Part I

Different Types of Efficiency in Language

1 Communicative Efficiency
Main Concepts

1.1 What Is Communicative Efficiency?

Generally speaking, efficiency means minimization of a cost-to-benefit ratio. In other words, being efficient means not spending more effort than necessary in order to achieve something. This idea is popular nowadays. We are taught to work smarter, not harder. We are advised to keep only things and human contacts that are meaningful to us. We are expected to practise time management and use energy-efficient cars and gadgets.

Efficiency is an inherent property of living organisms. It is a product of biological evolution. Individuals who behave efficiently will be more fit and ultimately will leave more copies of their genes (Ha 2010). There is plenty of evidence that humans and other animals behave efficiently in foraging, parental investment, cooperation and sibling rivalry. For example, the kinematic paths of motion of humans minimize the energy costs of movement (Anderson and Pandy 2001). Penguins waddle because it conserves energy. If they did not, this would result in more work being required from the muscles (Griffin and Kram 2000). 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 amount of calories gained from whelks and the amount of calories spent flying.

But efficiency is not only a result of biological evolution. It also comes with practice. For example, professional runners position their heels in such a way as to lower metabolic energy consumption (Scholz et al. 2008; see also Napoli and Liapis 2019). Since language is a very old and frequent human activity, we have had many opportunities to optimize it, both in phylogeny and ontogeny.

Human language as such can be regarded as a very efficient tool because it helps us to save time and effort when we need something from others.

Language has created huge benefits for us as a species, allowing us to build large and complex societies and cope with many challenges. At the same time, we tend to save our articulatory and processing effort while using language. For example, during the COVID-19 pandemic, people all over the world started using abbreviated names for *coronavirus*. The clipped form *corona* is particularly popular, being used in many languages, such as Bengali, Hebrew, Indonesian, Malayalam and Romanian. In Dutch, as well as in German, Danish and Swedish, *corona* is particularly frequent in compounds. The Dutch, for example, speak about *coronapatiënten* 'corona-patients', *coronadoden* 'corona-deaths' and *coronatests*. They must adhere to *coronaregels* 'corona-rules' and deal with the *coronacrisis*. In short, we are living in the *coronatijd* 'corona-time' at the moment. Speakers of Australian English are probably the champions in least effort. They have come up with a radically shortened form, *rona*. One would say, *I'm in iso [self-isolation] because of rona*.[1]

Moreover, we are aware of our tendency to save effort. We can even use it as an excuse. For example, at one meeting Donald Trump called Tim Cook, the Apple CEO, 'Tim Apple'. After the media started making fun of his gaffe, Trump posted a message on Twitter, saying that he had been trying to 'save time & words':

> At a recent round table meeting of business executives, & long after formally introducing Tim Cook of Apple, I quickly referred to Tim + Apple as Tim/Apple as an easy way to save time & words. The Fake News was disparagingly all over this, & it became yet another bad Trump story! Donald J. Trump (@realDonaldTrump) 11 March 2019

Thus, efficiency is an important aspect of language communication. But it is not easy to study. Unfortunately, it is impossible to tell exactly how efficient a particular utterance is in a specific context. The reason is that we cannot measure all costs and all benefits of communication (see more in Section 1.2). Instead, we can compare alternative expressions that convey similar meanings and say which one is more costly, and which is less. In many situations the speaker[2] can choose between expressions of different length. Some examples are given in (1). In (1a), one can use the lexical causative *stop* or the periphrastic causative *get X to stop*. Example (1b) illustrates the use of different referential expressions: the longer proper name *Jennifer*, and the shorter pronominal form *she*. In (1c), the difference between the sentences is

[1] I thank Peter Austin for this example. See more information in the MPI TalkLing blog: www.mpi-talkling.mpi.nl/?p=36&lang=en (last access 4 June 2022).

[2] In this book I discuss mostly spoken languages. Still, I expect the general principles and strategies of efficient spoken or written communication to be applicable in many cases of signed communication (but see Section 1.3). To what extent this working hypothesis holds is a question for future research.

in the use or absence of the complementizer *that*. In (1d), the speaker can choose between the clipped form *maths* and the full form *mathematics*. The example in (1e) contrasts the analytic and synthetic comparative forms of adjectives, which can sometimes be used interchangeably in English. The example in (1f) 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. The example in (1g) is an instance of the genitive alternation, where the Saxon genitive with *-'s* is shorter than the Norman genitive with *of* and also allows for omission of some determiners.

