20

Translation, information theory and cognition

Elke Teich, José Martínez Martínez and Alina Karakanta

20.1 Introduction

It is widely acknowledged that human language processing relies to a considerable degree on expectancy. In online processing, we make predictions about what comes next, and we choose particular linguistic encodings based on preceding events in the situation, the preceding discourse or assumptions about the addressee (e.g., shared knowledge or state of attention) (see Levy, 2008; Levy & Jaeger, 2007). Similarly, in offline (written) language use, we rely on the expectancy of particular linguistic encodings according to register, text type and/or genre.

The formal basis for modelling predictability in context (linguistic, situational or register) is provided by probabilistic models of language use—so-called language models, which are widely used in natural language processing. While it has been suggested that probabilistic explanations should provide valuable insights into translators’ behaviour and translational choice (Toury, 2004), in Translation Studies only a few attempts have been made to develop suitable formal approaches. Some earlier works are Levý’s account based on game theory (Levý, 1967; see Cronin, 1998) and Gutt’s account (Gutt, 1989) based on relevance theory (Sperber & Wilson, 1986).

The purpose of the present chapter is to sketch a formal basis for the probabilistic modelling of human translation that may provide a common basis for product-oriented and process-oriented research on translation and makes communicative explanations of translational choice and its linguistic traces possible: shining through, normalization, simplification, etc. (see Baker, 1996; Teich, 2003; Toury, 1995; Volansky et al., 2015). The approach is rooted in information theory, with Shannon’s notion of information at the core (Shannon, 1948). We provide a definition of Shannon information applied to linguistic communication—called surprisal in research on comprehension and information density in production research—and discuss its relevance for modelling translation (Section 20.2.1).

According to Shannon, a crucial property of the communication channel is that it is noisy, i.e. the signal is always distorted, which may result in loss of information through transmission. We explain the concept of the noisy channel and the types of linguistic problems that are typically addressed with noisy channel models—including machine translation, which provides the link to modelling human translational choice (Section 20.2.2). Thirdly, we suggest that a number of
translation-relevant variables, notably (dis)similarity between languages, level of expertise and translation mode (i.e. interpreting vs. translation), may be appropriately indexed by entropy, which in turn has been shown to indicate production effort (Section 20.2.3). What is crucial about an information-theoretic approach is that it provides a link to cognition: Surprisal has been shown to correlate with a number of behavioural and neurophysiological indices (e.g. response latencies, reading time, pupil dilation and event-related potentials (ERP); see Crocker et al., 2016) and is therefore a suitable measure of cognitive effort in human language processing. Section 20.3 concludes with a brief summary, a discussion of methods to estimate probabilities from corpora (Section 20.3.1) and an outlook on rational explanations of translation (Section 20.3.2).

**20.2 Core concepts**

**20.2.1 Shannon information and surprisal/information density**

In human language processing, context exerts various constraints upon the kinds of linguistic unit that may come up. For example, at the level of syntax, given a preposition, the unit most likely to follow is a nominal phrase; or, at the level of words, given the word “read”, a likely continuation is “book”. Information theory allows us to define such probabilities in context on the basis of the amount of information that is conveyed by a given unit measured in bits. This is commonly formalized by the measure of surprisal (Equation 20.1), which estimates the probability of a *unit* (e.g. a word) given some *context* (e.g. the preceding *n* words) as the negative logarithm to the base 2 (alternatively, base 10).

\[
\text{Surprisal} = - \log_2 p(\text{unit} | \text{context})
\] (20.1)

According to this model, linguistic events with high surprisal are low in probability and convey more information than events with low surprisal/high predictability in context. Crucially, predictability in context is inversely proportional to cognitive effort; i.e. higher surprisal incurs higher processing cost (Hale, 2001).
For illustration, consider the following two examples:

(1) Jane read a book.
(2) Jane bought a book.

