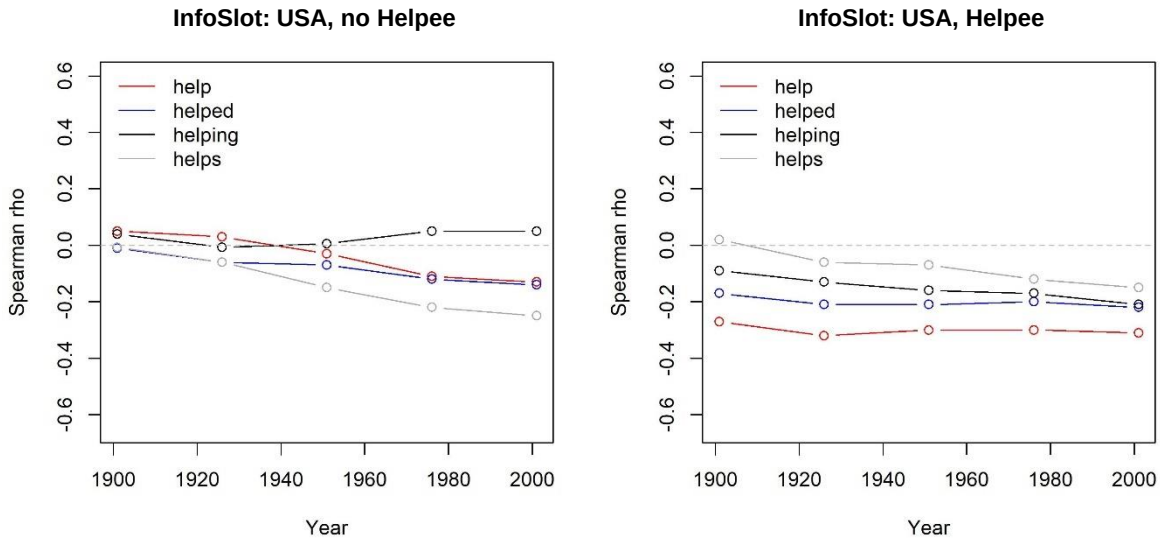

Figure 8.5. The USA data: correlations between the proportions of the bare infinitive and information content of the slot given a verb

In the British data (see Figure 8.6), we observe very similar changes. Note that the positive correlations of *helps* in the subschema with the Helpee in the right panel cease to be statistically significant with time.

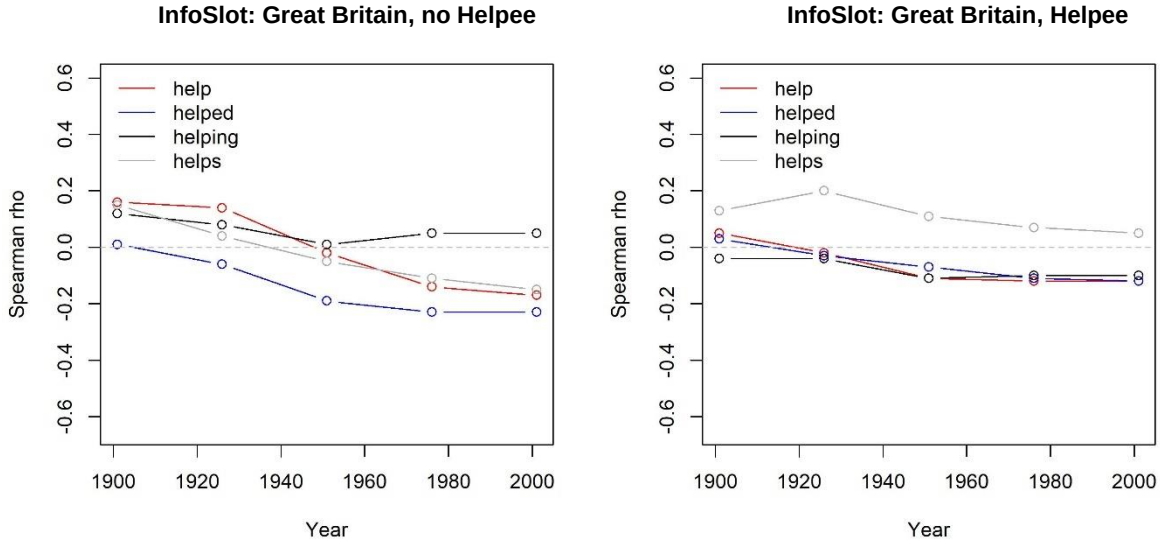

Figure 8.6. The British data: correlations between the proportions of the bare infinitive and information content of the slot given a verb

To summarize, the correlations between the proportions of bare infinitives and information content of the slot given a verb tend to become or remain negative. In the US subschemata with the Helpee these correlations are slightly stronger than in the corresponding British constructions. These differences and the ones reported above suggest that quite a few of the British subschemata lag behind their American counterparts, as far as the effect of the information content variables is concerned.

8.4. Summary and discussion

The present chapter has used the big data from the Google Books Ngram dataset to investigate the historical changes in the construction with *help* in texts from the USA and Great Britain in the 20th and the beginning of the 21st century. The frequencies of the constructional variants support the previous observations about the increasing frequency of *help* + (*to*) Infinitive, which happened mostly due to the gradual spread of the constructional variant with the bare infinitive. We have also observed that the British variant follows the path of the American construction in that regard, which supports Mair's (2002) account of auxiliarization and Americanization of *help*.

On the methodological side, one should mention that the frequencies and tendencies detected in the present study are very similar to the ones reported in previous comparative and diachronic studies (e.g. Mair 2002; McEnery and Xiao 2005; Rohdenburg 2009; Lohmann 2011) based on fully contextualized data and reputable corpora. This fact lends credibility to the Google Books Ngram data as a data source for grammatical research.

The main research question of this case study was if the use of the bare and *to*-infinitive can be predicted from information content, based on the probabilistic relationships between the construction and the verbs that fill in the infinitival slot, and whether these effects change with time in the data from both countries. The results of the correlational analyses show clearly that they do. As the bare infinitive becomes more and more frequent, the effect associated with information content of the slot given a verb becomes more prominent in both countries across most subschemata defined by the form of *help* and the presence or absence of the Helpee. This

is particularly obvious when the variant with the *to*-infinitive becomes restricted to the verbs that are highly frequent (e.g. *be*, *have*, *go*, *know* and *see*).

As for information content of a verb given the slot, all subschemata in the early British data and most subschemata in the early American data display a negative correlation. However, this effect disappears gradually in British English, especially with the base form *help* (in both countries). In the American data, the correlation even becomes positive. Such reversals happen sometimes when one frequency measure counterbalances the effect of another measure in the presence of a correlation between them (cf. Bell et al. 2003). More research is needed in order to explain this finding.

How can these changes be explained? Probably, they have to do with the gradual auxiliarization of *help* and the increasing frequency of *help* + (*to*) Infinitive. As the construction becomes more entrenched and conventionalized, one can increasingly rely on the probability of the slot given a verb, or $\mathbb{P}$ (Inf-slot$_{\text{HELP}}$|Verb), when interpreting the function of the verb. At the same time, when the construction is infrequent, this information may be less accessible. The probability on average grows with time, since the frequency of the construction increases. Apparently, there should be some frequency threshold for the speakers to become sensitive to this probabilistic information.

These conclusions are supported by the differences in the effects of the information-theoretic variables observed in different subschemata of the construction. These differences can be explained by the different degrees of auxiliarization of the corresponding forms of *help*, as one can judge from their total frequencies. For instance, the form *help* is the leader in 'forgetting' the effect of information content of the verb given the slot. It is also the form with the highest total frequencies in most periods, which probably explains why it also has the highest proportions of the bare infinitives. In contrast, the form *helping*, when followed immediately by an infinitive, exhibits the lowest relative frequency of the bare infinitives and overall a low frequency. It is also the subschema with the weakest effect of information content of the slot given a verb. In that respect, it is similar to the form *helps* followed by a pronominal Helpee, which also has the lowest proportions of bare infinitives in comparison with *to*-infinitives among the subschemata with a pronominal Helpee in both countries (see Appendix 3). These observations support the idea that a higher total probability of a subschema means a greater impact of the conditional probability of the slot given a verb on the formal variation within that subschema.

Generally speaking, the specific verb is the linguistic cue, and its function as a part of the infinitival complement of *help* is the interpretation that the hearer should arrive at. High cue validity allows one to reduce the form, whereas low cue validity increases the chances of the longer form. Apparently, this type of information becomes increasingly important when a constructional element becomes grammaticalized and the entire expression becomes more entrenched and conventional. This fits the results of the previous case study of another grammatical phenomenon, namely that of differential case marking. Based on these results, we can formulate a more general hypothesis. As a construction becomes conventionalized, the role of cue validity of this type increases. This hypothesis should be tested in future studies.

Chapter 9. Slot-filler predictability and efficiency: Locative (*at*) *home*:

9.1. Aims of this chapter

The present chapter continues to discuss slot-filler predictability and its effects on the use of near-synonymous constructions of different length.[48] It investigates the use of locative adverbials *home* and *at home* in American English. When the meaning is directional, e.g. *go/return/bring (someone) home* or *a long way home*, no preposition is used. The forms *home* and *at home* can only be interchangeable when the meaning is locative, as in (1):

(1) a. *Dads who stay **at home*** (COCA, Magazines)

b. *Stories abound of men **staying home** to look after newborns* (COCA, Magazines)

For brevity, this variation will be called the domative alternation. To the best of my knowledge, it has not been studied systematically by linguists. One of the few mentions of this variation can be found in Huddleston and Pullum (2002: 683). They claim that *home* marks location only as a subject-oriented complement, as in *Are you home? We stayed home*, but not in other contexts, e.g. **I kept my computer home* or **Home, the children were playing cricket*.

At the same time, the use of these expressions attracts language learners' attention, judging from numerous discussions on Internet fora.[49] One of the aims of the present study is to fill this gap and investigate the linguistic factors that influence the use of *(at) home*. I will

[48] A part of the work presented here (Section 9.2) has been published in a slightly different form in Levshina, Natalia. 2018. Anybody (at) home? Communicative efficiency knocking on the Construction Grammar door. *Yearbook of the German Cognitive Linguistics Association 6*: 71–90.

[49] E.g. https://english.stackexchange.com/questions/21286/im-home-or-im-at-home, https://www.quora.com/What-is-the-difference-between-I-am-home-and-I-am-at-home , https://www.usingenglish.com/forum/threads/68883-Correct-Usage-home-or-at-home and numerous others (last access 19.11.2018).

focus on American English, where this variation seems to be more common, as one can conclude from language users' intuitions and experts' comments.[50] In this study, I will test some of the factors that are mentioned in these discussions, such as figurative vs. literal meaning and the semantics of arrival. The data, which will be described below, come from the spoken component of the *Corpus of Contemporary American English* (COCA) (Davies 2008–). I use conditional inference trees (Hothorn et al. 2006) to test the impact of the above-mentioned factors, which turn out to be highly relevant for the domative alternation.

The main aim, however, is to test if slot-filler predictability, which was introduced in Chapter 7, can predict the use of the short and long forms. Zooming in on the uses of *(at) home* after intransitive predicates in the contexts which allow for sufficient variability in the use of the two variants, I check if the predictability of a predicate given *(at) home* and the other way round allows us to predict the presence or absence of *at*. Following the approach employed in Chapters 7 and 8, predictability is understood and measured as the conditional probability of a verbal predicate given the domative adjunct with or without *at*, and as the probability of the adjunct given the verb. The prediction, based on the Principle of Communicative Efficiency, is that the shorter form *home* is preferred when either the verb or the adjunct is more predictable.

This chapter begins with a multifactorial analysis in Section 9.2 in order to exclude the contexts where there is no or little variation between the forms. Section 9.3 zooms in on the variable contexts and tests the effects of slot-filler predictability. The results are summarized and discussed in Section 9.4.

9.2. Multifactorial analysis of *(at) home*

9.2.1. Data and variables

First, I extracted all instances of the wordform *home* from the spoken component of COCA. I also extracted the contexts, which included 25 words on the left and 25 words on the right from the target form, as well as the information about the broadcasting channel and TV or radio program where each observation occurs. Next, the instances were inspected manually, and all spurious hits were discarded. In addition to the contexts with verbs of self- and caused motion

[50] E.g. https://forum.wordreference.com/threads/home.883256/ and http://www.bbc.co.uk/worldservice/learningenglish/grammar/learnit/learnitv240.shtml (last access 19.11.2018).

(e.g. *go home, drive someone home*), I removed the instances with such verbs as *expect, get, want, allow, call, invite* and *welcome* (*someone home*), where the directional semantics was strong. I also excluded idiomatic expressions, such as *drive/hammer a point/message home, home and dry* and *home free.* In addition, names of films, books, songs and programmes (e.g. *Home Alone* and *Home on the Range*) were removed, as well as the lexicalized uses of *stay at home,* as in *a stay at home mom/dad*.

After that, I still had over 10,000 instances of locative *(at) home*. From this dataset, I took a random sample of 1,000 instances and coded them for several variables, which are described below. The longer variant *at home* in this random sample is almost twice as frequent as the variant with zero marking *home* (more exactly, 652 occurrences of *at home* and 348 occurrences of *home*).

The contextual variables, which are presented in Table 9.1, are the following.

Table 9.1. Overview of the contextual variables

Variable	**Label**	**Values**
1. Locative adverbial or particle before *(at) home*	*PlaceAdv*	*No*, *back*, *here*, *Other*
2. Literal or figurative meaning	*Figurative*	*Literal*, *Metaph* (metaphorical), Gener (generalized)
3. Semantics of arrival	*Arrival*	*Yes*, *No*
4. Syntactic function	*SyntFun*	*Pred_Intr* (intransitive predicate), *Pred_Tr* (transitive predicate), *Sent* (sentence adjunct), *Exist* (existential construction), *Attr* (NP attribute), *Ellipsis* (elliptical structure)
5. Broadcasting channel	*Channel*	*ABC*, *CNN*, *CBS*, etc.