(1) a. *John **stopped** the car. – John **got** the car **to stop**.*
b. ***Jennifer** entered the room. – **She** entered the room.*
c. *She believes you are here. – She believes **that** you are here.*
d. *I'm studying **maths**. – I'm studying **mathematics**.*
e. *Ann is **cleverer** than Mary. – Ann is **more clever** than Mary.*
f. *Dunno [dəˈnəʊ]. – I don't know [aɪ dəʊn(t) ˈnəʊ].*
g. *the emperor**'s** family – the family **of the** emperor*

In all these pairs, the costs of articulation are lower if the speaker uses the shorter variant. It also costs less time. But the shorter variant is not always the best one. Sometimes one needs to use a more effortful expression in order to make sure that the intended meaning is conveyed. For example, if there is a chance of phonetic misinterpretation, one will use hyperarticulation: *It's not a **p**in, it's a **b**in*. Also, when talking to a stranger, a local is unlikely to use an abbreviated variant of a toponym. For example, if a Berliner says *Alex* instead of *Alexanderplatz* when giving directions to a tourist, they are likely to be misunderstood. We speak about efficiency if people use less costly expressions, at the same time conveying the intended meaning. In many cases, this means using shorter forms to convey easily accessible meanings, and longer forms to convey less accessible ones. More examples of such contrasts can be found in Chapter 2.

But efficiency is not only about saving articulation effort and time. Different structures can be more or less efficient from the perspective of language processing. For example, (2) illustrates variation in the order of syntactic constituents. According to some theories, the sentence in (2a), where the short prepositional phrase precedes the long object, requires less processing effort than the sentence in (2b), where the order is reversed. The reason is that (2b) has longer syntactic dependencies, which create higher memory costs. These issues are discussed in detail in Chapter 3.

(2) a. *I met [on the street] [my eccentric aunt from San Francisco].*
b. *I met [my eccentric aunt from San Francisco] [on the street].*

In the above-mentioned examples, users have choice between more and less costly expressions. Very often, these choices become conventionalized and

associated with different meanings, grammatical categories or registers. They become obligatory. A typical example is the singular–plural distinction. Cross-linguistically, singular forms are less often marked formally than plural forms (Greenberg 1966), as illustrated by the pair *book – books* in (3a). In (3b), the shorter form *furniture* has a collective use, whereas the longer form *a piece of furniture* has a singulative meaning. In (3c), the comparative forms of adjectives are more costly than the positive forms.

(3) a. *(one) book-Ø – (five) book-**s***
b. *furniture – **a piece of** furniture*
c. *nice – nic**er**, expensive – **more** expensive.*

Unlike in (1) and (2), the speaker has no choice because the constructions convey different categories and meanings (although one can find languages where number marking is optional, for instance). Still, these asymmetries are efficient because more frequent meanings and categories are expressed by less costly forms. This saves the total amount of effort and time in the long run.

Finally, we can compare the costs of expressions which are not functionally or formally related at all, provided we can also compare their accessibility. According to Zipf's (1965 [1935]) Law of Abbreviation, more frequent words tend to be shorter than less frequent ones. Compare, for example, short and frequent words *I*, *in* and *be*, with long and rare words *harpsichord*, *archaeopteryx* and *gongoozle* 'to watch the passage of boats'. Although one can also find many pairs of words in which the frequent member is longer than the rare one (e.g., the word *understand* is more frequent in everyday language than a physics term *quark*, but the former is longer than the latter), any text of sufficient length will yield a significant negative correlation between frequency and length (see Section 2.6).