In example (1), the item “read” in the preceding context of “book” makes “book” highly predictable (there aren’t many other, similarly likely alternative completions); “book” has low information content, so surprisal is low. In example (2), in contrast, “buy” does not strongly license a particular continuation, so surprisal on “book” is relatively high: we get more information when we see/hear “book” in the context of “buy” than in the context of “read”. Figure 20.1 illustrates why “book” needs more bits for encoding in (2) compared with (1), and so it incurs a higher processing cost.

For estimating probabilities in context, we need linguistic usage data. These can be obtained from corpora (corpus probabilities; see Section 20.2.2). While this involves some challenges (see Section 20.3.1 for discussion), in the following we focus on the potential of an information-theoretic approach for modelling human translational choice and contributing to the theory of translation at large.

20.2.1.1 Applications in linguistics

The perspective of predictability in context is extremely fruitful for the study of language use, variation and change. Apart from language comprehension, in language production it has been shown that shorter, more condensed linguistic variants are preferred in more predictive contexts and longer, more expanded variants when the context is less predictive. Examples are expanded/reduced vowel space size in speech (Schulz et al., 2016), fragments in syntax (Reich, 2017), condensed syntactic expression (e.g., coercion: Jane began a book (Delogu et al., 2017)) or optional marking of discourse relations (Rutherford et al., 2017). Thus, language users seem to strive for an optimal encoding of a given message depending on predictability in context by modulating the amount and rate of information in a message through specific linguistic choices. This seems to be a valid communicative explanation of certain types of linguistic variation to be observed in online communication, notably the choice of fully expanded vs. condensed linguistic forms such as relative pronoun or complementizer omission, syntactic fragments or shorter syllable durations. In the analysis of written, offline communication, surprisal can act as a measure of the (relative) complexity of a text as an alternative to type-token ratio, lexical density or the Fog index, which are based on simple frequency counts. Surprisal has the added value of being context aware and, more importantly, directly cognitively relevant. In a theoretical perspective, adopting an information-theoretic approach opens up the opportunity to explain language use on the basis of rational communication, according to which interlocutors want their interactions to be successful, as seen in Grice’s (1975) maxims and in Sperber and Wilson’s (1986) relevance theory, while at the same time keeping their cognitive effort at a reasonable level. There is now increasing empirical evidence that communicative concerns play an important role in language variation and change at large, as seen by Hume and Mailhot (2013), Degaetano-Ortlieb and Teich (2016), Rubino et al. (2016) or Baayen et al. (2017). The structure and evolution of the linguistic system may thus be explainable in terms of communicative concerns, striking a balance between expressiveness and efficiency (see e.g. Plantadosi et al. (2011) on the optimization of word lengths across languages).

20.2.1.2 Application to translation

The perspective of rational communication can be straightforwardly applied to translation. It is reasonable to assume that translators, too, want communication to be successful, and, as far as
professional translation goes, they strive for a high-quality output. However, there are a number of specific constraints interfering with these goals. First, translators need to produce a translation that is true to the source-language text and conforms to the target-language expectations at the same time—the classic translation dilemma. Furthermore, compared with other linguistic processes (e.g. reading or shadowing), translation is a process with high resource limitations (time pressure and cognitive effort) (Hyönä et al., 1995), and it can therefore be assumed that translators, and even more so interpreters, have a vested interest in keeping their effort at a reasonable level. This includes efficient management of working memory, e.g. by maintaining only few hypotheses about a translation solution and trying to be as certain as possible about the best solution (Pym, 2008). Further interacting constraints are level of expertise (professional vs. learner), translation mode (translation vs. interpreting), language pair (i.e. (dis)similarity between source and target language) and translation direction (i.e. from/into native vs. non-native language).