1. *Locative adverbial or particle before (at) home*. This variable shows whether the domative adjunct was preceded by another locative adverbial or particle. The values were "No", "back", "here" and "Other". The adverbs *back* and *here* were both frequent (95 and 51 occurrences in the sample, respectively), and therefore were taken into account individually. Although they

displayed strong preferences for one or the other variant (i.e. *back home* and *here at home*), there were several exceptions. Compare, for example, (2a) and (2b):

(2) a. *I'd rather be poor back **home** than here…* (NPR Tell more)

b. *But back **at home**, the party was over*. (CBS 48 Hours)

2. *Literal or figurative meaning of (at) home.* The category "Literal" means that *(at) home* indicates being in a place where someone actually lives. An example is provided in (3).

(3) *I also have geraniums **at home***. (ABC Primetime)

The second sense is metaphorical. It is used when *(at) home* expresses one's feeling of being comfortable and at ease in a particular situation:

(4) a. *And he's probably more comfortable and **at home** with his stage makeup every day.* (Ind Geraldo)

b. *…so that if they know if anything goes wrong, they're going to be able to survive, and it's like being **home***. (CNN Talkback)

Although the *at*-variant is usually preferred in these situations, as in (4a), there are a few cases when the bare variant is used in a figurative meaning, as in (4b). The third type is a semantic generalization, when *(at) home* is used to refer to the city or country where one lives. An example of this type is (5a), which discusses a politician, who faces problems in his own country:

(5) a. *But for all his achievements on the international scene, the problems he faces **at home** seem insurmountable.* (ABC Nightline)

b. *…Republicans I talked to, lawmakers, several of them who are **home** with their constituents, home with potential voters…* (CNN Zahn)

In the generalized contexts like this, the *at*-variant is used more frequently, as in (5a) than the bare variant, as in (5b).

3. *Semantics of arrival*. This variable stands for the presence or absence of contextual clues that suggest that the person or object staying at home has recently arrived there. Compare (6a), where the speaker signals his or her arrival home, with (6b), where this information is not available or relevant:

(6) a. *Darling, I'm **home**!*

b. *Is anybody **home**?*

To code this variable, I relied on different contextual clues, including certain temporal expressions (e.g. *soon*, *by now*, *for Christmas*, *finally*, *after a journey to X*) and the previous location (from place X). An example is provided in (7).

(7) *I was **home** from college for the summer, and I said I'd do it.* (NPR Weekend)

One can expect that arrival should be more often implied in the contexts with *home* than in those with *at home*. The reason is the closeness of such uses to directional semantics, which is expressed by bare *home*.

4. *Syntactic function of (at) home*. This variable is inspired by Huddleston and Pullum's (2002) observation that the unmarked locative *home* can be used only with subject-oriented predicates (cf. Section 9.1). Coding the orientation of the predicate as subject- or object-oriented turned out to be very difficult in practice. As a proxy, I decided to use transitivity of the predicate because intransitive predicates are usually subject-oriented. There were also many other functions. The full list is as follows:

- adjunct of an intransitive predicate, i.e. one without a direct object: *I'm home*;

- adjunct of a transitive predicate, i.e. one with a direct object: *I build furniture at home*;
- sentence adjunct: *At home, I drink only tea*;
- attribute that post-modifies a nominal phrase: *Their stores at home are even emptier than here*;
- adverbial modifier in the existential construction *there + BE*: *There is too much stress at home*;
- part of an elliptic structure: *Finally, at home!*

If Huddleston and Pullum's (2002) claim is correct, we can expect *home* to be used predominantly as an adjunct of intransitive predicates.

5. *Channel*, which stands for the broadcasting channel that the observation comes from, e.g. ABC, CNN and Fox Broadcasting Company. This was done in order to take into account possible variation across the media, similar to random effects in regression analysis.

9.2.2. Conditional inference tree

This section tests the variables which were introduced in the previous section. I use a non-parametric method of conditional inference trees. This is a classification and regression method which has been used in sociolinguistics (e.g. Tagliamonte and Baayen 2012) and variational probabilistic grammar (e.g. Szmrecsanyi et al. 2016). One of the main advantages of this method is that it helps to model complex interactions between predictors in a very intuitive and easily interpretable way. Conditional inference trees are grown based on *binary recursive partitioning*. The algorithm starts with the entire dataset and tries to find the predictor that is the most strongly associated with the response. Then the algorithm makes a binary split in that variable, such that the strength of association or correlation between the predictor and the response is maximized. After that, the procedure is repeated again as long as certain criteria are met. Most importantly, a split can be made when a certain level of statistical significance is achieved, which serves as the minimum criterion for splitting.

To fit a conditional inference tree model, I used the package *party* (Hothorn et al. 2006).[51] The default settings were used (i.e. the minimum criterion for splitting was 0.95, which corresponds to the maximal *p*-value 0.05, the minimum number of observations in a node was seven, the minimum number of observations for a node to be considered for further splitting was 20). The resulting tree is shown in Figure 9.1.

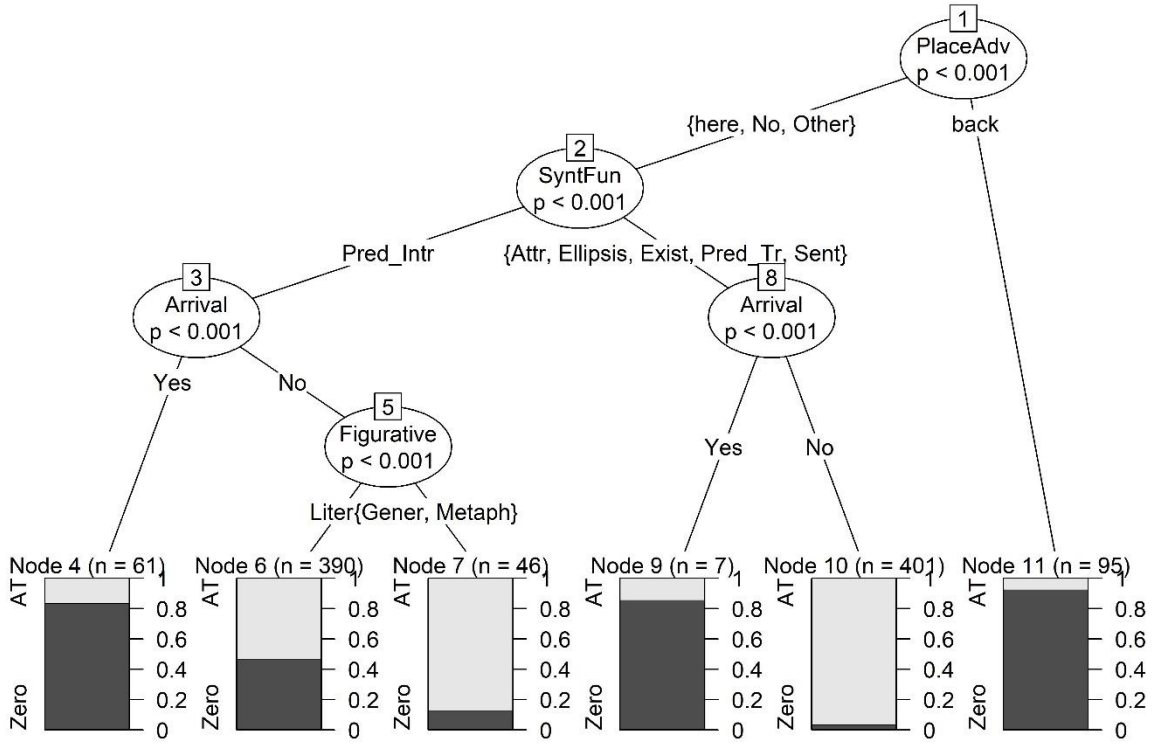

Figure 9.1. Conditional inference tree of *(at) home*

The first split at the very top (see Node 1) is made in the variable *PlaceAdv*, which stands for the presence of a location adverb modifying *(at) home*. If we have *back* before the domative adjunct, no other splits are made. The corresponding final Node 11 on the extreme right contains all observations with the adverb *back* + *(at) home* in the dataset. The proportions of the bare and *at*-variants are shown in the bar plot. As one can see, the variant with zero marking is predominant. This means that the chances of the bare variant after *back* are very high. Consider an example:

[51] I have also tried a more recent package *partykit*, which is claimed to have more up-to-date algorithms. The results were identical.

(8) *See, I tend bar back* ***home*** *in Indiana.* (NPR Fresh Air)

In the absence of *back*, other variables play a role. Let us examine them. First, consider Node 2, where the split is made in the syntactic function of *(at) home*. It separates adjuncts of intransitive predicates (*Pred_Intr*), e.g. *be (at) home*, which form the left branch, from the other functions, which form the right branch and are then split in the variable related to arrival (Node 8). If the semantics of arrival is prominent, the chances of the bare variant are very high (see Node 9). An example is given in (9).

(9) *We can't wait to have your kids* ***home*** *safe as well, and for that wedding, just incredible.* (ABC GMA)

If it is not prominent, *at home* is used almost exclusively (Node 10), for example:

(10) *And you* ***at home****, go have fun with eggs, and happy Easter, everybody.* (ABC GMA)

Let us now go back to the contexts with intransitive predicates. Here, again, a split is made in *Arrival* (Node 3). As in the previous case, the semantics of arrival is associated with a very high proportion of the bare variant (Node 4). Consider an example:

(11) *But he promised he would be* ***home*** *in a year and he never came home.* (NPR Morning)

In all other remaining cases, the distinction between figurative and literal meanings plays a role (Node 5). If the semantics is figurative (generalized or metaphoric), the *at*-variant is almost exclusively used (Node 7), as in the following example:

(12) *Well, you are a party animal. (…) You were right* ***at home***. (NBC Today)

If *(at) home* is used in the literal sense, the proportions of the bare and *at*-variants are almost equal (Node 6). Examples are given in (13).

(13) a. *I'm lucky because I'm a writer and I work* ***at home***. (NPR Talk of the Nation)

b. *…if you were sitting* ***home*** *on a sunny day while a lot of other boys were playing baseball…* (NPR Fresh Air)

There are 390 such observations.

The classification accuracy of this tree is 77.9%. This number stands for the proportion of observations where the predictions of the model and the actual variants used in the contexts coincide.[52] If an observation is in a node with predominantly *home*, e.g. Node 11, then the model will predict the bare variant for this observation, as well. In an observation is in a final node with predominantly *at home*, then all observations in that final node will obtain the *at*-variant. The accuracy is higher than the baseline of 65.2%, which represents the accuracy that can be achieved if one always predicts the more frequent variant *at home*.

Another statistic, which is based on predicted probabilities, is the concordance index *C*. Predicted probabilities for an individual observation are computed on the basis of the proportions of each variant in a given final node. In our case, $C = 86.3$ (with 0.5 for a useless model and 1 as perfect discrimination). This number shows that the model discriminates well between the variants.

9.3. Zooming in on the variable contexts

This section zooms in on the contexts that exhibit most variation, according to the multifactorial analysis presented above. Recall that these observations have no adverb *back*, contain only

[52] Based on out-of-bag prediction.

intransitive predicates, exhibit no semantics of arrival and occur only in the literal meaning. The variants *home* and *at home* are almost equally distributed in these contexts. I use my original larger dataset from the spoken component of COCA (see Section 9.2.1) in order to extract such contexts and test the effect of predictability on the choice between the variants without the danger that the effects are confounded by the other contextual variables.

In order to avoid that, set expressions (e.g. *charity begins at home*, *be home free*, *romp home*) were excluded, as well as the metaphoric expression *feel at home*. The contexts with the adverb *back* before the domative adjunct were excluded, as well. Next, all intransitive verbs (i.e. verbs without a direct object in the sentence) followed immediately by (*at*) *home* were selected. After manual cleaning, this resulted in 4,032 occurrences of the domative alternation with 71 different verbs. The bare variant was used 2,623 times (65%), while the prepositional form occurred 1,409 times (35%).

The frequencies of each verb with *(at) home* and the total frequencies of the verb in the spoken subcorpus were obtained. Based on that information, the information content scores were computed for each verb, using the same approach as the one employed in the previous chapters.

The data are distributed in a rather straightforward way. Most verbs are followed by *at home* only. Only seven verbs are followed both by the prepositional and bare variants in the sample: *be*, *stay*, *belong*, *sit*, *remain*, *wait* and *live*. Six of those verbs (with the exception of *belong*, which occurs only once with *home* and once with *at home*) are among the top twelve verbs that have the highest frequency with *(at) home* and which are displayed in Table 9.2. In particular, the verbs *be*, *stay* and *sit* have the highest proportions of the bare variant. Notably, they are also the top most frequent verbs that occur with both variants. This means that they have the lowest information content of the verb given the slot. Moreover, the verbs *stay* and *sit* also have very low information content of the slot given the verb in comparison with the other verbs.

Table 9.2. Frequencies and information content of top twelve verbs most frequently used with *(at) home*

Verb	Bare variant	Prepositional variant	Total frequency with domative	InfoVerb	InfoSlot
be	1796 (77.1%)	505 (22.9%)	2301	0.81	11.06
stay	760 (76.9%)	229 (23.1%)	989	2.03	4.83
sit	62 (24%)	196 (76%)	258	3.97	6.67
live	2 (1.5%)	135 (98.5%)	137	4.88	8.38
work	0 (0%)	66 (100%)	66	5.93	10.28
watch	0 (0%)	54 (100%)	54	6.22	9.22
die	0 (0%)	21 (100%)	21	7.58	10.15
start	0 (0%)	17 (100%)	17	7.89	11.79
wait	1 (5.9%)	16 (94.1%)	17	7.89	10.57
play	0 (0%)	15 (100%)	15	8.07	11.35
happen	0 (0%)	14 (100%)	14	8.17	12.64
remain	1 (7.1%)	13 (92.9%)	14	8.17	9.57

Table 9.3. Estimates and other parameters of the GAM for *(at) home*

Intercept (log-odds of the longer form)	Effective degrees of freedom (tensor product smooths)	*P*	Adjusted R^2	Explained deviance	Scaling parameter
4.63***	5.93	< 0.0001	0.997	98.8%	1

Figure 9.2 shows the results of a binomial Generalized Additive Model with frequencies of the prepositional variant vs. the bare variant as the response, and the information-theoretic variables as predictors (tensor product smooths). See Section 7.4.1 for more details about the method. As shown in Table 9.3, the adjusted pseudo-R^2 is very high (0.997), as well as the explained deviance (99%). This is not surprising, since we have observed a nearly perfect separation of the verbs that are highly frequent with *(at) home* and take part in the alternation,

from those that are much less frequent and allow variation only sporadically. The red islands on the plot correspond to the verbs *stay* (bottom left corner) and *be* (left). If one ignores *be*, the chances of *at home* seem to increase along the diagonal from the bottom left corner to the centre. This suggests that both informativeness measures play a role.