The idea of minimizing the costs of communication while keeping the benefits has a long tradition in linguistics. In fact, already in the 19th century similar ideas were used to explain the processes of grammaticalization and sound change. For example, Georg Curtius (1820–1985), a German philologist, explained phonetic attrition (*Verwitterung* 'weathering') by the drive to *Bequemlichkeit* 'comfort'. This drive is counterbalanced by the tendency to preserve meaning-bearing sounds and syllables, which resist attrition in order to be recognizable (Delbrück 1919: 143–144). Therefore, language users try to minimize their effort, at the same time making sure that the meanings are conveyed. Similarly, William Dwight Whitney (1875: 69) wrote about the tendency towards ease and economy as a driving force of the process of assimilation. He also mentioned that what is easy to the 'practised speaker' is not necessarily what is easy for second language learners and children, thus pointing to the potential conflict with learnability – another important factor in language evolution.

Zipf not only formulated the Law of Abbreviation (see above), but also contemplated the causes of efficient behaviour. He argued that language users act as rational 'artisans', who follow the Principle of Least Effort (Zipf 1949; see also Section 5.4.2). Among more recent approaches, one can mention the following closely related principles and hypotheses:

- Haiman's (1983) principle of economy;
- Du Bois' (1985) dictum 'Grammars code best what speakers do most';
- Cristofaro's (2003) principle of Information Recoverability;
- Hawkins' (2004) principle 'Minimize Forms';
- Givón's (2017: 157) code–quantity principle;
- Haspelmath's (2021a) form–frequency correspondence hypothesis.

Efficient word order has also received substantial attention. One of the earliest contributions is Behaghel's (1909) law of growing constituents, which says that of two constituents of different length, the longer constituent follows the shorter one. This provides advantages both for production and comprehension. Later, Yngwe (1960) wrote about efficient word orders generated by a formal language model, which put less demands on working memory. Hawkins (2014) formulated the principles 'Minimize Domains' and 'Maximize On-line Processing'. One manifestation of word order efficiency which has received a lot of attention recently is so-called dependency distance minimization (Ferrer-i-Cancho 2006; Liu 2008; Futrell, Mahowald and Gibson 2015b; see also Chapter 3).

The speaker's efficient choices are also discussed in pragmatics. In particular, they are captured by some of the Gricean and Neo-Gricean principles, maxims and heuristics (Grice 1975; Horn 1984; Levinson 2000), as will be shown in Section 1.4. I should also mention here Keller's 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' (Keller 1994: 107).

In the recent decades, these and similar ideas have been tested on large and typologically diverse corpora with the help of advanced quantitative methods (see Levshina and Moran 2021 for an overview). Examples are phonological studies of language production, focusing on duration of words and articulation or omission of certain sounds (e.g., Cohen Priva 2008; Bell et al. 2009; Seyfarth 2014), use and omission of optional grammatical markers, such as complementizers or relativizers (e.g., Jaeger 2006; Wasow, Jaeger and Orr 2011), or the above-mentioned dependency distances. In addition to corpora, we can rely on other methods, such as computational modelling, artificial language learning, communication games and traditional psycholinguistic experiments. In many studies, an important role is played by information theory (cf. Gibson et al. 2019).

All this wealth of ideas and evidence requires systematization and explanation, as well as some critical re-evaluation. In particular, the following questions require an answer:

- What are the different costs and benefits in language communication?
- What efficient linguistic strategies are there?
- What are the pragmatic and cognitive mechanisms of efficient linguistic behaviour for the speaker and the addressee?
- How do efficient conventionalized linguistic form–meaning pairings develop?

This book addresses these questions and provides many examples of efficient linguistic structures and patterns of use. Note that we will only speak here about communicative efficiency, that is, minimization of the cost-to-benefit ratio in language use, and leave out other possible types of efficiency in language (e.g., learning efficiency).

1.2 Benefits and Costs in Communication

1.2.1 Types of Benefits

If efficiency is minimization of a costs-to-benefits ratio, what are the costs and benefits of using language? We will begin with benefits. Surprisingly, they are rarely discussed in the literature on communicative efficiency.

Speaking very generally, the ultimate goal of all our activities as an organism is survival. For this purpose, we need to collaborate with some people and compete with others. This involves influencing other people, so that they give us some material goods, help us, attack our rivals or simply leave us alone. We also benefit from useful information that we request and obtain because it helps us to adjust our behaviour and adapt to the environment better. These are the benefits of communication in a very broad sense.