Notwithstanding its specific constraints, translation can be modelled in probabilistic terms just like other processes of language use. Taking the production perspective, in the remainder of this chapter we will focus on how to model the following goals of translation and the linguistic traces that may be left by striving to realize these goals (see translationese: Baker, 1995; Gellerstam, 1986):

1. Be as true as possible to the source-language message.
2. Adhere to the norms of the target language.

In product-oriented Translation Studies, the first relates to the notion of equivalence, which defines a relation between two products, a source-language text and a target-language text; the second corresponds to the notion of adequacy or appropriateness (see e.g. Reiss, 1983). Thirdly, a condition rather than a goal, it is in the interest of translators to keep their efforts (cognitive and temporal as well as technical) at a reasonable level.

We assume that the translation output is optimal when goals (1) and (2) are reached, but typically they are compromised due to effort levelling and the interacting constraints discussed earlier. Also, one goal may be favoured over the other, possibly resulting in shining through/interference (overemphasizing goal 1) or normalization/standardization (over-emphasizing goal 2).

Turning to probabilistic modelling, the successful outcome of goal 1 may be characterized as maximizing the probability of being able to retrieve the source-language expression given the chosen target-language translation; and goal 2 can be formalized in probabilistic terms as maximizing the probability of the chosen translation in the target-language context. To capture the gist of translation, we need to be able to model these two goals together. As will be explained in Section 20.2.2, this is exactly what the noisy channel model allows us to do.

Finally, regarding cognitive effort, we analyse the number and distribution of translation options considered at a given choice point as an important factor. Here, the hypothesis is that the more (equiprobable) options are being considered at a given choice point, the higher the cognitive effort involved. As will be explained in Section 20.2.3, the complexity of a choice (and associated cognitive effort) can be appropriately modelled by entropy.

### 20.2.2 Noisy channel

The noisy channel model of communication gives recognition to the fact that the transmission of a message is typically distorted. For example, an original input “first name” may be
phonetically distorted in such a way that what is understood is “first time”, or a given word may be orthographically distorted by misspelling, e.g. “embarrass” as “embarass”.

The noisy channel model is applied successfully in a number of natural language processing tasks, including spelling and optical character recognition (OCR) correction, speech recognition and machine translation. In these tasks, the common goal is to restore the original input from the distorted output of a noisy channel. Formally, this is represented as

$$\text{argmax}_c p(c|x) = \text{argmax}_c p(x|c) p(c)$$

(20.2)

for all candidate matches $c$, where $p(x|c)$ is called the error model (how likely is it that $x$ is a variant of $c$?) and $p(c)$ the language model (how likely is $c$ in a given language?).

For the sake of illustration, consider an example from spelling error correction (see Figure 20.2; based on Jurafsky & Martin, 2018, p. 480). A sender sends a word through the noisy channel (e.g. “embarrass”) which is distorted to a noisy word by misspelling (e.g. “embarass”, “emberrass”, “emberass”), and the receiver needs to “guess” what the original word was. In a noisy channel model, this guessing is done using knowledge about the original input $c$ as well as knowledge about possible distortions in the channel. When receiving the word “emberrass”, the receiver assesses how likely “a” is instead of “e” and how likely the word “embarrass” is. In many applications, this search process is performed with the help of a device called a decoder, which gathers candidate original words for the noisy word and decides on the basis of the underlying probabilistic model what the best match is.

To explain what the noisy channel model can do for us in modelling human translation, it is instructive to look at its application in machine translation.

### 20.2.2.1 Application in machine translation

The noisy channel is the underlying formal model in statistical machine translation. According to such a model, a translation itself is a distorted message; e.g. an intended message in a language $e$ is output by an expression in another language $f$ (see Figure 20.3). More technically, in statistical machine translation a text is translated according to a probability distribution $p(e|f)$, where $e$ is an expression in a target language and $f$ is an expression in a source language. For modelling, $e$ and $f$ are reversed to $p(f|e)$ (the latter is easier to estimate). Additionally, the probability of the expression $e$ on its own (i.e. in original texts of target language $e$) is taken into account. $p(e)$ is the so-called translation model (= error model of the noisy channel) and $p(e)$ the language model (see Equation 20.3).
The ultimate goal in machine translation is, then, to find the optimal translation \( e \), which maximizes the two probabilities:

\[
\text{argmax}_{e} \ p(e) \cdot p(f|e)
\]