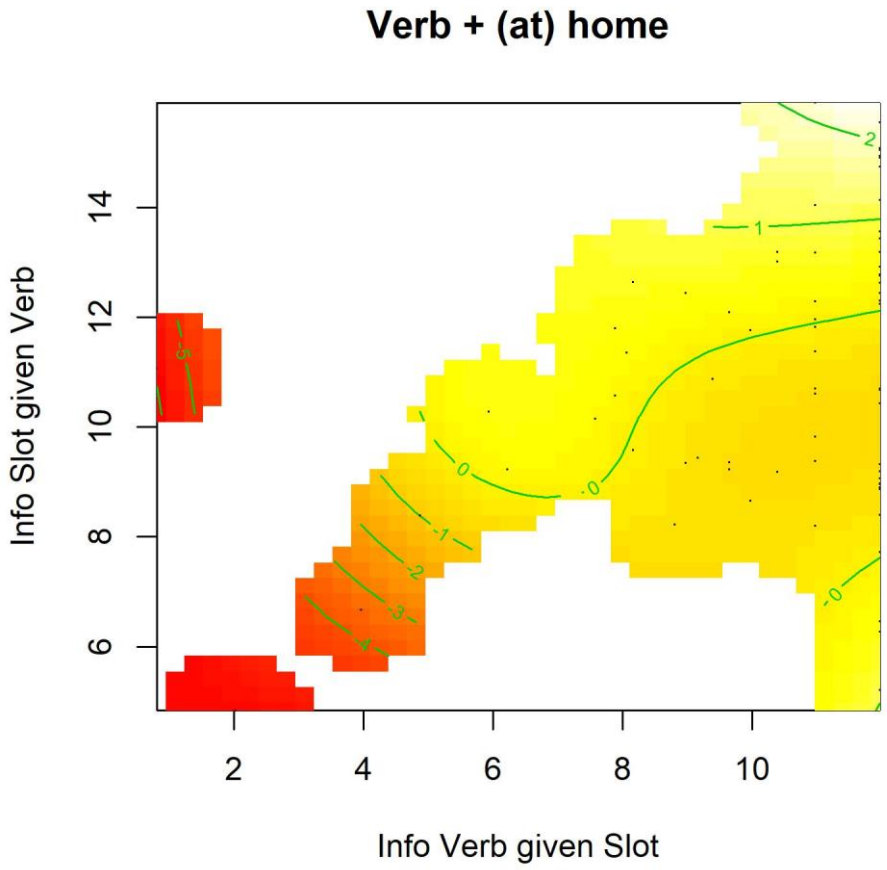

Figure 9.2. Effect of information content on the log-odds of the prepositional variant *at home* (as compared to the bare variant *home*) based on a GAM

9.4. Summary and discussion

This chapter has investigated the variation of locative *(at) home*. The results of the multifactorial analyses show that the variation is constrained by semantics (figurative meanings are expressed nearly exclusively by the longer variant; the semantics of arrival favours the

shorter variant) and syntax: importantly, the variation is mainly possible with intransitive predicates.

Moreover, additional analyses allowed us to identify some predictability effects, which are associated with the most frequent exemplars of the construction. Several verbs that appear with *(at) home* very frequently (i.e. *be*, *sit* and *stay*) are responsible for most variation in the data. The other verbs exhibit very little variability and are used nearly always with *at home*. We observe thus clear exemplar effects. There are also some indications that informativeness of the slot given a verb plays a role. In particular, the verb *stay* is less frequently used with *(at) home* than the verb *be*. However, the two verbs have almost the same proportions of the bare infinitives. The reason may be that *stay* has also the lowest information content of the slot given a verb in the sample. This supports the Hypothesis of Slot-Filler Predictability and Formal Length formulated in Chapter 7.

Chapter 10. Two more cases: *go (and)* Verb and *want to/wanna* + Infinitive

10.1. Aims of this chapter

This chapter concludes the fourth part of this study with a quantitative investigation of two other alternations: *go (and)* Verb and *want to/wanna* + Infinitive. As in the three previous chapters, we will test the Hypothesis of Slot-Filler Predictability and Formal Length. Section 10.2 presents the case study of *go (and)* Verb, whereas Section 10.3 examines *want to/wanna* + Infinitive. The results of the studies are discussed in Section 10.4, which also draws conclusions about slot-filler predictability effects examined in this study.

10.2. Variation of *go (and)* Verb

10.2.1. Theoretical background

This alternation is illustrated by the following example:

(1) *Let's go (and) get some pizza!*

This alternation has been widely discussed in the generativist literature. The main question of these studies is how the shorter variant is derived by formal operations from the construction with *and* (see, e.g., Pullum 1990 and Wulff 2006 for an overview). Some semantic differences have been observed, as well. In particular, Carden and Pesetsky (1979: 81) point out that *go and* + Verb, unlike *go* + Verb, can express unexpected events. They provide the following examples:

(2) a. *??As we had arranged, the President went and addressed the graduating class.*

b. *To our amazement, instead of addressing the graduating class, the President went and harangued the janitors.*

The use of *go and* + Verb is more suitable when the action is surprising, as in (2b) than when it is planned, as in (2a).

In addition, Shopen (1971) argued that *go* + Verb implies volitionality, which is not always observed with *go and* + Verb. The shorter variant is also associated with the motion away from the viewpoint location (*Go and come back to our house!* vs. **Go come back to our house!*).

More recently, Wulff (2006) found that the lexical overlap between the constructions is not very large. This finding undermines the derivational account. Her distinctive collexeme analysis also suggests that the verbs most distinctive of the *go* + Verb construction are process verbs (e.g. *run*, *work*, *walk* and *fly*).

Another relevant usage-based study is Flach (2017), where data about the syntactic environment of the constructions are provided. Her corpus frequencies suggest that *go* + Verb is more frequently used after adhortative *let's* and in the imperative mood, whereas *go and* + Verb is preferred after modals and as a *to*-complement. However, a full multifactorial analysis of this variation is still a task for the future.

According to the Hypothesis of Slot-Filler Predictability and Formal Length (see Chapter 7), one can expect the longer variant to be preferred when a verb has high informative content given the Verb slot of the *go* (*and*) + Verb construction, or the slot is highly informative given the verb. This prediction is tested below.

10.2. Data

From the spoken component of COCA, I extracted all instances of *go* in the base form or *go and.* This data source was chosen because *go* + Verb seems to be more popular in colloquial American English (e.g. Pullum 1990).[53] These strings could be followed by any word, with the exception of function words, adverbials (e.g. *go there*), adjectives (e.g. *go crazy*) and participles (e.g. *go unnoticed* and *go shopping*). After that, it was necessary to check the data manually because some of the verbs were annotated incorrectly (e.g. the second verb in *go figure* was tagged as a noun). The result was 6,540 instances of the alternation with 627 individual verbs. The *go* + Verb construction occurred more frequently than *go and* + Verb: 4,618 occurrences against 1,922 occurrences. I also extracted the token frequencies of individual verbs from the spoken component of COCA. Using this information, I computed the information-theoretic measures as was described in Section 7.2.

10.2.3. Results of quantitative analyses

As in the previous case studies, a quasi-binomial Generalized Additive Model was fitted (due to some overdispersion). The response variable was the frequency of *go and* + Verb vs. *go* + Verb for every individual verb. The predictors were again the information-theoretic variables with a bivariate tensor product smooth. The explanatory power of the model was moderate (see Table 10.1). The settings were as follows: binomial (logit) family, tensor product smooths, thin plate regression splines as the basis, wiggliness parameter gamma = 1.4. Further information about Generalized Additive Models and related concepts can be found in Section 7.4.1.

Table 10.1. Estimates and other parameters of the GAM for *go (and)* Verb

Intercept (log-odds of the longer form)	**Effective degrees of freedom (tensor product smooths)**	***P***	**Adjusted R^2**	**Explained deviance**	**Scaling parameter**
-0.26***	7.11	< 0.0001	0.25	29.3%	1.25

[53] In British English, the distribution is heavily skewed in favour of *go and* + Verb. Wulff (2006) finds only 454 instances of *go* + Verb in the entire British National Corpus, and 5,320 instances of *go and* + Verb.

The effects of both information-theoretical variables are shown in Figure 10.1. The yellow regions represent the values where the chances of *go and* + Verb are higher. They are observed at the top, where the information content of the constructional slot given a verb is higher. Many of those verbs are highly frequent, such as *be*, *become*, *come*, *give* and *join*, e.g. *Do you want to go and be in a coma?* The red regions show the values where the chances of *go and* + Verb are lower. They are observed below, where the information content of a verb given *go (and)* + Verb is lower. Some of the verbs with the lowest values, which are not hapax legomena, are *pee*, *hunt*, *check*, *rent*, *visit* and *golf*, e.g. *Why did you decide to go visit Saddam Hussein?* In such contexts, *go* + Verb represents one event with actual motion in space. Therefore, the shorter variant is preferred when the slot is more expected given the verb, and the longer variant is preferred in the contexts when the slot is less expected. This is exactly what is predicted by the Hypothesis of Slot-Filler Predictability and Amount of Coding. One can also see slightly more yellow in the right-hand side of the plot than in the left-hand side. The verbs on the left-hand side with low information given the slot (e.g. *get*, *do*, *look*, *see*) also have high proportions of the bare form, whereas the verbs on the extreme right with high information are more frequently used with *go and* + Verb. This is a very heterogeneous group of verbs that occur with *go* (*and*) + Verb only once.

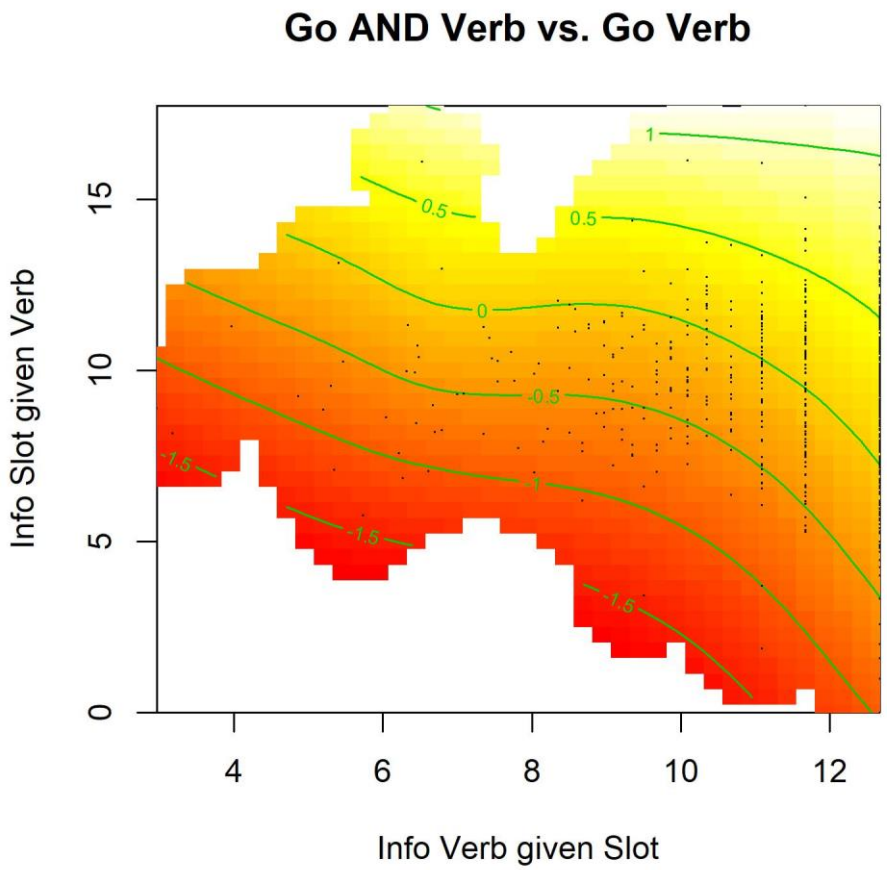

Figure 10.1. Effect of information content on the chances of *go and* + Verb vs. *go* + Verb, based on a GAM

The analyses show that the longer form *go and* + Verb is preferred with verbs that have high informativeness (in both directions). It is noteworthy that the semantics of unexpectedness, which, according to Carden and Pesetsky (1979), is associated with *go and* Verb, is accompanied by relative unexpectedness of its slot fillers, which can be measured quantitatively. More research is needed, of course, in order to disentangle these and other factors that may influence the use of the variants, although preliminary analyses of *go (and)* Verb with additional contextual variables reveal similar effects of the information content variables (Flach and Levshina, In preparation), which gives confidence that the observed effects are robust.

10.3. Variation of *want to/wanna* + Infinitive

10.3.1. Theoretical background

This alternation belongs to the same well-known family as *going to/gonna*, *got to/gotta* and a few other forms (Krug 2000). Examples from a song popular in the 1980s are provided in (3).[54]

(3) a. *Girls just **want to have** fun.*

b. *And girls they **wanna have** fun.*

The default verb that expressed volition in Old English was *willan* 'want'. Over time, it became a future marker. Its functions are now performed by *want.* According to Krug (2000: Chapter 4), *want* developed fully as a modal volitional verb in the 19th century. With the increasing frequency of use, the contracted form *wanna* emerged as a result of fusion of the matrix verb and *to*, which became reanalyzed as one unit. The fusion is also accompanied by formal reduction, as is often the case (cf. Bybee 2006).

According to the Hypothesis of Slot-Filler Predictability and Formal Length, the shorter form *wanna* will be preferred when either the infinitive is more probable given the constructional slot and/or the slot is more probable given the verb. In contrast, the longer form *want to* will be preferred in informative contexts.