Following Relevance Theory (Sperber and Wilson 1995; Wilson and Sperber 2004), we can also speak of benefits as positive cognitive (or contextual) effects for the addressee. Positive cognitive effects are worthwhile differences between the old (before communication) and new (after communication) representation of the world. They represent new conclusions based on the utterance and context, but also strengthening, revisions and abandonment of already available assumptions. Cognitive effects are similar to updating of prior beliefs in Bayesian inference. Human cognition is geared towards maximizing cognitive effects (Sperber and Wilson 1995). The changes in beliefs correspond to diverse cognitive processes in the addressee: learning new information or confirming previous beliefs about the world, bonding with the speaker, deciding to perform an action, empathizing with the speaker, enjoying the style or accepting new linguistic conventions. These diverse processes illustrate Jakobson's referential, phatic, conative, emotive, poetic and

metalingual functions of language use (Jakobson 1971 [1960]). Importantly, cognitive effects and the resulting processes represent benefits not only for the addressee but also for the speaker, who is interested in evoking them.

In order to evoke desired cognitive effects in the addressee, the speaker needs to ensure that the linguistic units and their functions (that is, lexical meanings, grammatical categories, roles and other information) are transferred more or less successfully. Using Relevance-Theoretic parlance, we can say that successful communication requires a recovery of what is explicitly said, or explicatures. This information is obtained by a combination of decoding and inference, with the help of such operations as reference resolution, semantic narrowing, loosening, speech act identification and others. Explicatures form the basis for recovery of implicated premises and conclusions, which represent cognitive effects for the addressee.

Of course, this is an idealization. We do not always recover all units and all meanings; nor do we need to. First of all, communication happens in a noisy channel, using Shannon's terminology (1948). Faithful transfer of linguistic units can fail due to physical impediments (e.g., speaking in a crowded pub) or processing difficulties (e.g., see F. Ferreira [2003] on 'good-enough' processing of sentences). Our language seems to be protected against noise by redundancy (cf. Hengeveld and Leufkens 2018), which means that not all units must be transferred perfectly. At the same time, it is obvious that linguistic units must be of some use for communicators. If we speak about the grammatical function of a case marker, for instance, we assume that this meaning helps the addressee to understand who did what to whom, even if this information can be partly inferred from other linguistic cues (e.g., lexical or semantic properties of the arguments). The working hypothesis is that human languages develop and retain conventionalized cues because these cues are normally useful for evoking cognitive effects.

The benefits, from more specific to very general ones, are displayed in Figure 1.1. We will assume that in most cases the transfer of linguistic units is successful, helping the addressee to obtain intended cognitive effects and adjust their own behaviour, as a result. From the speaker's perspective, triggering desirable cognitive effects in the addressee helps to influence the addressee's behaviour in a useful way. Finally, influencing other people's behaviour or adjusting one's own increases the chances of the language user's survival as a living organism.

1.2.2 Types of Costs

Communication costs have received more attention than benefits in the literature. They can be classified into several types, as shown in Figure 1.2. First of all, we can speak about costs related to the effort involved in communication. Two major types are processing and articulation (including sign languages) or