(20.3)

where the \text{argmax} operation is a search process in the space of possible target translations—this is the decoding procedure as explained earlier for the task of spelling correction. Challenges involved in statistical MT are to come up with efficient approaches to decoding, to find the best translation unit granularity (word based vs. phrase based) and to determine the optimal parameter weights for a given translation task (e.g. giving more weight to \( p(f|e) \) and less to \( p(e) \) or vice versa). Also, the output quality of a statistical MT system rests very much on the size and quality of parallel corpora (see Section 20.3.1 for discussion).

### 20.2.2.2 Application to human translation

We can now conceptualize the components of the noisy channel model for machine translation in terms of human translation. The translation model and the language model can be used directly for representing the goals of human translation discussed earlier. Humans try to reconcile goals 1 and 2—in terms of a noisy channel model, they seek the optimal balance between \( p(e|f) \) and \( p(e) \). Goal 1, henceforth referred to as source-language (SL) fidelity, is represented by the translation model, i.e. maximizing the probability of retrieving the source-language expression \( s \) given the chosen target-language translation \( t \). The type of data needed for modelling here is parallel corpora. In a statistical MT system, the probabilities based on the counts retrieved from the parallel corpus are represented in a so-called phrase table. For an example, see an extract of a phrase table in Figure 20.4, which lists the translations occurring in the underlying parallel corpus for the English source-language expression “prerequisite” into German with their translation probabilities.

According to the example, the best match for “prerequisite” would be “Voraussetzung” (\( p(e|t) = 0.516 \)).

Goal 2, henceforth referred to as target-language (TL) conformity, is represented by the target-language model, i.e. maximizing the probability of the target-language expression on its own. Here, we
need a monolingual corpus of the target language that is ideally as comparable as possible to the parallel
corpus (register, domain). The language model is a measure of a well-formed expression in the target
language. Moreover, it aids the translation model in difficult decisions by providing knowledge about
the context of the expression (preceding words). For example, the translation model gives the highest
probability to “Voraussetzung” as a translation of “prerequisite”. However, depending on the context,
the best match could be the option with the second highest probability, “Grundvoraussetzung”. In
this way, the language model is responsible for ensuring fluent output. See Section 20.3.1 for further
discussion on giving more/less weight to language models or translation models.

Provided we are able to obtain empirically sound models, we may use them in a wide spectrum
of relevant applications, ranging from translation quality assessment and translationese detection to
compiling material for translation training and obtaining stimuli for experimental translation process
research. For instance, in a given translation, we can assess whether it is more on the literal or on the
adaptive end of the translation cline, i.e. emphasizing SL or TL. For example, in the EuroParl-UdS corpus
we find two alternative translations for the English “development budget”, “Entwicklungsbudget” and
“Entwicklungshaushalt”, and the preferred translation is “Entwicklungsbudget”. However, in compara-
ble original texts, “Entwicklungshaushalt” is clearly preferred. Choosing “Entwicklungsbudget”
is thus an obvious case of literal translation, overemphasizing SL fidelity at the cost of TL conformity.
Considering accumulated effects, noisy channel analysis provides a diagnostic tool for translationese: a
tendency to favour high probabilities for \( s|t \) is interpreted as SL shining through, while a tendency
to favour high probabilities for \( t \) is indexical of TL normalization.