10.3.2. Data

For this case study I extracted all instances of *want to* or *wanna* immediately followed by an infinitive from the spoken component of the British National Corpus (BNC 2007). This variation seems to be present in the spoken component only. After removing a few obviously erroneous hits (e.g. *can, want, wan*), there were 6,125 instances of the alternation with 537 individual verbs. The variant *want to* + Verb, which occurred 4,042 times, was almost twice as frequent as *wanna* + Verb with 2,083 occurrences. The frequencies of each verb in the whole

[54] Lyrics of a song by Cyndi Lauper in 1983: https://en.wikipedia.org/wiki/Girls_Just_Want_to_Have_Fun.

spoken component were computed, as well. Based on these frequencies, the informative content measures were computed as shown in Section 7.3.

10.3.3. Results of quantitative analyses

The frequencies of *want to* vs. *wanna* for each verb (transformed into log-odds) served as the binomial response variable. Due to some overdispersion, a quasi-binomial Generalized Additive Model was fitted. The results are shown in Table 10.2. As in the previous cases, the bivariate tensor product smooths turned out to provide a better fit than univariate smooths. The settings were as in the previous model: binomial (logit) family; tensor product smooths, thin plate regression splines as the basis, gamma = 1.4.

Table 10.2. Estimates and other parameters of the GAM for *want to/wanna* + Infinitive

Intercept (log-odds of the longer form)	**Effective degrees of freedom (tensor product smooths)**	***P***	**Adjusted R^2**	**Explained deviance**	**Scaling parameter**
0.97***	12.78	< 0.0001	0.122	13%	1.55

The effects of the information content variables are shown in Figure 10.2. The yellow regions show the areas with the higher chances of *want to* vs. *wanna*, whereas the red regions correspond to lower chances of *want to*. The plot suggests that verbs with both high information content of the verb given the slot and high or medium information content of the slot given the verb have higher chances of the longer form. Many of the verbs located in the top left corner represent, interestingly, mental states or events, such as *agree*, *believe*, *decide*, *feel* and *remember*. The verbs with a relatively strong preference for *wanna* in the left-hand side of the plot are high-frequency verbs *go*, *do* and *get*. In the top left corner are located such high-frequency verbs as *be*, *think* and *have*, which tend to have quite a few uses with *wanna*, hence the slight decrease in the likelihood of *want to*. Many verbs in the 'red belt' at the bottom, which are frequently used with *wanna*, denote perception, such as *see*, *watch*, *hear*, *listen*, *taste*,

or various human activities and accomplishments, e.g. *sing*, *play*, *read*, *brush* (e.g. hair) and *cuddle*.

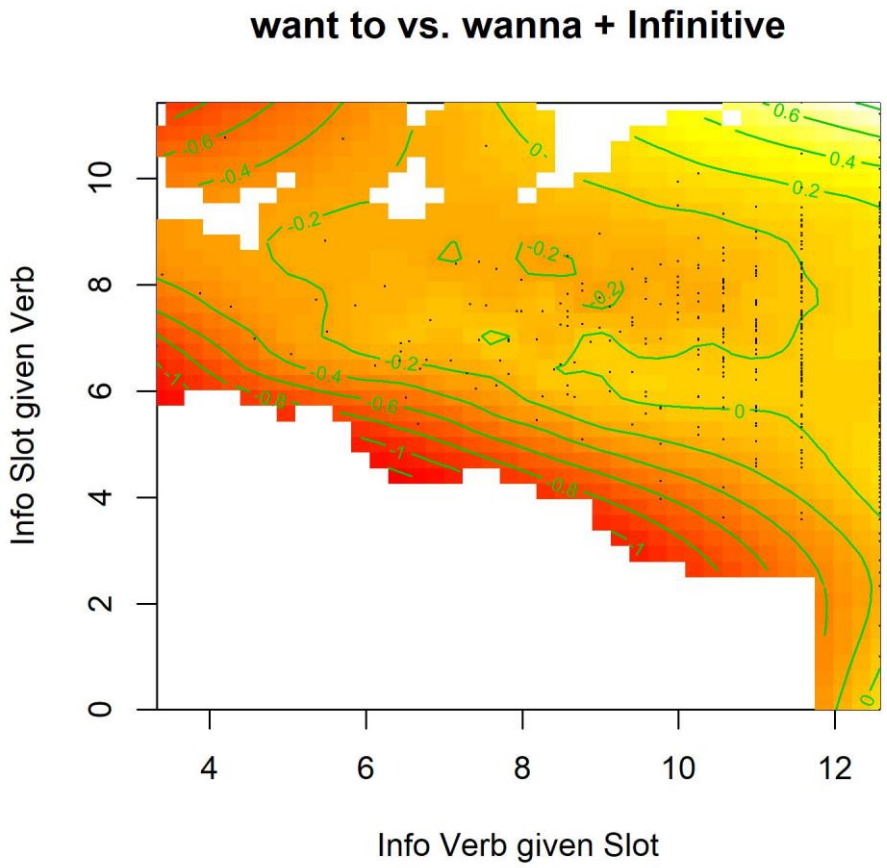

Figure 10.2. Effect of information content on the log-odds of *want to* vs. *wanna*, based on a GAM

The results show that lower information content is more associated with the shorter variant *wanna*, while higher informative content boosts the chances of the longer variant *want to*. In particular, the effect of information content of a verb given the slot is consistent across the entire range. There is also some effect of information content of the slot given a verb in the middle and right-hand parts of the plot. The fact that the explanatory power of the model is low suggests that there are other factors that influence the choice for one or the other variant. Previous research has revealed substantial variation between speakers of different generations, whereby younger speakers choose *wanna* more frequently than older speakers (Krug 2000:

161). There is also geographic variation between different parts of Britain, although the differences seem to be more subtle than the differences between the generations (*Ibid.*: 163–164). These factors need to be taken into account in the future work.

10.4. Summary and discussion

Two small case studies shown in this chapter provide further support for the Hypothesis of Slot-Filler Predictability and Formal Length. Needless to say, these analyses should be refined in multivariate studies of the alternations in question, so that other contextual variables can be taken into account.

More generally speaking, the effects of slot-filler predictability shown in this part can be added to the inventory of frequency effects in language use, structure and change (e.g. Diessel and Hilpert 2016). In addition, the results have consequences for the study of associations between constructions and collexemes. Construction Grammar has traditionally focused on semantic compatibility between those (e.g. Goldberg 1995; Stefanowitsch and Gries 2003). The present study demonstrates that the probabilistic relationships between constructions and collexemes can explain the choice between the shorter and longer constructional variants. This needs to be integrated in the constructionist theory. Another important finding is that the directionality of the relationships between a construction and its collexemes plays an important role in predicting the formal variation in the construction.

Conclusions

The aim of this study was to develop a theory of efficiency in human languages, based on general pragmatic principles. The main claims presented in this thesis are as follows.

1. Language users have a bias towards efficient communication. This means that they tend to maximize the benefit-to-cost ratio, where benefit is the intended cognitive effects in the hearer's mind that help the speaker to reach his or her goals, and the cost is the speaker's effort. Although other types of efforts, e.g. extraction of words from the long-term memory, may also be relevant, the speaker's effort was defined in this study in terms of articulation efforts because articulation is the costliest process in communication.

2. Communication is efficient when small effects have low costs in terms of language production, and large effects have high costs. Since language users assume that their interlocutors act rationally, there are two heuristics that the speaker and the hearer rely on in their interaction: the Low-Cost Heuristic ("Low costs – Small cognitive effects"), and the High-Cost Heuristic ("High costs – Large cognitive effects"). The size of cognitive effects can be defined by how much change should occur in the mental representations of the hearer. The more the intended message deviates from the already activated, easily accessible information and expectations based on the previous knowledge about the world and the language, the greater the cognitive effects.

3. From all that follows that highly predictable, accessible, activated, expected, stereotypical, etc. information tends to be expressed by less costly forms, which require less effort, whereas less predictable, accessible, etc. information tends to be expressed by costlier forms.

4. This correlation can be traced in pairs of related or unrelated forms that convey information with different degrees of activation and predictability:

- I- and M-implicatures, including particularized and generalized ones;
- more and less explicit anaphoric expressions;
- grammatical categories, such as singular and plural;
- grammatical alternations, e.g. the double object dative and the prepositional dative;
- the use and omission of function words and grammatical morphemes, e.g. complementizers and case markers;

- interchangeable analytic and synthetic forms, e.g. *cleverer* vs. *more clever*;
- different variants of the same word, syllable or phrase, reduced or hyper-articulated in language production;
- more and less predictable words and other units of different length, as in Zipf's Law of Abbreviation.

5. The Low-Cost and High-Cost Heuristics are the main mechanisms that are responsible for the emergence of efficient formal asymmetries in the language system. I argue that the Low-Cost Heuristic provides an elegant explanation of the correlation between formal and semantic reduction in the process of grammaticalization, while the High-Cost Heuristic is important for renewal.

6. This study has zoomed in on a set of diverse grammatical phenomena: causative constructions, differential case marking, variation in the use or omission of function words, and the use of full expressions and reduced chunks. In all these cases, the amount of coding negatively correlates with the probability of an intended interpretation given formal cues. In other words, the more probable the grammatical function of an element in question, the shorter the form. This correlation has been observed in cross-linguistic data, and in the variation of English constructions. We can conclude that we are dealing with a universal phenomenon, which is observed both in different languages and across different levels of language structure. More specifically, the results are as follows:

- The correlation between compactness and semantics (e.g. directness and indirectness of causation) in causative constructions of the world can be explained by the asymmetries in different probabilities of attested causation types. Evidence from a sample of typologically and genetically diverse languages, supported by corpus frequencies of different causation types, demonstrates that efficiency provides a better account of the formal differences than iconicity or productivity. Learners of an artificial 'alien' language show a bias towards efficient communication by choosing shorter causative forms to describe more frequent causative situations, and longer forms to describe less frequent ones.
- The typological distribution of differential case marking of A and P, in terms of the associations between the presence or absence of marking and different semantic, syntactic and pragmatic features of A and P, can be best explained by the tendency to mark the arguments that display features with a low cue validity with regard to the

syntactic role, and to leave unmarked the arguments whose semantic and other properties have high cue validity. In other words, the arguments are marked when the given syntactic role is less probable, and not marked with the syntactic role is more probable given the features of the argument. This conclusion is based on the data from nineteen corpora of different languages and genres.

- In language variation, the probability of constructional slots given slot fillers and/or the other way round correlates with the higher odds of the shorter form. This has been demonstrated in case studies of several alternations in English: *help* + (*to*) Infinitive, stative verb + (*at*) home, *go* (*and*) Verb and *want to/wanna* + Infinitive. In addition, the use of the shorter variant is more likely in more frequent subschemata of the *help*-construction (i.e. *help/helps/helped/helping* + (NP) + (*to*) Infinitive) than in less frequent ones. Also, there is intriguing variation in the direction of slot-filler predictability across the subschemata of the *help*-construction in different time periods. As the construction becomes more frequent, the importance of probability of the constructional slot given a lexical filler increases.

These results are important for functional and cognitive linguistics because they provide a coherent framework for discussing different manifestations of efficiency on all levels of language structure – from phonology to discourse patterns. The present study is a first attempt to integrate all these phenomena based on typological evidence, diachronic illustrations, corpus data and experimental evidence. It brings together insights from diverse accounts and theories: Neo-Gricean pragmatics, audience design, Ariel's (1990, 2008) Accessibility Theory, Zipf's Principle of Least Effort, some aspects of markedness theory (cf. Haspelmath 2006) and Optimality Theory (e.g. de Hoop and Malchukov 2008), and many other ideas.

The theory of communicative efficiency yields the following testable predictions for future research:

Prediction 1. Two grammatical or lexical values of the same category that demonstrate asymmetries in formal expression are likely to display differences in their probabilities in language use, so that the less costly form expresses the more probable category, and the costlier form represents the less probable category.

Prediction 2. When a category exhibits a formal split – that is, if there are several different forms that perform the same grammatical function (cf. Haspelmath Forthcoming-a), which

differ in the amount of effort and the choice between which is determined by particular lexical, semantic or formal properties of the corresponding expressions, – the less costly form exhibits a higher conditional probability of the function given these properties, whereas the costlier form exhibits a lower conditional probability.

Prediction 3. When a construction has two or more variants with closely related functions and several subschemata with different probabilities, the less costly variant will be more frequently preferred in the more probable subshemata, and the costlier variant will be more frequently chosen in the less probable subschemata.

Prediction 4. When a construction has two or more variants with closely related functions, the less costly variant will be used when the mutual predictability between the construction slots and their fillers is higher, and the costlier variant will be used when the mutual predictability between the slots and the fillers is lower. See a more exact definition of slot–filler predictability in Chapter 7.

Prediction 5. If a new reduced form competes for some functional domain with an older and non-reduced one, the reduced form will first take over the more probable functions or meanings in that domain. If the new form is an instance of enhancement, it will first occupy the least probable functions.

There remain some open questions and tasks for future research. First of all, the framework presented here needs to be tested on a wider range of linguistic phenomena in the future. I expect many more instances of efficient formal asymmetries to be found in languages of the world. There is a certain danger of being selective, picking up the phenomena that provide support to the Principle of Communicative Efficiency, and ignoring the violations. We need to find out how one can sample the grammatical functions and the corresponding forms systematically, in order to be able to measure the aggregate efficiency of a grammar, and to reject the null hypothesis of no correlation between benefits and costs.

Second, there are many open issues in the debate about the role of audience design and various cognitive processes in the speaker's mind. This study has argued that the higher-level pragmatic mechanisms based on the Low-Cost and High-Cost Heuristics can constrain the lower-level unconscious processes of reduction and enhancement due to automatization, planning difficulties or other speaker-internal mechanisms. According to Croft (2000: 163), "it is clear that the speaker chooses degree of reduction of a constructional form with the hearer in mind". This claim needs experimental support for different linguistic phenomena.

Third, the connection between production efficiency and learning should be explored. A system that is easy to use, may be difficult to acquire, and the other way round. One can hypothesize that there is a trade-off between articulatory efficiency and systemic pressure. The latter may facilitate the acquisition of grammatical patterns.

Fourth, the results should be formalized in the framework of probabilistic pragmatics (Franke and Jaeger 2015). In particular, the Bayesian approach is promising. The previous discourse experience of language users and the knowledge of context can be seen as prior probabilities of a certain interpretation. These priors interact with the cues provided by the speaker, and the resulting posterior probabilities may become priors for the next act of communication. This mechanism should be spelled out formally and tested empirically, e.g. in communicative games.