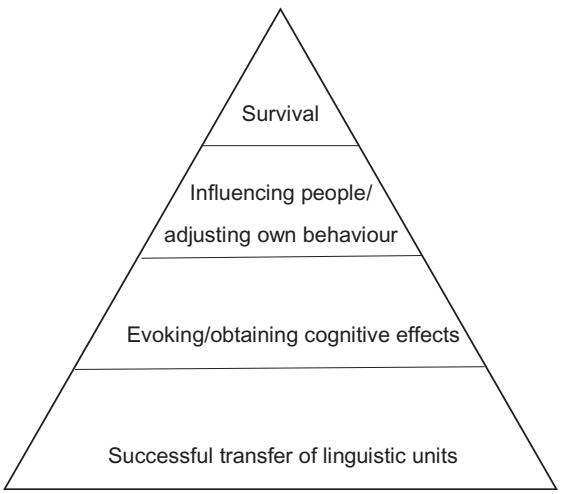

Figure 1.1 A hierarchy of benefits in linguistic communication

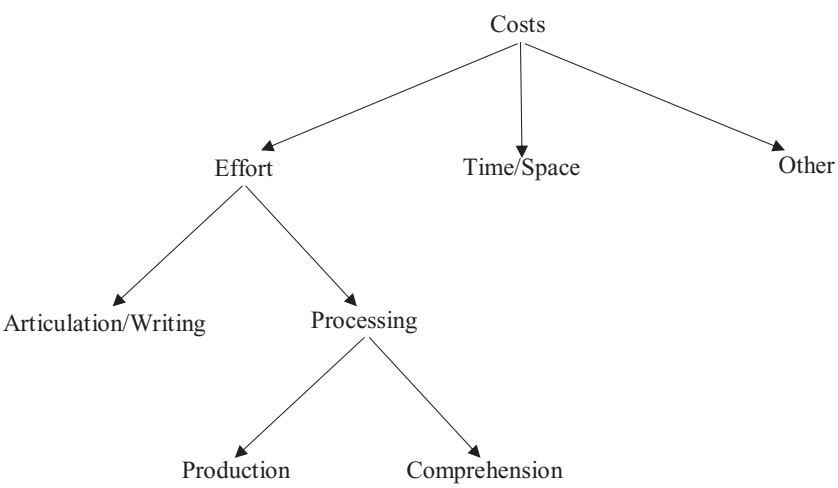

Figure 1.2 Different types of costs in linguistic communication

writing. Processing costs are associated with different cognitive processes required for language comprehension and production.

Time (or space in writing) is another type of cost. According to V. Ferreira (2008), speakers have the responsibility not only to say things their addressees can understand, but also to say things quickly. Similarly, Clark's (1996) 'temporal imperative' says that speakers need to use time in the conversations

wisely and responsibly. Since articulation takes time, these costs usually go together. However, there are situations in which time has an independent value. An important aspect of efficient use of time has to do with word order. Speakers tend to produce first the constituents that are more accessible. Accessibility is influenced by a range of factors, including frequency, givenness and animacy. By producing accessible material first, the speaker buys time for planning less accessible units (see more in Section 3.2.2).

Most studies of efficiency focus on the amount of effort and time, but other costs can be important, too. For example, poor communication can have severe social consequences, including loss of face, ruined reputation and broken relationships. Politicians know this all too well. For example, when the current US President Joe Biden once said, *I'm Irish but not stupid*, many Irish people were not amused. Why? The use of *but* signals that the speaker thinks that both he and the addressee are familiar with the cultural stereotype that Irish people are generally stupid. From that, it is easy to conclude that Biden actually thinks that his audience shares the stereotypical belief that the Irish are stupid.

It is difficult to measure social costs, but there are ways of quantifying the degree of miscommunication with the help of information theory. For example, Kemp, Xu and Regier (2018) operationalize what they call communicative costs as the difference between the speaker's and the addressee's probability distributions over referents that can be represented by a certain referential expression. See more on this approach in Chapter 6.

Social costs are closely related to effort. Misunderstanding can lead to additional articulation costs and loss of time, from a simple repair in a dialogue to extensive explanations and press releases. Articulatory and social costs can also be in conflict, as one can see in the current debate about the use of feminitives in German. In particular, the masculine plural form of nouns referring to human beings is considered ambiguous in the sense that it is not clear whether it names men only or both men and women. For example, *die Kollegen* 'the colleagues' and *die Lehrer* 'the teachers' can be interpreted in both ways. In order to be gender-inclusive, avoiding the male-only interpretation, it is considered appropriate by many people to use these forms along with the plural feminine forms, e.g., *Kolleginnen und Kollegen* 'colleagues (female) and colleagues (male)'. This can lead to very long forms, especially if there are attributes. For example, in a job advertisement, one would write something like this:

(4) *Wir suchen ein-e erfahren-e Buchhalter-in*
We search ART-F.ACC experienced-F.SG.ACC accountant-F
/ ein-en erfahren-en Buchhalter.
ART-M.ACC experienced-M.ACC accountant
'We are looking for an experienced accountant (male or female).'

In this example, the costs of writing and space are particularly high.