Furthermore, other methods from the same formal family of information-theoretic methods
can be naturally connected. For instance, a translation’s difference from a comparable target-
language text (or text collection/corpus) can be measured with relative entropy (e.g. by Jensen–
Shannon divergence or Kullback–Leibler divergence). Relative entropy is used as standard to
compare probability distributions in terms of the number of additional bits needed for encoding
when a non-optimal model is used (see Fankhauser et al. (2014) and Degaetano-Ortlieb & Teich
(2016) for applications in contrastive and diachronic analysis). For an example, see Figure 20.5
showing word clouds of the most distinctive words for English translation vs. interpreting based
on Kullback–Leibler divergence (KLD), (a) from German and (b) from Spanish. Size encodes
the information gain (additional bits). In this example, we can observe, for instance, that the most
distinctive features of interpreting (right) are markers of oral mode, such as interactant personal
pronouns, contractions and grammatical coordination, whereas translation (left) exhibits clear

| SL                  | TL                  | \( (s|t) \) |
|---------------------|---------------------|------------|
| prerequisite        | Bedingung           | 0.002433   |
| prerequisite        | Grundbedingung      | 0.002433   |
| prerequisite        | Grundvoraussetzung  | 0.096603   |
| prerequisite        | Voraussetzung dafür | 0.002433   |
| prerequisite        | Voraussetzung       | 0.51612    |

Figure 20.4 Extract from a phrase table
signs of written mode, such as topical words and the definiteness marker “the”. Note that the difference between translation and interpreting outputs seems to be independent of the source language, as indicated by the example in Figure 20.5 (see Shlesinger & Ordan, 2012 for similar observations).

20.2.3 Entropy

Entropy quantifies the amount of uncertainty related to the outcome of an event. For illustration, let us consider a simple example. There are three bowls with four apples in each of them: in bowl A there are four green apples, bowl B contains one red and three green apples, and bowl C has two green and two red apples. With bowl A, we can be 100% certain that we get a green apple; with bowl B, there is a 75% chance that the apple we get is green and a 25% chance that it is red;
whereas with bowl C, green and red are equally probable (50%). Comparing the three cases, in A no uncertainty about the outcome is involved—therefore entropy is zero. C has the highest uncertainty, with both outcomes being equally likely, and entropy is 1. B is in between (0.8).

Technically, entropy measures the size of the search space of possible values of a random variable \( X \) and its associated probabilities (Manning & Schütze, 2003, p. 63). The mathematical formulation of entropy \( H \) is given in Equation 20.5, where \( p(x) \) is the probability mass function of the random variable \( X \) (see preceding example: kind of apple) over all its possible values \( x \in X \) (see example: green or red apple).

\[
H(X) = \sum_{x \in X} p(x) \log_2 p(x)
\] (20.5)

Entropy, like surprisal, is measured in bits. When we are absolutely certain about the outcome, entropy is zero bits (as in A in the apple example). In contrast, the most uncertain situation, with the highest entropy possible, occurs when all outcomes are equiprobable (as in C). Then entropy equals \( \log_2 n \) (where \( n \) is the total number of possible outcomes). Like surprisal, the measure of entropy will always yield a positive number, with values ranging from 0 to \( \log_2 n \).

### 20.2.3.1 Application in linguistics

Entropy has been applied in linguistics to capture the complexity of linguistic choices. For instance, it has been shown that there is a relation between entropy and cognitive effort in language comprehension, which may be described as uncertainty on the part of readers or listeners about what comes next (e.g. a word). In sentence comprehension, entropy (uncertainty) typically decreases with each incoming word, and the amount of information that each word gives may be defined as the reduction in entropy due to that word (see Blache & Rauzy, 2011; Frank, 2010; Hale, 2001, 2003, 2006). To mention a more specific example, Linzen and Jaeger (2016) look at probabilities of complementation patterns of verbs with multiple subcategorization options (e.g., “He forgot my birthday”, “He forgot about my birthday”, “He forgot that it was my birthday”). When probability is evenly distributed across subcategorization frames, verbs with more frames have higher entropy; for the same number of frames, the less balanced the distribution, the lower the entropy. For processing, the prediction is that the first is more effortful than the second.

Also, language production has been shown to work with predictability (e.g. Jaeger, 2010). Here, entropy in the search space of linguistic options is an important indicator of processing complexity: as a tendency, the higher the entropy of the space of options, the higher the effort incurred in processing (measured e.g. by response latency or production time). In fact, this partly explains longer latencies in word search in older adults—due to lifelong experience, the space of options is simply much larger than in younger people. Hence, rather than general cognitive decline, a higher entropy in the lexicon may better explain word search problems in older people in some tasks (see e.g. Blanco et al. (2016) for relevant results from behavioural experiments).

### 20.2.3.2 Applications to human translation

If we define the task of translation as a search task in a space of alternative linguistic options in a target language (see \( p(s|t) \) in the noisy channel model), we can apply entropy directly to formalize the set of possible translation outputs as:

\[
H(T) = -\sum_{t \in T} p(t) \log_2 p(t)
\] (20.6)
where \( T \) stands for the translation space, i.e. the set of all possible translations \( t \) for a given source-text unit as found in a parallel corpus.

For illustration, consider two examples from an excerpt of the proceedings of the European Parliament in English translated into Spanish by translation trainees in Figures 20.6 and 20.7, showing two translation spaces with low and high entropy, respectively (see Martínez Martínez & Teich, 2017). The x-axis shows the individual options and the y-axis plots surprisal, i.e. how probable a translation is (low surprisal = high probability). Figure 20.6 illustrates the translation space of “Council” showing a very low entropy due to few options (five), with one option being the clearly preferred translation (“Consejo”). We may infer that the translation of “council” into Spanish is quite straightforward (on the basis of the underlying parallel corpus) and not associated with high cognitive effort. In the other example (Figure 20.7), in contrast, there are many alternative options (35), and they show fairly similar surprisal scores; i.e. they are all similarly likely. Again, given that entropy is positively correlated with cognitive effort, it may be concluded that translational choice from English to Spanish regarding “gross breach” is associated with relatively high cognitive effort, as indicated by the underlying parallel corpus, while the translation of “Council” is associated with low effort.

Estimates of entropy based on parallel corpora may be directly used to assess translation difficulty from the production perspective, connecting up with existing work on translation in a natural way. The number of variants available to and considered by the translator is mentioned as an indicator of cognitive effort by both Krings (2001, pp. 536–537) and Englund Dimitrova (2005, p. 26). See also Angelone’s work on uncertainty management in translation associated with problem solving (Angelone, 2010); or Campbell’s choice network analysis, which confirms a correlation between the number of choices and cognitive effort (Campbell, 1999, 2001).

One of the most elaborate approaches using entropy in translation process research is word translation entropy, as initially proposed by Schaeffer and Carl (2014) and taken further in subsequent research as in Carl and Schaeffer (2017a, 2017b), Schaeffer et al. (2016) and Bangalore et al. (2016). This work emanates from exploiting data from the CRITT translation process database (Carl et al., 2016). Also, there are related tasks in lexical natural language processing similar to
choice in translation, such as ambiguity detection based on semantic similarity between source and target text (see e.g. Cap, 2017; Villada Moirón & Tiedemann, 2006), that are relevant to consider for modelling human translation.

Other applications of entropy in Translation Studies include assessment of a given translation (or set of translations) in terms of degree of literalness or degree of expertise. For instance, Carl and Schaeffer (2017a) show that more literal translations are easier to produce than less literal translations, linking back to evidence from priming studies. Similar effects are shown by Bangalore et al. (2016) for syntactic choices (e.g. word order). Martínez Martínez and Teich (2017) show how entropy can be used to characterize learner vs. expert behaviour in translation, interpreting low entropy as an indicator of routine behaviour (high certainty about translational choice).