Finally, some of the variational English models should be extended in order to exclude additional contextual factors. An open question is how to interpret the relationships between different types of slot-filler predictability in synchrony and diachrony. A cross-linguistic perspective on constructional variation would also provide useful insights.

Appendix 1. List of languages in the typological sample

1) List of languages (families) that have pairs of causatives with (in)directness distinction (in a broad sense), used for the case studies in Chapter 4 and Chapter 3. The genetic classification is provided according to the WALS.

Africa (5):

Gumuz (isolate?), Humburi Senni (Songhay), Khoekhoe/Nama (Khoe-Kwadi), Ma'di (Central Sudanic), Noon (Niger-Congo)

Australia (3):

Diyari (Pama-Nyungan), Garrwa (Garrwan), Kayardild (Tangkic)

Eurasia (13):

Ainu (isolate), Basque (isolate), Betta Kurumba (Dravidian), Finnish (Uralic), Great Andamenese (isolate), Hebrew (Afro-Asiatic), Hindi (Indo-European), Japanese (isolate), Korean (isolate), Kusunda (isolate), Lahu (Sino-Tibetan), Nivkh (isolate), Yukaghir Kolyma (Yukaghir)

North America (12):

Caddo (Caddoan), Cherokee (Iroquoian), Chimariko (Hokan), Creek (Muskogean), Filomeno (Totonacan), Lakhota (Siouan), Mutsun (Penutian), Northern Paiute (Uto-Aztecan), Slave (Na-Dene), Takelma (unclear), Teribe (Chibchan), Wappo (Wappo-Yukian)

Papua and Austronesia (4):

Indonesian (Austronesian), Motuna (East Bougainville), Skou (Skou), Yimas (Lower Sepik-Ramu)

South America (9):

Aguaruna (Jivaroan), Apinayé (Macro-Ge), Hup (Nadahup), Mocoví (Guaicuruan), Mosetén (Mosetenan), Trumai (isolate), Urarina (isolate), Waimiri-Atroarí (Cariban), Yagua (Peba-Yaguan)

2) Additional languages with other distinctions related to compactness (relevant for the case study in Chapter 3 only):

Africa (2):

Ik (Eastern Sudanic), Tubu/Dazaga (Saharan)

Papua and Austronesia (3):

Adang (Timor-Alor-Pantar), Manambu (Sepic), Tidore (West Papuan)

South America (2):

Cavineña (Tacanan), Karó (Tupian)

3) Languages where no semantic distinctions between causative constructions have been found (relevant for the case study in Chapter 3 only):

Africa (1):

Sandawe (isolate)

Eurasia (1):

Udihe (Altaic)

Papua and Austronesia (1):

Meninggo/Moskona (Sentani)

South America (3):

Mapuche (Araucanian), Parisi-Haliti (Arawakan), Yuracaré (isolate)

Appendix 2. Corpus frequencies of different A and P from previous studies

Corpus	Study	Role	Total	Lexical	Human	1st & 2nd person	New	Notes
Sakapultek Pear Story	Du Bois 1987	A	187 or 180	11/180 (6.1%)	187/187 (100%)	NA	6/187 (3.2%)	
		P	177 or 170	81/177 (45.8%)	17/170 (10%)	NA	42/170 (24.7%	
Chinese narratives (retelling the film *Ghost*)	Chui 1992	A	407	155 (38.1%)	NA	NA	12 (2.9%)	
		P	363	306 (84.3%)	NA	NA	122 (33.6%)	
French monologic speech	Ashby & Bentivoglio 1993	A	481	32 (6.7%)	NA	NA	0 (0%)	
		P	481	324 (67.4%)	NA	NA	143 (29.7%)	
Spanish monologic speech	Ashby & Bentivoglio 1993	A	571	35 (6.1%)	NA	NA	2 (0.4%)	
		P	571	341 (59.7%)	NA	NA	142 (24.9%)	
Written Swedish	Dahl & Fraurud 1996	A	3127	NA	1766 (56.5%)	NA	NA	Only clauses with overt subject and object
		P	4476	NA	580 (13%)	NA	NA	
Four oral narratives in Modern Hebrew	Sutherland-Smith 1996	A	270	18 (6.7%)	NA	NA	6 (2.2%)	Syntactic transitives + semantic transitives
		P	197	111 (56.3%)	NA	NA	47 (23.9%)	
Spoken Swedish	Dahl 2000	A	2991	NA	Animate 2789 (93.2%)	Egophoric 1815 (60.7%)	NA	
		P	3244	NA	Animate 531 (16.4%)	Egophoric 139 (4.3%)	NA	
Inuktitut child language	Allen & Schröder 2003	A	617, 613 or 616	7/617 (1.1%)	Animate 610/616 (99%)	601/617 (97.4%)	4/613 (0.7%)	Mostly affixal referring expressions
		P	617 or 603	37/617 (6%)	Animate 128/606 (21.1%)	88/617 (14.3%)	163/603 (27%)	
Mapudungun narrative texts	Arnold 2003	A	161	24 (14.9%)	NA	NA	2 (1.2%)	Only main clauses
		P	161	137 (85.1%)	NA	NA	77 (47.8%)	
English talk shows	Everett 2009	A	392	38 (9.7%)	360 (91.8%)	NA	NA	
		P	397	237 (59.7%)	50 (12.6%)	NA	NA	
Portuguese talk shows	Everett 2009	A	155	27 (17.4%)	135 (87.1%)	NA	NA	

		P	163	138 (84.7%)	10 (6.1%)	NA	NA	
Chinese conversations	Lin 2009	A	100	20 (20%)	NA	NA	15 (15%)	
		P	100 or 97	80/100 (80%)	NA	NA	54/97 (55.6%)	
Chinese spoken narratives	Lin 2009	A	80	15 (18.8%)	NA	NA	10 (12.5%)	
		P	83 or 80	78/83 (94%)	NA	NA	58/80 (72.5%)	
Chinese written texts	Lin 2009	A	88	14 (15.9%)	NA	NA	18 (20.5%)	
		P	88 or 87	72/88 81.8%	NA	NA	61/87 (70.1%)	
English autobiographical narratives	Schiborr 2016	A	1046	86 (8.2%)	971 (92.8%)	617 (59%)	NA	
		P	1114	532 (47.8%)	138 (12.4%)	54 (4.8%)	NA	
Northern Kurdish traditional narratives	Haig & Thiele 2016	A	422	55 (13%)	408 (96.7%)	137 (32.5%)	NA	
		P	428	234 (54.7%)	111 (25.9%)	29 (6.8%)	NA	
Persian stimulus-based narratives	Abidifar 2016	A	603	82 (13.6%)	580 (96.2%)	22 (3.6%)	NA	
		P	628	331 (52.7%)	113 (18%)	0 (0%)	NA	
Teop traditional narratives	Mosel & Schnell 2016	A	797	77 (9.7%)	760 (95.4%)	204 (25.6%)	NA	
		P	616	265 (43%)	267 (43.3%)	29 (4.7%)	NA	
Vera'a traditional narratives	Schnell 2016	A	1360	101 (7.4%)	1288 (94.7%)	210 (15.4%)	NA	
		P	917	514 (56.1%)	324 (35.3%)	79 (8.6%)	NA	

Appendix 3. Normalized frequencies of individual subschemata with the bare and to-infinitive.

Table A3.1. The USA: Subschemata without the Helpee (frequencies per million 1-grams)

Period	**Type of infinitive**	**help**	**helped**	**helping**	**helps**
1901-1925	bare	4.12	1.04	0.16	0.31
	with *to*	5.07	6.33	1.92	2.72
	both forms	9.2	7.38	2.08	3.02
1926-1950	bare	9.15	3.17	0.36	1.11
	with *to*	7.32	9.32	2.81	4.16
	both forms	16.47	12.49	3.17	5.27
1951-1975	bare	16.24	6.1	0.52	2.48
	with *to*	8.65	10.34	3.24	5.08
	both forms	24.89	16.44	3.76	7.56
1976-2000	bare	30.71	10.74	0.82	5.99
	with *to*	8.72	8.17	3.49	6.08
	both forms	39.43	18.91	4.31	12.06
2001-2009	bare	32.97	10.71	0.98	6.27
	with *to*	7.85	7.04	3.45	5.65
	both forms	40.82	17.75	4.43	11.92

Table A3.2. The USA: Subschemata with the Helpee (frequencies per million 1-grams)

Period	**Type of infinitive**	**help**	**helped**	**helping**	**helps**
1901-1925	bare	3	0.44	0.13	0.09
	with *to*	4.58	1.36	0.52	0.69
	both forms	7.58	1.8	0.64	0.77
1926-1950	bare	6.15	1.08	0.39	0.31
	with *to*	4.81	1.55	0.73	0.9
	both forms	10.96	2.62	1.12	1.21
1951-1975	bare	9.6	1.51	0.72	0.77
	with *to*	4.43	1.42	0.78	1.04
	both forms	14.03	2.93	1.5	1.81
1976-2000	bare	23.89	3.85	1.64	2.73
	with *to*	4.55	1.56	0.78	1.24
	both forms	28.45	5.41	2.42	3.97
2001-2009	bare	31.53	5.37	2.26	3.75
	with *to*	5.66	1.89	0.89	1.39
	both forms	37.19	7.26	3.15	5.14

Table A3.3. Great Britain: Subschemata without the Helpee (frequencies per million 1-grams).

Period	**Type of infinitive**	**help**	**helped**	**helping**	**helps**
1901-1925	bare	0.74	0.09	0.01	0.03
	with *to*	5.01	6.89	1.48	2.23
	both forms	5.75	6.99	1.49	2.27
1926-1950	bare	1.89	0.45	0.04	0.11
	with *to*	6.7	9.68	2.1	3.27
	both forms	8.59	10.12	3.38	3.38
1951-1975	bare	5.86	1.81	0.1	0.64
	with *to*	9.41	12.48	2.69	5.09
	both forms	15.27	14.29	2.78	5.73
1976-2000	bare	15.77	5.7	0.34	2.68
	with *to*	11.61	12.3	3.75	6.67
	both forms	27.38	18	4.09	9.35
2001-2009	bare	22.22	7.26	0.54	3.96
	with *to*	12.83	10.67	4.18	7.57
	both forms	35.05	17.94	4.72	11.53

Table A3.4. Great Britain: Subschemata with the Helpee (frequencies per million 1-grams).

Period	**Type of infinitive**	**help**	**helped**	**helping**	**helps**
1901-1925	bare	0.61	0.05	0.01	0.01
	with *to*	4.07	1.2	0.4	0.6
	both forms	4.67	1.25	0.41	0.61
1926-1950	bare	1.45	0.19	0.03	0.02
	with *to*	4.53	1.61	0.46	0.75
	both forms	5.98	1.8	0.49	0.77
1951-1975	bare	3.23	0.47	0.13	0.12
	with *to*	4.24	1.64	0.54	0.84
	both forms	7.47	2.12	0.67	0.97
1976-2000	bare	8.18	1.22	0.42	0.67
	with *to*	4.31	1.5	0.7	1.13
	both forms	12.49	2.72	1.12	1.8
2001-2009	bare	13.35	1.8	0.68	1.26
	with *to*	6.08	1.63	0.89	1.63
	both forms	19.43	3.42	1.57	2.89

References

Adibifar, Shirin. 2016. Persian. In: Geoffrey Haig and Stefan Schnell. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts), https://lac.uni-koeln.de/multicast-english/, accessed 11.09.2018.

Aissen, Judith. 1999. Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory* 17: 673–711.

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy, *Natural Language and Linguistic Theory* 21: 435–483.

Allen, Shanley E.M., and Heike Schröder. 2003. Preferred Argument Structure in early Inuktitut spontaneous speech data. In: John W. Du Bois, Lorraine E. Kumpf and William J. Ashby (eds.), *Preferred argument structure: Grammar as architecture for function*, 301–338. Amsterdam: John Benjamins.

Amberber, Mengistu. 2000. Valency-changing and valency-encoding devices in Amharic. In: R. M. W. Dixon and Alexandra Y. Aikhenvald (eds.), *Changing valency: Case studies in transitivity*, 312–332. Cambridge: Cambridge University Press.

Andersen, Henning. 2016. Abduction. In: Ian Roberts and Adam Ledgeway (eds.), *The Cambridge Handbook of Historical Syntax*, 301–321. Cambridge: Cambridge University Press.

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge.

Ariel, Mira. 2001. Accessibility theory: An overview. In: Ted Sanders, Joost Schliperoord and Wilbert Spooren (eds.), *Text representation*, 29–87. Amsterdam: John Benjamins.

Ariel, Mira. 2008. *Pragmatics and Grammar*. Cambridge: Cambridge University Press.

Aristotle. *The Organon and Other Works*. Opensource collection. Translated under the editorship of W.D. Ross. https://archive.org/details/AristotleOrganon.

Arnold, Jennifer E. 2001. The effects of thematic roles on pronoun use and frequency of reference. *Discourse Processes* 31: 137–62.

Arnold, Jennifer E. 2003. Multiple constraints on reference form: Null, pronominal, and full reference in Mapudungun. In: John W. Du Bois, Lorraine E. Kumpf and William J.

Ashby (eds.), *Preferred argument structure: Grammar as architecture for function*. 225–245. Amsterdam: John Benjamins.

Arnold, Jennifer E. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass* 4(4): 187–203. doi.org/10.1111/j.1749-818x.2010.00193.x

Arnold, Jennifer E. and Zenzi M. Griffin. 2007. The effect of additional characters on choice of referring expression: everyone counts. *Journal of Memory and Language* 56: 521–36.

Ashby, William B., and Paola Bentivoglio. 1993. *Language Variation and Change* 5: 61–76.

Asr, Fatemeh Torabi, and Vera Demberg. 2012. Implicitness of discourse relations. In: *Proceedings of COLING 2012: Technical Papers*, 2669–2684. COLING 2012, Mumbai, December 2012.

Aylett, Matthew, and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1): 31–56.

Aylett, Matthew, and Alice Turk. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of Acoustical Society of America* 119(5): 3048–3058.

Baese-Berk, Melissa, and Matthew Goldrick. 2009. Mechanisms of interaction in speech production. *Language and Cognitive Processes* 24(4): 527–554.

Bates, Douglas, Martin Maechler, Ben Bolker and Steve Walker. 2015. Fitting Linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1-48. doi.org/ 10.18637/jss.v067.i0.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13: 145–204.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113(2): 1001-1024.

Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1): 92-111.

Bentz, Christian, and Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. Capturing Phylogenetic Algorithms for Linguistics. In: Christian Bentz, Gerhard Jäger and Igor Yanovich (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen, online publication system, https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

Bickel, Balthasar, Alena Witzlack-Makarevich and Taras Zakharko. 2015. Typological evidence against universal effects of referential scales on case alignment. In: Ina Bornkessel- Schlesewsky, Andrej L. Malchukov and Marc Richards (eds.), *Scales*, 7–43. Berlin: de Gruyter Mouton.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga and John B. Lowe. 2017. *The AUTOTYP typological databases*. Version 0.1.0 https://github.com/autotyp/autotyp-data/tree/0.1.0

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Blakemore, Diane. 1987. *Semantic Constraints on Relevance*. Blackwell, Oxford.

Blumenthal-Dramé, Alice, and Bernd Kortmann. 2017**.** Causal and concessive relations: Typology meets cognition. Paper presented at the 39th Annual Conference of the German Linguistic Society, March 8–10 2017, Saarbrücken.

Bolinger, Dwight. 1963. Length, vowel, juncture. *Linguistics* 1: 5–29.

Bossong, Georg. 1985. *Empirische Universalienforschung: Differentielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Narr.

Bouma, Gosse. 2016. Om-omission. In: Martijn Wieling, Martin Kroon, Gertjan van Noord and Gosse Bouma (eds.), *From Semantics to Dialectometry: Festschrift in honor of John Nerbonne,* 65–74. College Publications.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and Harald Baayen. 2007. Predicting the dative alternation. In: Gerlof Bouma, Irene Krämer and Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bruno, Ana Carla. 2003. *Waimiri Atroari Grammar: Some Phonological, Morphological, and Syntactic Aspects*. Doctoral dissertation, University of Arizona.

Buz, Esteban, Michael K. Tanenhaus and T. Florian Jaeger. 2016. Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language* 89: 68-86.

Bybee, Joan L. 1994. The grammaticization of zero. Asymmetries in tense and aspect systems. In: William Pagliuca (ed.), *Perspectives on grammaticalization*, 235-254. Amsterdam: John Benjamins.

Bybee, Joan L. 1999. Usage-based phonology. In: Michael Darnell, Edith A. Moravcsik, Frederic J. Newmeyer, Michael Noonan and Kathleen Wheatley (eds.), *Functionalism and formalism in linguistics. Volume I: General papers*, 211–242. Amsterdam: John Benjamins.

Bybee, Joan L. 2003. Cognitive processes in grammaticalization. In: Michael Tomasello (ed.), *The New Psychology of Language: Cognitive and functional approaches to language structure. Vol. 2*, 145–167. Mahwah, NJ: Erlbaum.

Bybee, Joan L. 2006. From usage to grammar: the mind's response to repetition. *Language* 82(4): 711–733.

Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, Joan L. 2010. *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.

Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World.* Chicago: The University of Chicago Press.

Bybee, Joan, and Sandra A. Thomspson. 1997. Three frequency effects in syntax. *Berkeley Linguistics Society* 23: 378–388.

Bybee, Joan, and Joanne Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics* 37(4): 575–596.

Caldwell, Christine A., and Kenny Smith. 2012. Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS ONE* 7(8): e43807. doi.org/10.1371/journal.pone.0043807

Carden, Guy, and David Pesetsky. 1979. Double-verb constructions, markedness, and a fake co-ordination. *Papers from the Thirteenth Regional Meeting of the Chicago Linguistic Society* 13, 82–92. Chicago: Chicago Linguistic Society.

Chatterji, Suniti Kumar. 1926. *The Origin and Development of the Bengali Language*. Calcutta: Calcutta University Press.

Chui, Ka-Wai. 1992. Preferred argument structure for discourse understanding. In: *Proceedings of COLING-92, Nantes, August 23-28 1992*, 1142–1146.

Clark, Herbert H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Clark, Herbert H., and Edward F. Schaefer. 1989. Contributing to discourse. *Cognition* 13: 259–294.

Clark, Herbert H., and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22: 1–39.

Cohen Priva, Uriel. 2008. Using information content to predict phone deletion. In Natasha Abner and Jason Bishop (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 90–98. Somerville, MA: Cascadilla Proceedings Project.

Comrie, Bernard. 1976. The syntax of causative constructions: Cross-language similarities and divergencies. In: Masayoshi Shibatani (ed.), *Syntax and Semantic 6: The Grammar of Causative Constructions*, 261–312. New York: Academic Press.

Comrie, Bernard. 1978. Ergativity. In: W. P. Lehmann (ed.), *Syntactic Typology. Studies in the Phenomenology of Language*, 329—394. Austin: The University of Texas Press.

Comrie, Bernard. 1981. *Language Universals and Linguistic Typology*. Oxford: Blackwell.

Comrie, Bernard. 1986. Markedness, grammar, people, and the world. In: Eckman, Fred R., Edith A. Moravcsik and Jessica R. Wirth (eds.), *Markedness*, 85–106. New York: Plenum Press.

Corbett, Grevile, Andrew R. Hippisley, Dunstan Brown and Paul Marriott. 2001. Frequency, regularity, and the paradigm: A perspective from Russian on a complex relation. In:

Joan L. Bybee and Paul J. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure, 201–226.* Amsterdam: John Benjamins.

Croft, William. 2000. *Explaining language change: an evolutionary approach*. Harlow, Essex: Longman.

Croft, William. 2003. *Typology and universals*. 2nd edn. Cambridge: Cambridge University Press.

Cristofaro, Sonia. In press. Taking diachronic evidence seriously: Result-oriented vs. source-oriented explanations of typological universals. In: Karsten Schmidtke-Bode, Natalia Levshina, Susanne M. Michaelis and Ilja Seržant (eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the evidence.* Berlin: Language Science Press.

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7(1): 37–77.

Dahl, Östen, and Kari Fraurud. 1996. Animacy in grammar and discourse. In: Thorstein Fretheim and Jeannette Gundel (eds.), *Reference and referent accessibility*, 47–64. Amsterdam: John Benjamins.

Dalrymple, Mary, and Irina Nikolaeva. 2011. *Objects and information structure*. Cambridge: Cambridge University Press.

Davies, Mark. (2008–) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/.

de Hoop, Helen, and Peter de Swart (eds.). 2008b. *Differential subject marking*. Dordrecht: Springer.

de Hoop, Helen, and Andrej L. Malchukov. 2008. Case-marking strategies. *Linguistic Inquiry* 39(4): 565–587.

Diessel, Holger, and Martin Hilpert. 2016. Frequency effects in grammar. In: Mark Aronoff (ed.), *Oxford Research Encyclopedia of Linguistics*. New York: Oxford University Press. http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-120

Divjak, Dagmar, Natalia Levshina and Jane Klavan. 2016. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* 27(4): 447–464.

Dixon, Robert M. W. 1979. Ergativity. *Language* 55: 59–138.

Dixon, Robert M.W. 1991. *A New Approach to English Grammar, on Semantic Principles*. Oxford: Clarendon Press.

Dixon, Robert M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Dixon, R. M. W. 2000. A typology of causatives: Form, syntax and meaning. In R. M. W. Dixon and Alexandra Y. Aikhenvald (eds.), *Changing valency: Case studies in transitivity*, 30–83. Cambridge: Cambridge University Press.

Dressler, Wolfgang U. 1990. The Cognitive Perspective of 'Naturalist' Linguistic Models. *Cognitive Linguistics* 1: 75–98.

Dryer, Matthew S., and Martin Haspelmath (eds.). 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2018-11-09.)

Du Bois, John W. 1985. Competing motivations. In: John Haiman (ed.), *Iconicity in syntax*, 343–65. Amsterdam: John Benjamins.

Du Bois, John W. 1987. The discourse basis of ergativity. *Language* 63: 805—855.

Du Bois, John W., Lorraine E. Kumpf and William J. Ashby (eds.). 2003. *Preferred argument structure: Grammar as architecture for function*. (Studies in discourse and grammar 14.) Amsterdam: John Benjamins.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson and Nii Martey. 2000–2005. *Santa Barbara Corpus of Spoken American English, Parts 1–4*. Philadelphia: Linguistic Data Consortium.

Duffley, Patrick J. 1992. *The English Infinitive*. London: Longman.

Eksell Harning, K. 1980. *The Analytical Genitive in Modern Arabic Dialects*. Doctoral dissertation, Orientalia Gothoburgensia 5. Gothenburg: University of Gothenburg Press.

Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1): 1-24.

Ellis, Nick C., and Fernando Ferreira-Junior. 2009. Construction Learning as a Function of Frequency, Frequency Distribution, and Function. *The Modern Language Journal* 93(3): 370–385. doi.org/10.1111/j.1540-4781.2009.00896.x

Enfield, N.J. 2007. *A Grammar of Lao*. Berlin: Mouton de Gruyter.

Escandel-Vidal, Victoria, and Manuel Leonetti. 2011. On the rigidity of procedural meaning. In: Victoria Escandell-Vidal, Manuel Leonetti and Aoife Ahern (eds.), *Procedural Meaning: Problems and Perspectives*, 81–102. Bingley: Emerald/Brill.

Evans, Nicholas D. 1995. *A Grammar of Kayardild: With Historical-Comparative Notes on Tangkic* (Mouton Grammar Library 15). Berlin: Mouton de Gruyter.

Everett, Caleb. 2009. A reconsideration of the motivation for preferred argument structure. *Studies in Language* 33(1): 1–24.

Faltz, Leonard M. 1985. *Reflexivization: A study in universal syntax*. New York: Garland.

Fauconnier, Stefanie. 2011. Differential agent marking and animacy. *Lingua* 121(3): 533– 547.

Fay, Nicolas, and T. Mark Ellison. 2013. The cultural evolution of human communication systems in different sized populations: Usability trumps learnability. *PLoS ONE*: e71781. doi.org/10.1371/journal.pone.0071781

Fedzechkina, Maryia, T. Florian Jaeger and Elissa L. Newport. 2012. Language learners restructure their input to facilitate efficient communication. *PNAS* 109(44): 17897–17902. doi.org/10.1073/pnas.1215776109

Fenk, August, and Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Pshychologie* XXVII (3): 400–414.

Ferrer-i-Cancho, Ramon. 2017. The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach. *Glottometrics* 39: 38–71.

Ferrer-i-Cancho, Ramon, and Fermín Moscoso del Prado Martín. 2011. Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*: L12002.

Ferrer-i-Cancho, Ramon, Łukasz Dębowski and Fermín Moscoso del Prado Martín. 2013. Constant conditional entropy and related hypotheses. Journal of Statistical Mechanics: Theory and Experiment 7: L07001. doi:10.1088/1742-5468/2013/07/L07001.

Filimonova, Elena. 2005. The noun phrase hierarchy and relational marking: Problems and counterevidence. *Linguistic Typology* 9(1): 77–113.

Fillmore, Charles J. 1986. Pragmatically Controlled Zero Anaphora. In: *Proceedings of the Berkeley Linguistics Society* 12: 95–107.

Fischer, Olga. 1995. The distinction between bare and to-infinitival complements in late Middle English. *Diachronica* 12: 1–30.

Flach, Susanne K. 2017. *Serial Verb Constructions in English: A usage-based approach.* Doctoral dissertation, Freie Universität Berlin.

Fodor, Jerry A. 1970. Three *reasons* for not deriving "kill" from "*cause to die*". *Linguistic Inquiry* 1(4): 429–438.

Fowler, Carol A., and Jonathan Housum. 1987. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 25: 489–504.

Franke, Michael, and Gerhard Jäger. 2015. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft* 35(1): 3–44.

Gahl, Susanne, and Susan Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80: 748–775.

Garcia, Erica C., and Florimon van Putte. 1989. Forms are silver, nothing is gold. *Folia Linguistica Historica* VIII (1-2): 365–384.

García García, Marco. 2018. Nominal and verbal parameters in the diachrony of differential object marking in Spanish. In: Ilja Seržant and Alena Witzlack-Makarevich (eds.), *Diachrony of Differential Argument Marking*, 209–242. Berlin: Language Science Press.

Geeraerts, Dirk. 2016. The sociosemiotic commitment. *Cognitive Linguistics* 27(4): 527– 542.

Geeraerts, Dirk, Stefan Grondelaers and Peter Bakema. 1994. *The Structure of Lexical Variation. Meaning, naming, and context*. Berlin: Mouton de Gruyter.

Gilligan, Gary Martin. 1987. *A Cross-linguistic Approach to the Pro-drop-parameter.* Doctoral dissertation, University of Southern California.

Givón, Talmy. 1980. The binding hierarchy and the typology of complements. *Studies in Language* 4(3): 333–377.

Givón, Talmy (ed.). 1983. Topic Continuity in Discourse: A quantitative cross-language study. Amsterdam: John Benjamins.

Givón, Talmy. 1984. *Syntax. A Functional-Typological Introduction. Vol. I*. Amsterdam: John Benjamins.

Givón, Talmy. 1990. *Syntax. A Functional-Typological Introduction. Vol. II.* Amsterdam: John Benjamins.

Givón, Talmy. 1995. Markedness as meta-iconicity: distributional and cognitive correlates of syntactic structure. In: Talmy Givón, *Functionalism and Grammar*, 25–69. Amsterdam: John Benjamins.

Givón, Talmy. 2017. *The Story of Zero*. Amsterdam: John Benjamins.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2005. Argument realization: the role of constructions, lexical semantics and discourse factors. In: Jan-Ola Östman and Mirjam Fried (eds.), *Construction Grammars: Cognitive grounding and theoretical extensions*, 17– 44. Amsterdam: John Benjamins.

Goldberg, Adele E., Devin Casenhiser and Nitya Sethuraman. 2004. Learning Argument Structure Generalizations. *Cognitive Linguistics* 15: 289–316.

Goldberg, Adele E., Devin Casenhiser and Nitya Sethuraman. 2005. The Role of Prediction in Construction Learning. *Journal of Child Language* 32: 407-426.

Gordon, Peter C., and Davina Chan. 1995. Pronouns, passives, and discourse coherence. *Journal of Memory and Language* 34: 216–231.

Greenberg, Joseph. 1966. Language Universals, with Special Reference to Feature Hierarchies. (Janua Linguarum, Series Minor, 59.) The Hague: Mouton.

Gregory, Michelle, William D. Raymond, Alan Bell, Eric Fosler-Lussier and Daniel Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. *Proceedings of the Chicago Linguistic Society* 35: 151–166.

Grice, H. Paul. 1975. Logic and Conversation. In: Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, Vol.3. Speech Acts*, 41–58. New York: Academic Press.

Gries, Stefan Th., and Anatol Stefanowitsch. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97–129.

Guirardello, Raquel. 1999. *A Reference Grammar of Trumai*. Doctoral dissertation, Rice University.

Ha, Renee R. 2010. Cost–Benefit Analysis. In: Michael D. Breed and Janice Moore (eds.), *Encyclopedia of Animal Behavior, Vol. 1*, 402–405 Oxford: Academic Press.

Haig, Geoffrey. 2018. The grammaticalization of object pronouns: Why differential object indexing is an attractor state. *Linguistics* 56(4): 781–818. doi.org/10.1515/ling-2018-0011.

Haig, Geoffrey, and Stefan Schnell. 2016. The discourse basis of ergativity revisited. *Language* 92(3): 591–618.

Haig, Geoffrey, and Hanna Thiele. 2016. Northern Kurdish. In: Geoffrey Haig and Stefan Schnell. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts), https://lac.uni-koeln.de/multicast-english/, accessed 11.09.2018.

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

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

Haiman, John. 1991. From V/2 to subject clitics: evidence from Northern Italian. In: Elizabeth Closs Traugott and Bernd Heine (eds.), *Approaches to Grammaticalization. Vol. 1. Focus on theoretical and methodological issues*, 135–157. Amsterdam: John Benjamins.

Haiman, John. 1994. Ritualization and the development of language. In: William Pagliuca (ed.), *Perspectives on Grammaticalization*, 3–28. Amsterdam: John Benjamins.

Hampe, Beate. 2011. Metaphor, constructional ambiguity and the causative resultatives. In: Sandra Handl and Hans-Jörg Schmid (eds.), *Windows to the Mind*, 185–215. Berlin, Mouton de Gruyter.

Harmon, Zara, and Vsevolod Kapatsinski. 2017. Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology* 98: 22–24.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations**.** In: Bernard Comrie and Maria Polinsky (eds.), *Causatives and transitivity*, 87-120. Amsterdam: John Benjamins.

Haspelmath, Martin. 1999. Why is grammaticalization irreversible? *Linguistics* 37: 1043–68.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(1): 25–70.

Haspelmath, Martin. 2008a. A frequentist explanation of some universals of reflexive marking. *Linguistic Discovery* 6(1). DOI:10.1349/PS1.1537-0852.A.331.

Haspelmath, Martin. 2008b. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1): 1–33.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3): 663–687.

Haspelmath, Martin. 2017. Explaining alienability contrasts in adpossessive constructions: Predictability vs. iconicity. *Zeitschrift für Sprachwissenschaft* 36(2): 193-231. doi.org/10.1515/zfs-2017-0009

Haspelmath, Martin. Forthcoming-a. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. Available from https://www.academia.edu/s/b9e468a978/explaining-grammatical-coding-asymmetries-form-frequency-correspondences-and-predictability

Haspelmath, Martin. Forthcoming-b. Role-reference associations and the explanation of argument coding splits. Available at http://ling.auf.net/lingbuzz/004047

Haspelmath, Martin, and Andres Karjus. 2017. Explaining asymmetries in number marking: Singulatives, pluratives and usage frequency. *Linguistics* 55(6): 1213-1235.

Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog and Elif Bamyacı. 2014. Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics* 50(3): 587-625.

Haude, Katharina, and Alena Witzlack-Makarevich. 2016. Referential hierarchies and alignment: An overview. *Linguistics* 54(4): 433–441.

Hawkins, John A. 1994. *A Performance Theory of Order and Constituency.* (Cambridge Studies in Linguistics, 73.) Cambridge: Cambridge University Press.

Hawkins, John. 2014. *Cross-linguistic Variation and Efficiency*. Oxford: Oxford University Press.

Helmbrecht, Johannes, Lukas Denk, Sarah Thanner and Ilenia Tonetti. 2018. Morphosyntactic coding of proper names and its implications for the Animacy Hierarchy. In: Sonia

Cristofaro and Fernando Zúñiga (eds.), *Typological hierarchies in synchrony and diachrony*, 377–401. Amsterdam: Benjamins.

Hilpert, Martin. 2012. *Constructional Change in English: Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University Press.

Hollmann, Willem B. 2003. *Synchrony and Diachrony of English Periphrastic Causatives: A cognitive perspective*. Doctoral dissertation, University of Manchester.

Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William M. Christie (ed.), *Current Progress in Historical Linguistics*, 96–105. Amsterdam: North Holland.

Hopper, Paul J., and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2): 251–299.

Hopper, Paul J., and Elizabeth C. Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Horn, Laurence R. 1984. Towards a new taxonomy for pragmatic inference: Q-based and R-based implicature. In: Deborah Schiffrin (ed.), *Georgetown University Round Table on Languages and Linguistics*, 11–42. Washington, DC: Georgetown University Press.

Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3): 651–674.

Howes, David. 1968. Zipf's law and Miller's random-monkey model. *The American Journal of Psychology* 81(2): 269–272.

Hualde, José Ignacio and Ortiz de Urbina, Jon (eds.). 2003. *A Grammar of Basque.* Berlin: De Gruyter Mouton.

Huang, Yan. 2007. *Pragmatics*. Oxford: Oxford University Press.

Huddleston, Rodney, and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hudson Kam, Carla L., and Elissa L. Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology* 59: 30–66. doi.org/10.1016/j.cogpsych.2009.01.001

Jaeger, T. Florian. 2010. Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology* 61 (1): 23–62. doi.org/10.1016/j.cogpsych.2010.02.002.

Jaeger, T. Florian, and Esteban Buz. 2017. Signal reduction and linguistic encoding. In: Eva M. Fernández and Helen Smith Cairns (eds.), *Handbook of Psycholinguistics*, 38–81. Wiley-Blackwell.

Jakobson, Roman. 1932 [1971]. Zur Structur des russischen Verbums. In: Roman Jakobson. *Selected Writings. Vol. II. Word and Language*, 3–15. Berlin: De Gruyter Mouton.

Jakobson, Roman. 1960 [1971]. Linguistics and Poetics. In: Roman Jakobson. *Selected Writings. Vol. III. Poetry of Grammar and Grammar of Poetry*, 18–51. Berlin: De Gruyter Mouton.

Keenan, Edward L. 1975. Variation in Universal Grammar. In: Ralph Fasold and Roger Shuy (eds.), *Analyzing Variation in Language*, 136–148. Washington, DC: Georgetown University Press.

Keenan, Edward L., and Bernard Comrie. 1977. *Noun Phrase Accessibility and Universal Grammar*. *Linguistic Inquiry* 8(1): 63–99.

Keller, Rudi. 1994. *On Language Change: The invisible hand in language.* London: Routledge.

Kemmer, Susanne, and Arie Verhagen. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5: 115–156.

Kim, Seongho. 2015. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for Statistical Applications and Methods* 22(6): 665-674.

Kirby Simon, Hannah Cornish and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *PNAS* 105: 10681–10686. doi.org/10.1073/pnas.0707835105 .

Kirby, Simon, Tom Griffiths and Kenny Smith. 2014. Iterated learning and the evolution of language. *Current Opinion in Neurobiology* 28: 108–114. doi.org/10.1016/j.conb.2014.07.014

König, Ekkehard, and Letizia Vezzosi. 2004. The role of predicate meaning in the development of reflexivity. In: Walter Bisang, Nikolaus Himmelmann and Björn Wiemer (eds.), *What Makes Grammaticalization? A look from its fringes and its components*, 213–244. Berlin: Mouton de Gruyter.

Koptjevskaja-Tamm, Maria. 1996. Possessive noun phrases in Maltese: Alienability, iconicity, and grammaticalization. *Rivista di Linguistica* 8(1): 245–274.

Krug, Manfred. 2000. *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: Mouton de Gruyter.

Kulikov, Leonid I. 2001. 2001. Causatives. In Martin Haspelmath, Ekkehard König, Wolfgang Oesterreicher and Wolfgang Raible (eds.), *Language typology and language universals: An international handbook*, 886–898. Berlin: Mouton de Gruyter.

Kurumada, Chigusa, and T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83: 152–178.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents.* Cambridge: Cambridge University Press.

Landauer, Thomas K. and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2): 211–240.

Langacker, Ronald W. 1977. Syntactic Reanalysis. In: Charles N. Li (ed.), *Mechanisms of Syntactic Change*, 57–139. Austin: University of Texas Press.

Langacker, Ronald W. 2011. Grammaticalization and Cognitive Grammar. In: Heike Narrog and Bernd Heine (eds.), *The Oxford Handbook of Grammaticalization*, 79–91. Oxford: Oxford University Press.

LaPolla, Randy J., and Chenglong Huang. 2003. *A Grammar of Qiang with annotated texts and glossary*. Berlin: De Gruyter Mouton.

Lastra, Yolanda, and Pedro Martin Butrogueño. 2010. Futuro perifrástico y future morfológico en el Corpus Sociolingüístico de la ciudad de México. *Oralia* 13: 145–171.

Leben, William. 1973. *Suprasegmental phonology*. Doctoral dissertation, *MIT*.

Lee, Hanjung, and Haejeong Choi. 2010. Focus Types and Subject-Object Asymmetry in Korean Case Ellipsis: A New Look at Focus Effects. In: Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, Yasunari Harada (eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, PACLIC 24, Tohoku University, Japan, 4-7 November 2010: 213–222.

Lehmann, Christian. 2015. *Thoughts on Grammaticalization*. 3rd edn. Berlin: Language Science Press.

Levinson, Stephen C. *Presumptive Meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Levshina, Natalia. 2011. *Doe wat je niet laten kan [Do what you cannot let]: A usage-based study of Dutch causatives*. Doctoral dissertation, University of Leuven.

Levshina, Natalia. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50(2): 507–542.

Levshina, Natalia. 2018. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of *help* + (to) Infinitive in varieties of web-based English. *Glossa* 3(1): 55. doi.org/10.5334/gjgl.294

Levshina, Natalia, and Alena Witzlack-Makarevich. 2018. Explaining scale effects in differential case marking: Evidence from dialogue corpora. Paper presented at the 51st Annual Meeting of the Societas Linguistica Europaea. Tallinn, 1 September 2018. https://www.academia.edu/37325241/Explaining_scale_effects_in_differential_case_marking_Evidence_from_dialogue_corpora

Levy, Elena T., and David McNeill. 1992. Speech, gesture and discourse. *Discourse Processes* 15: 277–301.

Levy, Roger, and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt and Thomas Hoffman (eds.), *Advances in Neural Information Processing Systems (NIPS)*, vol. 19, 849–856. Cambridge, MA: MIT Press.

Lin, Wan-hua. 2009. Preferred Argument Structure in Chinese: A Comparison Among Conversations, Narratives and Written Texts. In Yun Xiao (ed.), *Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21)* 2009. Vol. 2, 341 – 357. Smithfield, Rhode Island: Bryant University.

Lind, Age. 1983. The variant forms of *help to*/*help Ø*. *English Studies* 64: 263–275.

Lindblom, Björn. 1984. Economy of speech gestures. In Peter F. MacNeilage (ed.), *The Production of Speech*, 217–245. New York: Springer.

Lindblom, Björn. 1990. Explaining phonetic variation: a sketch of the H & H theory. In: W. J. Hard-castle and A. Marchal (eds.), *Speech Production and Speech Modeling*, 403–439. Dordrecht: Kluwer Academic.

Little, Hannah, Kerem Eryılmaz and Bart de Boer. 2017. Signal dimensionality and the emergence of combinatorial structure. *Cognition* 168: 1–15. doi.org/10.1016/j.cognition.2017.06.011

Lohmann, Arne 2011. Help vs. help to - a multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics* 15(3): 499–521.

Lowrey, Brian. 2012. Early English causative constructions and the "second agent" factor. *Studies in Variation, Contacts and Change in English 10. Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources.* http://www.helsinki.fi/varieng/journal/volumes/10/lowrey/ (last access 9.11.2018).

MacKay, David J.C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126: 313–318.

Mair, Christian. 2002. Three changing patterns of verb complementation in Late Modern English: a real-time study based on matching text corpora. *English Language and Linguistics* 6(1): 105–131.

MacKay, Donald G. 1987. *The Organization of Perception and Action: A theory for language and other cognitive skills*. New York: Springer.

Malchukov, Andrej L. 2008. Animacy and asymmetries in differential case marking. *Lingua* 118(2): 203–221.

Maslova, Elena. 2003. *A Grammar of Kolyma Yukaghir*. [Mouton Grammar Library 27]. Berlin: De Gruyter Mouton.

Matisoff, James A. 1976. Lahu causative constructions: case hierarchies and the morphology/syntax cycle in a Tibeto-Burman perspective. In: Masayoshi Shibatani (ed.), *Syntax and Semantic 6: The Grammar of Causative Constructions*, 413-42. New York: Academic Press.

McCawley, James D. 1978. Conversational implicature and the lexicon. In: Peter Cole (ed.), *Syntax and Semantics. Vol. 9. Pragmatics*, 245–59. New York: Academic Press.

McEnery, Anthony, and Zhonghua Xiao. 2005. HELP or HELP to: What do corpora have to say? *English Studies* 86(2): 161–187.

McFarland, Teresa Ann. 2009. *The phonology and morphology of Filomeno Mata Totonac.* Doctoral dissertation, University of California Berkeley.

McGregor, William B. 2018. Emergence of optional accusative case marking in Khoe languages. In: Ilja Seržant and Alena Witzlack-Makarevich (eds.), *Diachrony of Differential Argument Marking*, 243–279. Berlin: Language Science Press.

Michaelis, Susanne M. 2017. Asymmetry in path coding: Creole data support a universal trend. Paper presented at the SPCL meeting Tampere, June 2017. doi.org/10.5281/zenodo.1456803

Michaelis, Susanne M. In press. Support from creole languages for functional adaptation in grammar: Dependent and independent possessive person-forms. In: Karsten Schmidtke-Bode, Natalia Levshina, Susanne M. Michaelis and Ilja Seržant (eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the evidence.* Berlin: Language Science Press.