Entropy thus offers the possibility to capture effects of some important factors involved in translational choice, such as source- and target-language similarity or level of expertise, and it may be used for assessing translation output in terms of translation consistency (e.g. when multiple translators are involved). As a measure, entropy turns out to be reasonably stable across languages and modes of translation (e.g. from scratch vs. post-edited), thus providing a reliable instrument for comparative studies. Crucially, as we have shown with selected examples from online language processing, entropy provides a direct link to cognitive effort—the higher the entropy at a given choice point, the higher the effort incurred, as indicated by several behavioural measures (e.g. response latency and production time).

### 20.3 Discussion and future directions

In this chapter, we have sketched a formal basis for the probabilistic modelling of human translation based on information theory, adopting Shannon’s noisy channel model and selected entropy-based measures of information. The new perspective that is opened up by applying information theory in Translation Studies is to consider translation as rational communication, according to which interlocutors aim to understand the sender’s message and to
be understood by the receiver while the cognitive effort spent remains reasonable. However rational language users may actually be in real life, it is clear that many constraints interact with interlocutors’ assumed rational goals, including noise in the channel, uncertainty about the correct interpretation of a message or the adequacy of a produced message for a given audience, as well as cognitive resource limitations. Translation being a type of language use, albeit a special one, by implication operates under the same kinds of constraints. The specific goals of an interlocutor-as-translator impose additional, translation-specific constraints, notably to find a good balance between fidelity to the SL expression and conformity to TL expectations (see Chapter 2.4 on the notion of equivalence). We have shown how these objectives can be formally represented by the components of a noisy channel model and suggested a number of uses for such models (Section 20.2.2). Furthermore, we have shown that selected constraints (level of expertise and translation difficulty) may be captured by entropy as a suitable operationalization of the notion of (un)certainty about the best translational choice (Section 20.2.3).

The common assumption in applications of information theory to the study of language is that language use can be modelled as a probabilistic process, according to which interlocutors rely to a large extent on predictability in context (see Section 20.2.1). While probabilistic models of language use have been successfully applied in a number of natural language processing tasks, applications to linguistic or translatological questions are still rare.

Embarking on this line of research involves some methodological challenges (see Section 20.3.1) but bears a lot of promise for theoretical advancement (see Section 20.3.2).

### 20.3.1 Methods: Estimating probabilities

As in other data-driven, corpus-based accounts, there are a number of issues involved in estimating probabilities, such as representativeness of the data set, data sparseness, difference in size of data sets when comparing e.g. parallel and comparable monolingual corpora, as well as reliability of (pre-)processing, notably word alignment. Also, to enable rational explanations including cognitive interpretations, we want to integrate contextual constraints into our models (see surprisal, Section 20.2.1). Apart from the ambient linguistic context, which is known to have an immediate effect on online processing effort, there are many other relevant contextual variables that may exert an influence on the expectancy of a given linguistic unit, including discourse context, register/text type and world knowledge. With the approach sketched here, we can potentially incorporate any kind of contextual variable.

To start with, for approximating $p(\text{unit} | \text{context})$, standard n-gram language models can be used, where context is simply the preceding context of $n-1$ words. Other variables relevant for translation are brought into play by choosing specific corpora, e.g. interpreting vs. translation, different language pairs and translation directions, or levels of translation expertise. As in other data-driven approaches, very much rests upon the choice of adequate data sets. For a noisy channel model of translation, we need substantial amounts of translation data to estimate $p(s | t)$ as well as target-language data to estimate $p(t)$. A specific challenge is that the probability scores for the translation model and the language model will be in quite different ranges and not directly comparable. Comparing the rankings of the translation options based on probabilities of the translation model and of the language model can provide us with information on whether a given translation choice tends towards SL fidelity or TL conformity. If we want the two components to have equal influence on the output, we can experiment with different weights on the two components of the noisy channel model; i.e. we can induce greater TL conformity by giving more weight to the target-language model or greater SL fidelity by putting a larger
weight on the translation model. In this way, we can explore the effects on translational choice of shifting between SL fidelity and TL conformity. See again the example of “Entwicklungsbudget” (higher probability according to translation model) vs. “Entwicklungshaushalt” (higher probability according to target-language model).

### 20.3.2 Theory: Translation as rational communication

Casting translation in terms of rational communication with information theory as a formal basis allows us to formalize some core notions of translation theory as well as provide the theoretical underpinnings for some less explained areas in Translation Studies. Theoretical notions such as equivalence (see Chapter 19, this volume), creativity (see Chapter 17, this volume), expertise (see Chapter 26, this volume) or competence (see Chapter 22, this volume) are all graded concepts and may adequately be represented in probabilistic terms. Furthermore, context is regarded as a crucial factor in translation, but it often remains too vaguely defined. In an information-based model, context is given in the very definition of information as the immediate linguistic context, register context and/or source- or target-language context (see discussion on methods in Section 20.3.1). Here again, probabilities are an adequate means for modelling the influence of context on translational choice. Finally, a number of translational phenomena that are hard to explain may be captured within a rational communicative framework. For instance, omission is a common choice in translation, and even more so in interpreting, but for a given occurrence it is often unclear when the choice is due to audience design (TL orientation) or to cognitive resource limitations on the part of the translator/interpreter. Here, the factors assumed to be involved in a given translational choice may be represented as multiple independent variables in a regression model.

In summary, while adopting an information-theoretic framework for modelling human translation presents some challenges, it bears the promise of integrating insights from descriptive corpus analysis and experimental results on selected cognitive processing aspects involved in translation for a more encompassing explanation of translation under the perspective of rational communication. The goals of rational interlocutors—successful message transmission with reasonable cognitive effort—are goals for translators, too. The formal apparatus we have described in this chapter provides the means to model rational communicative goals and the constraints acting upon them, including the specific constraints acting upon translation. With its link to evidence from cognitive processing, an information-based, rational communication approach provides an adequate level to formulate translation universals and a unifying framework for capturing translationese effects (see also Halverson, 2003 on the cognitive basis of translation universals). Finally, a rational communicative approach to translation opens up the opportunity of closer interaction with relevant linguistic research on language use, variation and change.

### Notes

1 Formally, surprisal and information density are identical, but the term *surprisal* tends to be used in the context of language comprehension, while information density tends to be used in studies of language production. Commonly employed related measures are *Mutual Information*, a popular measure for characterizing collocations, and *Information Gain*, commonly used for comparing probability distributions.

2 The corpora used for illustration here and in other examples are EuroParl-UdS (Karakanta et al., 2018), available from CLARIN-D at http://fedora.clarin-d.uni-saarland.de/europarl-uds/, and the Translation and Interpreting Corpus (TIC) (Kajzer-Wietrzny, 2015).

3 We could have used frequency here instead, i.e. how often a translation is used.
Further reading


In his pragmatic theory, Grice introduced four principles of linguistic interaction that describe the rational principles that people follow in effective communication. The four “maxims” of quality, quantity, relation (or relevance) and manner are the pillars of an overarching cooperative principle that Grice stipulated.


The book introduces a formal theory of language on the basis of information theory. It includes an information-theoretically oriented account of register variation using the example of scientific language and offering a communicative explanation for register/sublanguage formation.


This is a standard textbook in natural language processing (NLP). Different chapters introduce NLP techniques mentioned in the present article, including various kinds of computational language models.


This is a standard textbook in statistical machine translation (SMT). Different chapters introduce the basic workings of language models in the context of translation, including word-based and phrase-based models, text–translation alignment and evaluation of MT output.


This article from the early 1950s provided an asymmetric measure of relative entropy, called Kullback–Leibler Divergence after the authors.


This is the original article by Claude E. Shannon on information theory, including the definition of general communication systems by a noisy channel model.

References