Miller, George A. 1957. Some effects of intermittent silence. *The American Journal of Psychology* 70 (2): 311–314.

Mithun, Marianne. 2002. An invisible hand at the root of causation: The role of lexicalization in the grammaticalization of causatives. In: Ilse Wischer and Gabriele Diewald (eds.), *New Reflections on Grammaticalization*, 237–257. Amsterdam: John Benjamins.

Mondorf, Britta. 2003. Support for *more*-support. In: Günther Rohdenburg and Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*, 251–304. Berlin/New York: Mouton de Gruyter.

Mondorf, Britta. 2014. (Apparently) competing motivations in morpho-syntactic variation. In: Edith A. Moravcsik, Andrej Malchukov and Brian MacWhinney (eds.), *Competing Motivations in Grammar and Usage*, 209-228. Oxford: Oxford University Press.

Montaut, Annie. 2018. The rise of differential object marking in Hindi and related languages. In: Ilja Seržant and Alena Witzlack-Makarevich (eds.), *Diachrony of Differential Argument Marking*, 281–313. Berlin: Language Science Press.

Moriya, Akira. 2017. Causative “make” in the King James Bible (1611): Possible factors influencing the choice of bare and *to*-infinitives. *Zephyr* 29: 44-58. doi.org/10.14989/227415

Moscoso del Prado, Fermin. 2014. Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax. *Proceedings of the Annual Meeting of the Cognitive Science Society* 36: 2657–2662. Available from https://escholarship.org/uc/cognitivesciencesociety?volume=36;issue=36 (last access 27.07.2017).

Mosel, Ulrike and Schnell, Stefan. Teop. 2016. In: Geoffrey Haig and Stefan Schnell. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts), https://lac.uni-koeln.de/multicast-english/, accessed 11.09.2018.

Næss, Åshild. 2007. *Prototypical Transitivity*. Amsterdam: John Benjamins.

Nedjalkov, Vladimir P., and Galina A. Otaina. 2013. *A Syntax of the Nivkh Language: The Amur dialect*. Amsterdam: John Benjamins.

Newman, John, and Sally Rice. 2006. Transitivity schemas of English EAT and DRINK in the BNC**.** In: Stefan Th. Gries and Anatol Stefanowitsch (eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, 225–260. Amsterdam: John Benjamins.

Newmeyer, Frederick J. 2003. Grammar is grammar and usage is usage. *Language* 79: 682–707.

Okrand, Marc. 1977. *Mutsun grammar*. Doctoral dissertation, University of California Berkeley.

Olawsky, Knut. 2006. *A Grammar of Urarina* (Mouton Grammar Library 37). Berlin: Mouton de Gruyter.

Pechenick, Eitan A., Christopher M. Danforth and Peter Sh. Dodds. 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS ONE* 10(10): e0137041. doi.org/10.1371/journal.pone.0137041

Petré, Peter. 2017. The extravagant progressive: an experimental corpus study on the history of emphatic [*be* Ving]. *English Language and Linguistics* 21(2): 227–250.

Piantadosi, Steven, Harry Tily and Edward Gibson. 2011. Word lengths are optimized for efficient communication, *Proceedings of the National Academy of Sciences* 108(9): 3526.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In: Joan Bybee and Paul Hopper (eds.), *Frequency effects and the emergence of lexical structure*, 137–157. Amsterdam: John Benjamins.

Poplack, Shana, and Sali Tagliamonte. 1996. Nothing in context: Variation, grammaticization and past time marking in Nigerian Pidgin English. In: Philip Baker and Anand Syea (eds.), *Changing Meanings, Changing Functions. Papers related to grammaticalization in contact languages*, 71–94. Westminster, UK: University Press.

Popper, Karl. 1972. *Objective Knowledge: An Evolutionary Approach*, Oxford: Clarendon Press.

Pullum, Geoffrey K. 1990. Constraints on intransitive quasi-serial verb constructions in modem colloquial English. *The Ohio State University Working Papers in Linguistics* 39: 218–239.

Quine, Willard V. O. 1969. *Ontological Relativity and Other Essays*, New York: Columbia University Press.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reali, Florencia, and Thomas L. Griffiths. 2009. The evolution of frequency distributions: Relating regularization to inductive biases through iterative learning. *Cognition* 111: 317–328. doi.org/10.1016/j.cognition.2009.02.012

Rohdenburg, Günther. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2): 149–182.

Rohdenburg, Günther. 2003. *Horror aequi* and cognitive complexity as factors determining the use of interrogative clause linkers. In: Günther Rohdenburg and Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 205–250. Berlin: Mouton de Gruyter.

Rohdenburg, Günther. 2009. Grammatical divergence between British and American English in the nineteenth and early twentieth centuries. In: Ingrid Tieken-Boon van Ostade and

Wim van der Wurff (eds.), *Current issues in Late Modern English* (Linguistic Insights 77), 301–330. Bern: Peter Lang.

Rosch, Eleanor, and Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7: 573–605.

Royen, Gerlach. 1929. *Die nominalen Klassifikations-Systeme in den Sprachen der Erde. Historische Studie, mit besonderer Berücksichtigung des Indogermanischen*. Wien: Anthropos.

Schnell, Stefan. 2016. Vera'a. In: Geoffrey Haig and Stefan Schnell. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts), https://lac.uni-koeln.de/multicast-english/, accessed 11.09.2018.

Schiborr, Nils Norman. English. 2016. In: Geoffrey Haig and Stefan Schnell. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts), https://lac.uni-koeln.de/multicast-english/, accessed 11.09.2018.

Schlüter, Julia. 2009. The conditional subjunctive. In: Günter Rohdenburg and Julia Schlüter (eds.), *One language, two grammars?: Differences between British and American English*, 277–305. Cambridge: Cambridge University Press.

Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells. From corpus to cognition*. Berlin: Mouton de Gruyter.

Schmid, Hans-Jörg. 2015. A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association* 3: 1–27.

Schmid, Hans-Jörg, and Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3): 531–577.

Schmidtke-Bode, Karsten. In press. Attractor states and diachronic change in Hawkins' "Processing Typology". In: Karsten Schmidtke-Bode, Natalia Levshina, Susanne M. Michaelis and Ilja Seržant (eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the evidence.* Berlin: Language Science Press.

Schmidtke-Bode, Karsten, and Natalia Levshina. 2018. Reassessing scale effects on differential case marking: Methodological, conceptual and theoretical issues in the quest for a universal. In: Ilja A. Seržant and Alena Witzlack-Makarevich

(eds.), *Diachrony of Differential Argument Marking, 509–537*. Berlin: Language Science Press.

Seiler, Walter. 1984. *The main structures of Imonda: a Papuan language.* Doctoral Dissertation, The Australian National University.

Seržant, Ilja. In press. Weak universal forced: The discriminatory function of case in differential object marking systems. In: Karsten Schmidtke-Bode, Natalia Levshina, Susanne M. Michaelis and Ilja Seržant (eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the evidence.* Berlin: Language Science Press.

Seyfarth, Scott. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133(1): 140-155.

Seyfarth, Scott, Esteban Buz and T. Florian Jaeger. 2016. Dynamic hyperarticulation of coda voicing contrasts. *Journal of the Acoustical Society of America* 139(2): EL31–37.

Shibatani, Masayoshi, and Prashant Pardeshi. 2002. The causative continuum. In: Masayoshi Shibatani (ed.), *The Grammar of Causation and Interpersonal Manipulation*, 85–126. Amsterdam: John Benjamins.

Siewierska, Anna. 2004. *Person*. Cambridge: Cambridge University Press.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In: R. M. W. Dixon (ed.), *Grammatical Categories in Australian Languages*, 112—171. Canberra: Australian Institute for Aboriginal Studies, Canberra.

Sinnemäki, Kaius. 2014. A typological perspective on Differential Object Marking. *Linguistics* 52(2): 281–313.

Shopen, Timothy. 1971. Caught in the act: An intermediate stage in a would-be historical process providing syntactic evidence for the psychological reality of paradigms. *Papers from the Seventh Regional Meeting* of the *Chicago Linguistic Society* 7: 254–263.

Smith, Kenny, Amy Perfors, Olga Fehér, Anna Samara, Kate Swoboda and Elizabeth Wonnacott. 2017. Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B* 372: 20160051. doi.org/10.1098/rstb.2016.0051

Smith, Kenny, and Elizabeth Wonnacott. 2010. Eliminating unpredictable variation through iterated learning. *Cognition* 116: 444–449. doi.org/10.1016/j.cognition.2010.06.004

Song, Jae Jung. 1996. *Causatives and Causation: A Universal-Typological Perspective*. London: Addison Wesley Longman.

Sóskuthy, Márton. 2017. Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv preprint arXiv:1703.05339*.

Sperber, Dan and Deirdre Wilson. 1986/1995. *Relevance: Communication and Cognition*. Oxford: Blackwell Publishers.

Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–43.

Stent, Amanda J., Marie K. Huffman and Susan E. Brennan. 2008. Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication* 50(3): 163–178.

Sutherland-Smith, Wendy. 1996. Spoken narrative and Preferred Argument Structure: Evidence from modern Hebrew discourse. *Studies in Language* 20(1): 163–189.

Szmrecsanyi, Benedikt. 2003. Be going to versus will/shall: does syntax matter? *Journal of English Linguistics* 31(4): 295–323.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In: Geeraerts, Dirk, Gitte Kristiansen and Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 141–166. Berlin/New York: De Gruyter Mouton.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller and Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2): 109–137.

Tagliamonte, Sally, and R. Harald Baayen. 2012. Models, forests and trees of York English: *Was*/*were* variation as a case study for statistical practice. *Language Variation and Change* 24(2): 135–178.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. *Vol. 1*. Cambridge: MIT Press.

Tamariz, Monica. 2016. Experimental studies on the cultural evolution of language. *Annual Review of Linguistics* 3: 389–407. doi.org/10.1146/annurev-linguistics-011516-033807

Tamura, Suzuko. 2000. *The Ainu language* (ICHEL Linguistic Studies 2). Tokyo: Sanseido.

Taylor, John R. 2012. *Mental Corpus: How language is represented in the mind*. Oxford: Oxford University Press.

Thomson, Alexander. 1909. Beiträge zur Kasuslehre. *Indogermanische Forschungen* 24: 293–307.

Tiersma, Peter Meijes. 1982. Local and general markedness. *Language* 58(4): 832–849.

Tomasello, Michael. 2008. *The Origins of Human Communication*. Cambridge, MA: MIT Press.

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

Vajrabhaya, Prakaywan. 2016. *Cross-Modal Reduction: Repetition of words and gestures*. Doctoral dissertation, University of Oregon.

Verhagen, Arie, and Suzanne Kemmer. 1997. Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics* 27: 61–82.

Verhoef, Tessa. 2012. The origins of the duality of patterning in artificial whistled languages. *Language and Cognition* 4(4): 357–380. doi.org/10.1515/langcog-2012-0019

van der Horst, Joop M. 1998. *Doen* in Old and Early Middle Dutch: A comparative approach. In: Ingrid Tieken-Boon van Ostade, Marijke van der Wal and Arjan van Leuvensteijn (eds.), *"Do" in English, Dutch and German. History and present-day variation*, 53–64. Münster: Nodus Publicationen.

von Heusinger, Klaus, and Georg A. Kaiser. 2007. Differential object marking and the lexical semantics of verbs in Spanish. In: Georg A. Kaiser and Manuel Leonetti (eds.), *Proceedings of the Workshop "Definiteness, Specificity and Animacy in Ibero-Romance Languages"*, 83–109. Universität Konstanz: Fachbereich Sprachwissenschaft.

von Heusinger, Klaus, and Edgar Onea Gáspár. 2008. Triggering and blocking effects in the diachronic development of DOM in Romanian. *Probus* 20(1): 67 – 110. doi.org/10.1515/PROBUS.2008.003.

Walter, Mary Ann, and T. Florian Jaeger. 2008. Constraints on optional *that*: A strong word form OCP effect. In: Rodney L. Edwards, Patrick J. Midtlyng, Colin L. Sprague and Kjersti G. Stensrud (eds.), *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 505–519. Chicago, IL: CLS.

Wasow, Thomas, T. Florian Jaeger and David M. Orr. 2011. Lexical variation in relativizer frequency. In: Horst J. Simon and Heike Wiese (eds.), *Expecting the unexpected: Exceptions in grammar*, 175–195. Berlin: De Gruyter Mouton.

Wasow, Thomas, Roger Levy, Robin Melnick, Hanzhi Zhu and Tom Juzek. 2015. Processing, prosody, and optional to. In: Lyn Frazier and Edward Gibson (eds.), *Explicit and implicit prosody in sentence processing*, 133–158. New York: Springer.

Wieling, Martijn. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70: 86–116.

Wilkes-Gibbs, Deanna, and Herbert H. Clark. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language* 31: 183–194.

Wilson, Deirdre, and Dan Sperber. 1993. Linguistic form and relevance. *Lingua* 90: 1–25.

Witkowski, Stanley R., and Cecil H. Brown. 1983. Marking-reversals and cultural importance. *Language* 59(3): 569–582.

Witzlack-Makarevich, Alena, and Ilja Seržant. 2018. Differential argument marking: Patterns of variation. In: Ilja Seržant and Alena Witzlack-Makarevich (eds.), *Diachrony of Differential Argument Marking*, 1–40. Berlin: Language Science Press.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

Wulff, Stefanie. 2006. *Go*-V vs. *go-and*-V in English: A case of constructional synonymy? In: Stefan Th. Gries and Anatol Stefanowitsch (eds.). *Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis*, 101–125. Berlin: Mouton de Gruyter.

Yang, Suying. 1995. *The Aspectual System of Chinese*. Doctoral dissertation, University of Victoria.

Zach, Reto. 1979. Shell Dropping: Decision-Making and Optimal Foraging in Northwestern Crows. *Behaviour* 68: 106–117. doi.org/10.1163/156853979X00269

Zemskaja, Elena L., and L. A. Kapanadze (eds.). 1978. Russkaja Razgovornaja Reč. Teksty [Russian Colloquial Speech. Texts]. Moscow: Nauka.

Zipf, George K. 1935 [1965]. *The Psychobiology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.

Zipf, George K. 1949. Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley.