The boundaries of meaning: a case study in neural machine translation[*]

Yuri Balashov[†]

Department of Philosophy, University of Georgia, Athens, GA 30602, USA

yuri@uga.edu

ORCID ID: 0000-0001-7369-2122

September 12, 2022

Abstract

The success of deep learning in natural language processing raises intriguing questions about the nature of linguistic meaning and ways in which it can be processed by natural and artificial systems. One such question has to do with subword segmentation algorithms widely employed in language modeling, machine translation, and other tasks since 2016. These algorithms often cut words into semantically opaque pieces, such as 'period', 'on', 't', and 'ist' in 'period|on|t|ist'. The system then represents the resulting segments in a dense vector space, which is expected to model grammatical relations among them. This representation may in turn be used to map 'period|on|t|ist' (English) to 'par|od|ont|iste' (French). Thus, instead of being modeled at the lexical level, translation is reformulated more generally as the task of learning the best bilingual mapping between the sequences of subword segments of two languages; and sometimes even between pure character sequences: 'p|e|r|i|o|d|o|n|t|i|s|t|' → 'p|a|r|o|d|o|n|t|i|s|t|e'. Such subword segmentations and alignments are at work in highly efficient end-to-end machine translation systems, despite their allegedly opaque nature. The computational value of such processes is unquestionable. But do they have any linguistic or philosophical plausibility? I attempt to cast light on this question by reviewing the relevant details of the subword segmentation algorithms and by relating them to important philosophical and linguistic debates, in the spirit of making artificial intelligence more transparent and explainable.

Keywords: Opacity; Deep learning; Computational Linguistics; Neural machine translation; Subword segmentation

[*] Published in *Inquiry*. doi:10.1080/0020174X.2022.2113429

[†] I wish to thank the referees for the very helpful comments and suggestions. This work was partially supported by a M. G. Michael Award from the Franklin College of Arts and Sciences at the University of Georgia, AY 2023.

1 Introduction: Quine and Kaplan on the insignificance of ‘nine’ in ‘canine’

Words can be split into smaller segments in different ways. Some of them are illustrated below:

(1) a. canines → canine|s canine.PL
b. canine → ca|nine
c. canine → can|in|e
d. canine → cani|ne
e. canine → c|a|n|i|n|e

(1a) is a typical case of *morphemic segmentation* dividing words into morphemes, the smallest units of meaning contributing to the whole according to the rules of morphosemantics. The segments in (1b – 1d), on the other hand, cut across morpheme boundaries and are, in this sense, *accidental.* (1e) is the limit case of purely *orthographic* or *character* segmentation which appears to have nothing to do with semantics. (1b) and (1c), and perhaps (1e), are different from (1d) in that some segments in the former, but not in the latter, are meaningful when considered on their own: witness ‘nine’ in (1b), ‘can’ and ‘in’ in (1c), and ‘a’ (a definite article) and ‘i’ (a lowercased personal pronoun) in (1e). But these items do not contribute their usual meaning to the whole and are, for that reason, semantically inert or irrelevant. The contexts in which they appear are usually deemed to be semantically *opaque.*

Cases like (1b), (1c), and (1e) were made famous by Quine (1960, §30) and Kaplan (1969). Quine drew a stark contrast between the occurrence of singular terms like ‘nine’ in semantically transparent contexts such as

(2) Nine is greater than seven.

and in modal and propositional-attitude contexts, which he regarded as hopelessly opaque due to their resistance to substitution and existential generalization:

(3) Necessarily, nine is greater than seven.
(4) Frank believes that nine is greater than seven.

Thus (4) may be true (assuming Frank knows his arithmetic) and (5) false (if he is astronomically challenged):

(5) Frank believes that the number of planets is greater than seven.

Hence, we cannot coherently speak of *some number*, no matter how it is designated, that the predicate ‘λx (Frank believes that x is greater than seven)’ is true of:

(6) #($\exists x$) (Frank believes that x is greater than seven).

Quine motivated his pessimism about (3) – (6) by assimilating the occurrence of ‘nine’ and other similar expressions in contexts such as (3) and (4) to their occurrence in (1b) and their analogs:

> We are not unaccustomed to passing over occurrences that somehow “do not count” — ‘mary’ in ‘summary’, ‘can’ in ‘canary’; and we can allow similarly for all non-referential occurrences of terms, once we know what to look out for (Quine 1960, 144).

Kaplan’s approach, in contrast, was more optimistic. Getting inspiration from Frege’s notion of *referential shift* (Frege 1892), he took the occurrences of ‘nine’ in (3) and (4) to be fully transparent but denoting, not the number nine, but *themselves* (i.e. the expression ‘nine’, as in Kaplan (1969)), or their *sense* (in his version of intensional logic). With the aid of additional resources this allows one to make full sense of (3) – (6):

(3′) $\exists\alpha$ ($\Delta(\alpha$, nine) & $\mathbf{N}$ $\ulcorner\alpha$ is greater than five$\urcorner$).
(4′) $\exists\alpha$ ($\Delta(\alpha$, nine) & Frank $\mathbf{B}$ $\ulcorner\alpha$ is greater than five$\urcorner$).
(5′) $\exists\beta$ ($\Delta(\beta$, nine) & $\neg$ Frank $\mathbf{B}$ $\ulcorner\beta$ is greater than five$\urcorner$).
(6′) $\exists x$ (x is a number $\wedge$ $\exists\alpha$ ($\Delta(\alpha$, x) $\wedge$ Frank $\mathbf{B}$ $\ulcorner\alpha$ is greater than five$\urcorner$)).

where α and β range over expressions, ‘**N**’ and ‘**B**’ are sentential analogs of the necessity and belief operators, and ‘Δ’ is Church’s denotation predicate adapted by Kaplan. One can fully expect all of (3′) – (6′) to be true.

As Kaplan notes (Kaplan 1969), this is only the first step in a good, Fregean direction, “ripe with insight.” And his early response to Quine is just the tip of an iceberg.[1] I began with this classic exchange because it provides a useful background and a point of reference for my case study. What really matters for it is not where Quine and Kaplan disagree but where they agree: that no semantic sense can be made of the occurrence of ‘nine’ in ‘canine’ — see (1b) above — or, for that matter, of the occurrences of the subword segments in (1c – 1e). To paraphrase Kaplan, semantic concerns — substitution, existential generalization, and contribution to the meaning of the whole — are simply inappropriate to (1b – 1e) alike.[2,3] This seems to be a reasonable common ground.

The goal of this paper is to argue that recent developments in computational linguistics may prompt us to be more open-minded about this common ground. As recently noted by a leading researcher (Koehn 2020, 229),

[1] I.e. the ongoing debate on propositional attitude reports. For a recent overview, see Nelson (2022).

[2] Presumably, neither Quine nor Kaplan would object to a standard morphosemantic analysis of (1a).

[3] Quine’s take on character segmentation such as (1e) is notable in the present context. He intimates (Quine 1960, 143–4, 189–90) that *spelling* or *orthographic transcription* may be preferable to *quotation* because, unlike quotation, orthographic transcription generates not even an illusion of transparency. I revisit character segmentation in Section 3.2

> In the onslaught of deep learning on natural language processing, the linguistic concept of the *word* has survived, even as concepts such as morphology or syntax have been relegated to be just latent properties of language that can be discovered automatically in the intermediate representations of neural network models. But even that may fall. Maybe the atomic unit of language should be just the consonants and vowels, or in their written form, a character in the writing system — a letter in Latin script, a logograph or just a stroke in Chinese.

Koehn is speaking of the *semantic* import of the "intermediate hidden vector representation" of subword pieces and separate characters, such as (1a – 1e) above, not simply of their initial encoding in the form of useful numerical indices.

Such claims require careful examination, and the devil may be in the details. Language translation, I submit, is a natural place to examine them. According to the conventional wisdom, *meaning representation* and *meaning transfer* are at the very core of translation.[4] Thinkers as different as Schleiermacher, Heidegger, Benjamin, Quine, and Davidson approached this idea from rather different angles.[5] Jakobson (1959, 232) put it in a slogan: "The meaning of any linguistic sign is its translation into some further, alternative sign." On this view, the meaning of 'dog' has much, if not everything, to do with the fact that it is variously translated as *chien*, *Hund*, and *perro*. But this is just a starting point. 'dog|s' is translated as *chien*|*s* or *chien*|*nes*, and 'kick the bucket' as *casser sa pipe* ("break his pipe"). Signs, or "semantic atoms," therefore, may be word-internal functional morphemes such as '-s', or entire idiomatic phrases; they may be smaller or larger than words. In a broader perspective, different languages describe (model, represent) the extra-linguistic reality (i.e. who did what to whom) in very different ways reflected in numerous and often crosscutting typologies. For example, the "one morpheme per word" pattern of isolating analytic languages, such as Chinese (Dawson and Phelan 2016, 171):

(7) a.

[wɔ	mən	tɑn	tçin]
I	plural	play	piano

"We are playing the piano"

b.

[wɔ	mən	tɑn	tçin	lə]
I	plural	play	piano	past

"We played the piano"

[4] The conventional wisdom has been challenged from several directions usefully characterized by the translation scholar Rachel Weissbrod as follows: "[1] translation cannot transfer meaning; [2] meaning is not what translators are supposed to transfer; [3] translators are authorized to create meaning rather than transferring it; (4) translation studies is not about meaning" (Weissbrod 2018, 289). I return to the relationship between translation and meaning at the end of the paper.

[5] For recent discussions of their views on translation, see Rawling and Wilson (2018).

is contrasted with the morphological processes in synthetic agglutinative languages like Turkish in which words are formed by concatenating multiple morphemes with clean boundaries:[6]

(8) masalarımda (Turkish)

masa ('desk') + *lar* (plural) + *ım* ('my': possessive) + *da* ('at/on': locative)

"on my desks"

In polysynthetic languages such as Yupik, Chukchi, Sora, and Tiwi, highly complex words may be formed by combining several stems and affixes (i.e. both lexical and functional morphemes):[7]

(9) [angyaghllangyugtuq] (Yupik)

[angya- ghlla- ng- yug- tuq]

Boat. AUGMENT. ACQUIRE. DESIDERATIVE. 3SG

"He wants to acquire a big boat"

(10) [ŋɛnədʒdʒadarsiəm] (Sora)

[ŋɛn- ədʒ- dʒa- dar- si- əm]

I not received cooked-rice hand you.SG

"I will not receive cooked rice from your hands"

Translating between languages of different types is far from straightforward. In many cases it requires mapping subword sequences to word or phrase sequences and vice versa, and sometimes mapping a single long word into an entire sentence. It requires dealing with structures both below and above the word level whose relationship, traditionally studied in morphosyntax, may be complicated.

Even more intriguingly, Marian NMT[8] — a state-of-the-art neural machine translation engine, developed primarily by the Microsoft Translator team and widely used in production — translates the word 'periodontist' from English to its nearest neighbor French as *parodontiste*, with the source and target segmented and aligned as shown below:[9]

[6] `http://www.turkishtextbook.com/adding-word-endings-agglutination`

[7] Examples from Veselovská (2009, 47) and Dawson and Phelan (2016, 175).

[8] `https://marian-nmt.github.io`

[9] Segmentation and alignment based on the OPUS-CAT implementation of Marian NMT (Nieminen 2021); trained on over 100M English-French sentence pairs from the OPUS collection of multilingual corpora (`https://opus.nlpl.eu`).

period | on | t | ist
par | od | ont | iste

Both Quine and Kaplan would categorize such segmentations as purely "accidental"; the occurrence of 'period' in 'periodontist', in particular, is very similar to the occurrence of 'nine' in 'canine'. However, despite their allegedly opaque nature, such subword segmentations and alignments are at work in highly efficient end-to-end machine translation systems. The computational value of such processes is unquestionable. But can any sense be made of them outside machine learning?

In the remainder of this paper I attempt to cast light on this question, in the spirit of making artificial intelligence more transparent and explainable. The plan is as follows: Section 2 presents the key ideas of neural machine translation (NMT) in a way that avoids excessive technicalities but highlights the relevant details. Section 3 is a brief overview of the subword and character segmentation methods which have become part and parcel of NMT and related natural language processing applications, followed by a discussion of their theoretical significance in Section 4. I summarize the lessons of my case study and explore the broader linguistic and cognitive plausibility of some non-traditional ways of thinking about subword and character meaning in Section 5.

2 Neural machine translation

Neural Machine Translation (NMT) is one of the most impressive success stories of deep learning and artificial intelligence (AI).[10] Revolutionary innovations in the computational architectures made in 2014–2017 have led to dramatic improvements in the quality of machine translation and transformed the field forever. Despite its very real limitations, NMT keeps changing our everyday lives, even as we speak. Although the field is developing at rocket speed, with new NMT systems introduced and deployed virtually every day, some landmark achievements made in the space of three years are likely to remain part of any future history. Introduced initially in the framework of NMT, these key innovations — the original encoder-decoder model,[11] the attention mechanism,[12] and the transformer[13] — were almost immediately put to work in almost every other area of deep learning.

[10] NMT is, of course, one of a family of the latest developments in natural language processing (NLP) and computational linguistics. Others include language modeling and autoregressive text generation, document classification and summarization, question answering, speech recognition, dialog systems and personal assistants, as well as more linguistically-oriented tasks of parts-of-speech tagging, parsing (morphological, syntactic, dependency, and semantic), named entity recognition, and more. Of note are also multimodal systems combining NLP with image processing and generation, from automatic captioning to text-to-art, as well as music generation which are successfully adopting recent NLP algorithms.

[11] Sutskever, Vinyals, and Le (2014).

[12] Bahdanau, Cho, and Bengio (2014).

[13] Vaswani et al. (2017).

Importantly, NMT is also the birthplace of *subword segmentation algorithms*.[14] Initially introduced to address the problem of rare and unknown words,[15] they have proven incredibly efficient and indispensable to machine translation, as well as many other natural language processing (NLP) tasks. The origin, rapid evolution, and deployment of subword segmentation methods combine highly theoretical and sometimes speculative ideas with very practical considerations and ad hoc engineering innovations. This makes them ripe for analysis. Coming to terms with these developments may help linguists, philosophers, cognitive scientists, and researchers working in related fields catalyze their thinking about linguistic meaning and about ways in which it can be encoded and processed in natural and artificial systems, by suggesting new approaches to traditional problems.

Below I summarize some of these developments. The rest of this section introduces the NMT architecture and motivates the need for subword segmentation, which is described in Section 3. In Sections 4 and 5 I relate these developments to broader concerns about linguistic meaning.

2.1 Neural machine translation architecture

NMT models are based on *neural networks* — computational algorithms which have dominated AI research and applications since their resurgence in the form of deep learning architectures around 2006.[16] Neural networks are composed of increasingly complex and interconnected layers of basic feed-forward and recurrent units, or "neurons." At some level of approximation, a typical NMT model can be said to comprise three major components: (i) an *encoder*, which takes a source sentence (such as 'She promised me it') $s = (s_1, ...s_m)$ and applies an (ii) *attention*[17] or *self-attention* (transformer)[18] mechanism to generate a highly contextualized representation of the input, which then primes and continuously informs, in a way that may be complex and non-modular (the simplicity of Figure 1 notwithstanding), (iii) a *decoder* that generates a target sequence ('Elle me l'a promis') $t = (t_1, ...t_n)$.

The translation task can be framed as estimation of the conditional probability of translating s as t using the final softmax output and the chain rule of probability:

$$p\left(t|s;\theta\right) = \prod_{i=1}^{n+1} p\left(t_i|t_{i-1}, \ldots, t_0, s_m, \ldots, s_1; \theta\right)$$

where t_0 and t_{n+1} are conventional sequence delimiter tokens that mark the beginning and the end of a target sentence, while θ represents the whole set of model parameters which are learned jointly, by maximizing the log-likelihood of a parallel corpus D:

[14] Sennrich, Haddow, and Birch (2016); Kudo (2018); Kudo and Richardson (2018).

[15] Luong et al. (2015).

[16] See Goodfellow, Bengio, and Courville (2016).

[17] Sutskever, Vinyals, and Le (2014).

[18] Vaswani et al. (2017).

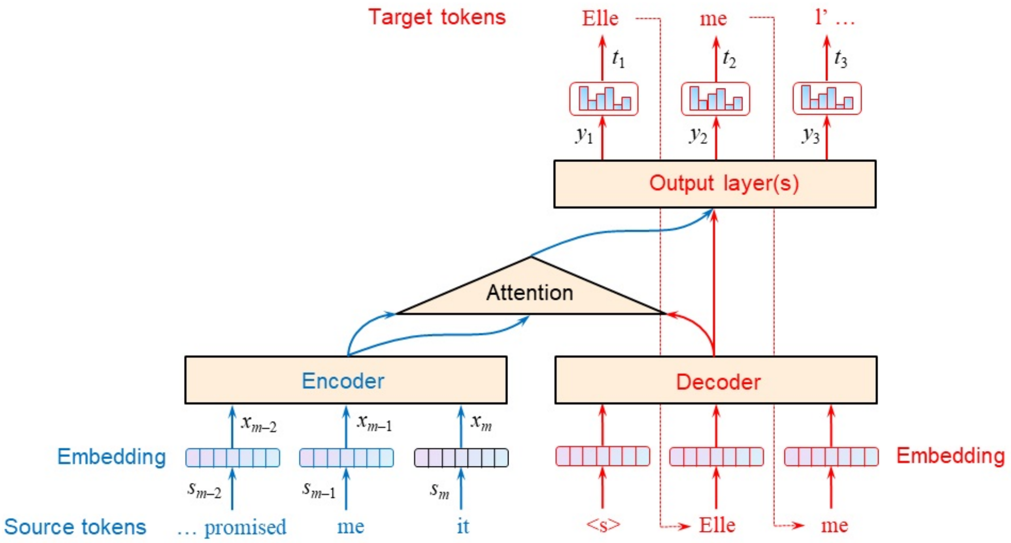


Figure 1: NMT architecture in broad outline

$$L\left(D,\theta\right) = \sum_{(s,t)\in D} p\left(t|s;\theta\right)$$

during training with backpropagation. At inference time the model translates new sentences by analogy with autoregressive generation in monolingual language models.[19]

2.2 It all begins with embedding. . .

Importantly, the whole process involves the *word embedding* operation on both sides, which takes an actual token s_i (such as 'dog' or 'chien') and projects its index (e.g. 2425) to a dense vector x_i (e.g. $(-1.452, 3.57, 0.058, \ldots 4.259)^{\mathrm{T}}$) that resides in a multidimensional embedding space[20] and can be passed on to the encoder or the decoder for further processing. Such column vectors are usually assembled into *embedding matrices* (one can think of them as giant lookup tables), and their components are learned by the model along with other parameters as part of end-to-end training.

[19] For details see, e.g., Koehn (2020) and Jurafsky and Martin (2022, Ch. 10).

[20] The geometric relations among the vectors of this space are expected to model semantic and other grammatical relations between words in a given corpus, in a way broadly similar to word embeddings in the much simpler algorithms such as Word2Vec (Mikolov et al. 2013) or GloVe (Pennington et al. 2014) which yield the famous results such as *vector* ('*King*') − *vector* ('*Man*') + *vector* ('*Woman*') = *vector* ('*Queen*') and *vector* ('*Paris*') − *vector* ('*France*') + *vector* ('*Italy*') = *vector* ('*Rome*'). But in NMT, the embedding parameters (i.e. the values of the embedding matrix) are typically learned along with other network parameters as part of *end-to-end* training.

To get a sense of the size of such a matrix and hence of the number of the corresponding model parameters to be learned, consider that every vocabulary item used in translation (on the source or target side) must be included as a separate column in the embedding matrix; otherwise the model won't know what to do with it. Moreover, every inflectional or derived form of a dictionary lemma such as 'go' takes a separate column; so there should be columns for 'goes', 'going', 'gone', and 'went' in addition to 'go'. The total number of columns needed to cover even simple domains such as news or subtitles could well exceed 20,000. Multiplying this with a typical dimensionality of the embedding space (e.g. 512) gives a conservative lower bound of the number of the embedding matrix elements (and hence, of the model parameters to learn) in the range of >10,000,000. They constitute a substantial portion, and sometimes the majority, of the model parameters and learning each of them comes with a computational cost.

For these reasons, a hard limit must be put on the vocabulary size, typically ≈ 50,000 word types. But this leads to major problems when it comes to translation.

2.3 The problem of rare and unknown words

A 50,000-word vocabulary may be too small to cover even the training corpus to begin with. In addition, the model may encounter new words at inference time. Translation is best viewed as an *open-vocabulary task*, especially for languages with highly productive morphological processes such as agglutination or compounding.[21] New words come into being all the time even in languages such as English; witness 'googling' or 'retweeting', let alone 'reeeaally' and 'sUUUpercooool'. Names of new companies (e.g. 'Trados' or 'OpenAI') or products (e.g. 'Tiguan') are introduced every day. Finally, according to Zipf's law, the distribution of words in languages is very uneven. Some of them — 'the', 'to', and other function words — could make up 30% of the entire corpus, while others such as 'latitudinous' may occur only once. There is a long tail of rare words in a typical lexical distribution.

Do items such as *muvaffakiyetsizleştiricileştiriveremeyebilecekler*[22] or *Rechtsschutzversicherungsgesellschaften*,[23] or even 'latitudinous' deserve to be included as separate columns in the computationally expensive embedding matrix? Definitely not. But then, what is the model supposed to do when it encounters, when the chips are down (i.e. at inference time), an unknown out-of-vocabulary word absent from the matrix?

[21] Consider this: every Turkish verb has over a million different inflected forms (Haspelmath 2011, 58), which far exceeds the limits of any realistic embedding matrix or, for that matter, of human memory!

[22] Turkish: "those who will not be able to make one easily/quickly a maker of unsuccessful ones." https://en.wikipedia.org/wiki/Longest_word_in_Turkish

[23] German: "insurance companies that provide legal protection." https://www.iamexpat.de/lifestyle/lifestyle-news/7-hilariously-long-german-words

3 Subword and character segmentation in neural machine translation

3.1 Byte pair encoding and other subword segmentation methods

Subword segmentation methods,[24] introduced concurrently with other developments in NMT, were designed to deal with this problem. The idea behind many of them is to start with a small vocabulary of the individual alphabet characters along with a special end-of-word symbol (e.g. '·') and then iterate over the entire training corpus (say, 20M sentences) progressively merging the most frequent pairs of adjacent characters into new symbols and adding them to the vocabulary until it reaches a preset target size (say 50,000 items).

To use a toy example,[25] consider a mini-corpus of 18 word tokens along with the frequencies of their occurrence, and the seed vocabulary of 10 characters plus '·'. Initially, the tokens are split into their individual characters:

Step	Frequency	Corpus	Merge	Vocabulary
0	5	l o w ·		·, d, e, i, l, n, o, r, s, t, w
	2	l o w e s t ·		
	6	n e w e r ·		
	3	w i d e r ·		
	2	n e w ·		

Note that 'lower' is an out-of-vocabulary (OOV) word which is absent from this training corpus. Assuming it does occur in a new sentence at inference time, the algorithm aims, among other things, to learn its segmentation. This, in turn, should allow the NMT model to leverage word-internal morphosemantic properties of 'lower'. Indeed, after eight merging operations, the corpus and the vocabulary look as follows:

[24] Such as *byte pair encoding* (Sennrich, Haddow, and Birch 2016), *SentencePiece* (Kudo 2018; Kudo and Richardson 2018), and *WordPiece* (Schuster and Nakajima 2012).

[25] Sennrich, Haddow, and Birch (2016); Jurafsky and Martin (2022, §2.4.3).

Step	Frequency	Corpus	Merge	Vocabulary
8	5	low·	(e, r) → er	·, d, e, i, l, n, o, r, s, t, w, er, er·, ne, new, lo, low, newer·, low·
	2	low e s t ·	(er, ·) → er·	
	6	newer·	(n, e) → ne	
	3	w i d er·	(ne, w) → new	
	2	new ·	(l, o) → lo	
			(lo, w) → low	
			(new, er·) → newer·	
			(low, ·) → low·	

The algorithm then applies the learned merge operations, in the order it learned them, to new sentences starting with pure character segmentation. The OOV word 'lower', in particular, gets segmented as 'low|er' (i.e. 'low er·'). This enables the NMT model to encode its compositional meaning by (i) learning the important morphosyntactic and semantic relations between the occurrences of 'new', 'er', and 'newer' on the one hand, and those of 'low' and 'er' on the other, as well as of their target counterparts, (ii) internalizing this knowledge in the form of the corresponding embeddings and attention weights, and (iii) applying it to the embedding and translation of new words, leading to marked improvements in performance.

In practice, most words end up as separate unsplit vocabulary items. But segmenting other words into smaller pieces allows the system to deal with rare and unknown words by learning and exploiting their grammatical properties and composition which are invisible when treating such words as entire unrelated tokens or, worse, as *unk*.[26] Accordingly, NMT need not be modeled at the lexical level. Instead, translation can be reformulated more generally as the task of learning the best bilingual mapping between the *sequences of subword segments* of two languages.

But it is not always that neat. Our toy algorithm segments 'worst' as 'w|o|r|s|t' and 'deer' as 'd|e|er'. Similar phenomena happen in real-life applications. As already noted at the end of Section 1, a state-of-the-art NMT engine[27] trained on around 100M English-French sentence pairs and using SentencePiece as the subword segmentation algorithm splits the new word 'periodontist' (unseen during training) as 'period|on|t|ist' and translates it into French as *par*|*od*|*ont*|*iste*. While the whole word is translated correctly, most of us would probably join Quine and Kaplan in classifying the above segmentations as semantically inert and hopelessly opaque, or even as mere "orthographic accidents." The authors of the seminal paper on *byte pair encoding* (BPE) describe their main motivation as follows:

[26] A special symbol used to replace all the OOV items in earlier solutions to the problem of rare and unknown words (Luong et al. 2015; Jean et al. 2015).

[27] Marian NMT (`https://marian-nmt.github.io`)

> Translation of some words is transparent in that they are translatable by a competent translator even if they are novel to him or her, based on a translation of known subword units such as morphemes or phonemes. . . . Our hypothesis is that a segmentation of rare words into appropriate subword units is sufficient to allow for the neural translation network to learn transparent translations, and to generalize this knowledge to translate and produce unseen words (Sennrich, Haddow, and Birch 2016, 1716).

but note that some splits fail to be transparent, and that no performance benefit is to be expected from "opaque segmentations, i.e. segmentations where the units cannot be translated independently" (*ibid.*, fn. 2).

Interestingly, this early remark proved to be overly pessimistic. Subsequent developments in subword and even pure *character* segmentation have demonstrated *performance gains* in some cases. I review character segmentation in Section 3.2 and explore the broader linguistic, cognitive, and philosophical implications of both methods in Sections 4 and 5.

3.2 Character segmentation methods

The integration of character segmentation methods into NMT started as early as 2015, concurrently with the adoption of BPE and other subword segmentation algorithms. Character methods continue to be explored, most recently in the framework of the transformer architectures.[28] At some approximation, they can be divided into *pure character* approaches and *hierarchical character-word* approaches.

A typical character-based architecture (Figure 2) is a version of the generic encoder-decoder schema (cf. Figure 1) in which word tokens are replaced with character tokens including punctuation and white spaces. This could be done on the encoder side only while keeping the decoder output at the word (or subword) level, or vice versa, or on both sides. On this approach, word segmentation becomes redundant: the sentences are treated as sequences of characters, and translation is framed as direct mapping between such sequences.

The *hierarchical* methods (Figure 3), on the other hand, seek to *supplement* word (or subword) segmentation and embedding with character segmentation and embedding (on one or both sides). The idea is to encourage the network to learn word embeddings as a *compositional function* of character embeddings during training and then apply this knowledge to generate the embeddings of unknown or rare words at inference time, feed them to the main NMT block and, if needed, perform a similar

[28] An incomplete list of important contributions to character segmentation in NMT includes Ling et al. (2015); Costa-jussà and Fonollosa (2016); Chung et al. (2016); Lee et al. (2017); Cherry et al. (2018); Gupta et al. (2019); Banar et al. (2020); Libovický and Fraser (2020); Gao et al. (2020); Li et al. (2021).

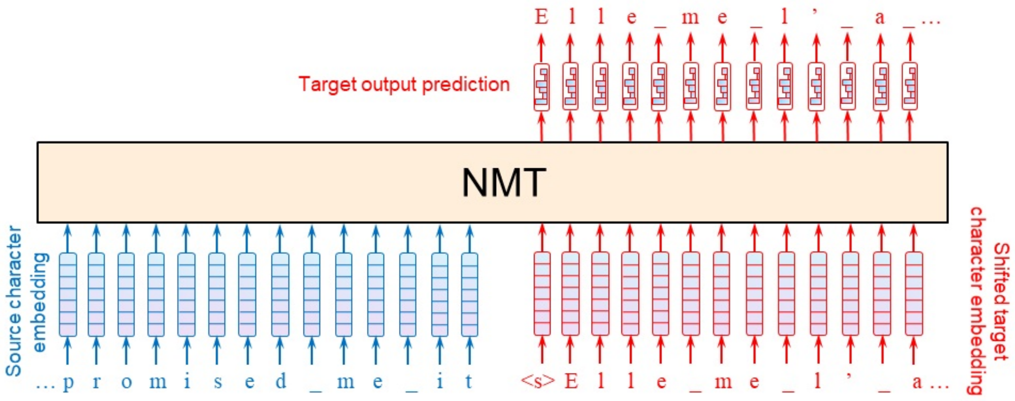

E l l e _ m e _ l ' _ a _...
Target output prediction
NMT
Source character embedding
... p r o m i s e d _ m e _ i t
<s> E l l e _ m e _ l ' _ a ...
Shifted target character embedding

Figure 2: Character-based NMT

(but potentially more complicated and context-dependent) two-step operation on the output side.[29]

An early and very successful hierarchical model was a hybrid system (Luong and Manning 2016) which translated mostly at the word level and resorted to character segmentation only when encountering unknown words. Informing word embeddings with the underlying character embeddings in various hierarchical architectures introduced since 2016 has been motivated, in no small part, by the desire to allow the network to bridge the gap between two levels of grammatical organization, that of words and that of characters,[30] with the hope of integrating orthography into the overall semantics of the sentence. It should be clear that pure character methods require no "segmentation"[31] and can operate with very small "vocabularies" of about 200 characters for both languages.[32]

On a more theoretical side, the seminal work on subword and character segmentation was followed by studies exploring the broader linguistic and conceptual implications of these methods. They raise intriguing questions to which I now turn.

[29] To model word composition from characters, a convolutional neural network or an additional recurrent neural network layer can be used, possibly supplemented with feed-forward, "residual" or "highway" connections.

[30] And in-between: "We can first break up words into subwords and then model these subwords with character-based models" (Koehn 2020, 232).

[31] Other than the initial tokenization. The text is already "segmented" into characters!

[32] But they have to deal with much longer sequences of tokens (500+ versus 20–50 in word and subword methods). This comes with a computational cost (thus Luong and Manning's baseline hierarchical model took about 3 months to train on state-of-the-art GPUs (Luong and Manning 2016)), which calls for compression (Cherry et al. 2018) and other work arounds.

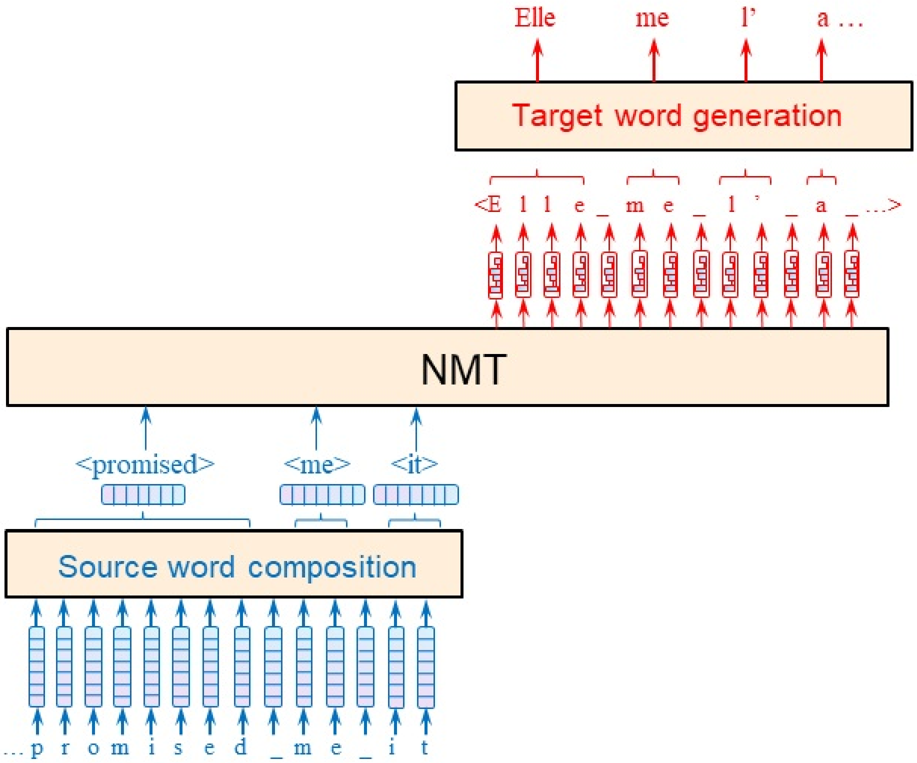


Figure 3: Hierarchical NMT

4 Subword segmentation and the boundaries of meaning

The questions are many and diverse. Below is a partial list followed by a discussion of some important observations made by the practitioners. How grammatical is character-based NMT? Can attention (and self-attention) over character sequences be motivated — linguistically, cognitively, or at least philosophically? What about subword segmentations that cut across morpheme boundaries? Can such sequences constitute reasonable semantic units and contribute to the meaning of the whole of which they are parts? Relatedly, can a neural network, even a very deep and sophisticated one, learn a highly nonlinear mapping from a sequence of characters or non-morphemic subword pieces to the meaning of a sentence? And if it can, does it make linguistic or conceptual sense?[33]

[33] Notably, some of these concerns are reflected in the titles of research publications: "From characters to words to in between: do we capture morphology?" (Vania and Lopez 2017); "Meaningless yet meaningful: morphology grounded subword-level NMT" (Banerjee and Bhattacharyya 2018); "Learning to segment inputs for NMT favors character-level processing" (Kreutzer and Sokolov

One important strand of work deals with the question of whether subword or character segmentation methods can successfully model morphology; for example, learn to split 'canines' as 'canine|s' rather than 'can|in|e|s' or 'ca|nine|s'. There are two sides to this question: (i) Is a method such as BPE or character segmentation capable of learning morphology *on its own*, from pure strings of text? (ii) Can an NMT model as a whole *benefit* from incorporating explicit *morphological tools*[34] into the pre-processing segmentation step (and/or post-processing de-segmentation step)? The first question becomes particularly tangible in the context of recent work demonstrating neural networks' ability to learn, from pure strings of text, complex *syntactic* phenomena such as center embedding and other long-distant dependencies.[35] In view of the almost continuous nature of morphosyntax,[36] one should expect phenomena of that sort to have counterparts in the corresponding morphological processes.[37]

Can one go below the level of characters? Surprisingly, or perhaps not, the answer is yes. All Unicode characters are composed of *bytes*, so there is a sense in which all the world language "vocabularies" eventually bottom out at just 256 tokens! This feature was exploited in some earlier work on segmentation in NMT (Costa-jussà, Escolano, and Fonollosa 2017). Chinese, Japanese, and other *logographic* languages present an especially interesting case of subword semantics heavily informed by both

2018); "Learning morphology for open-vocabulary neural machine translation" (Ataman 2019); "One size does not fit all: comparing NMT representations of different granularities" (Durrani et al. 2019).

[34] Such as Morfessor (Smit et al. 2014; Grönroos et al. 2014).

[35] See, in particular, Gulordava et al. (2018) and Linzen and Baroni (2021).

[36] A world-renowned typologist Martin Haspelmath has argued that word is a poorly defined concept which has much to do with our "bias towards written language and the strong influence of the habit of word separation by spaces in Western languages" (Haspelmath 2011, 33) and that there is "no good basis for a general, cross-linguistically viable word concept, and hence no basis for a general bifurcation between morphology and syntax." They are best thought of as a *unitary* theoretical domain (*ibid.*, 32, 72). Cf. a related discussion at the end of Section 1.

[37] Translating a single word of a polysynthetic language such as Sora into a long sentence of English is probably the most striking example of morphosyntax interpolating between two very different semantic decompositions (cf. examples (9) and (10) in Section 1 above). While modeling such cases in MT is not yet practically possible due to the virtual absence of training data for polysynthetic languages (but see Ortega, Castro Mamani, and Cho (2020) for important first steps in this direction), much effort has gone recently into studying *low-resource* language directions for which there is some data (say, 100–300K sentence pairs), even if it is two orders of magnitude smaller than the data for English-German or French-English. Low-resource settings can also be recreated by limiting the amount of data otherwise available to the system.

orthography and phonology, with a distinctive morphosyntax arising from their complicated interaction.[38,39]

The primary goal of the foregoing brief and selective survey of recent subword, character, and sub-character segmentation methods was to highlight important theoretical issues emerging from their application. I conclude this section with two general comments.

(1) While each of these results is valid and important, their cross-comparison is difficult because of the different settings, goals, corpora, domains, language pairs, base model varieties, training regimes, evaluations methods, and many other confounding factors involved.[40]

(2) Despite this diversity, the efficacy of many of these methods is beyond doubt. This calls for an explanation and a broader reflection, especially in cases of seemingly "opaque" segmentations breaking the performance records.

To take stock and prepare the ground for such reflection, I summarize the methods considered in this section in a schematic form. At some level of abstraction, every NMT model maps a sequence of indexed source tokens to a corresponding target sequence (Figures 1–3). The tokens may be words, subwords, characters or

[38] Chinese words, in particular, are composed of characters (*hanzi*) representing separate and unchangeable free morphemes. There is no inflexion and no natural spaces between words, so tokenization is often necessary to mark their boundaries for downstream NLP tasks such as translation. The resulting "words" are already too short (2.4 characters on average) for any subword segmentation method such as BPE to be useful. Most characters, however, are *structured logograms* whose parts (*ideographs* or *radicals*) combine in a systematic way to encode semantic and phonetic information. For example, Yeh, Chou, and Ho (2017) note that the character 猜 ('guess') is composed of the semantic radical 犭 denoting a categorical unit of meaning ('wild animal') and the phonetic radical 青 providing a pronunciation cue [qing1], which is needed due to the widespread homophony in Chinese. Importantly, some phonetic radicals can have *meaning* on their own and can even function as standalone characters. Thus 青 means 'cyan' when occurring in isolation. This blending of phonology with semantics is quite remarkable, and I briefly revisit it in Sections 5.2 and 5.3 below.

[39] Building on earlier studies, Zhang and Komachi recently investigated the performance of RNN- and transformer-based NMT systems for six language pairs involving Chinese, Japanese, and English (Zhang and Komachi 2021) at different levels of sub-character segmentation — "raw ideographs," "finest ideographs," and strokes — while making flexible use of the resulting shared vocabulary tokens between Chinese and Japanese (such as many common strokes historically imported from *hanzi* to *kanji*). Their results show that finer granularity of sub-character segmentation for both Chinese and Japanese consistently improves MT performance peaking at the stroke level on the source side and "ideo-finest" on the target side. The latter may be due to the semantic opacity of strokes or, as the authors suggest, to the decoding challenges presented by the much longer stroke sequences. Or both. The question might be worth exploring further.

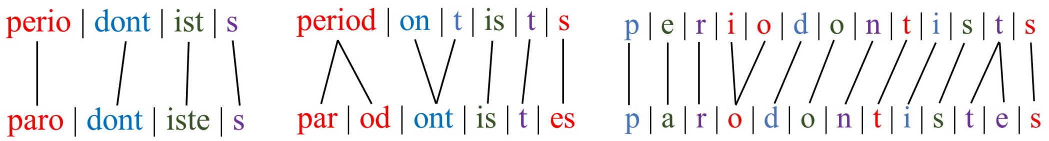
[40] Unlike in many other shared machine translation tasks, there are still no uniform benchmarks for segmentation, which is hardly surprising given the sheer diversity of the approaches. This makes the overall picture rather mosaic. The field is developing very rapidly, with entire conferences and sessions devoted to segmentation in NMT and other NLP tasks. At the time of writing, various character and sub-character methods continue to perform best on some language pairs, domains, or datasets while subword methods outperform them on most others. And in many cases, what matters may indeed be hiding in the numerous details.

sub-characters. The first pre-processing step transforms the actual input string of source words into a sequence of such tokens which is then fed to the embedding layer generating their dense representations. This transformation may be as simple as *identity* (in word-based models) or as complicated as a separate neural network layer. A chosen encoder-decoder model maps the resulting representations to the target token predictions (this is where most of the deep learning magic happens) while using their embeddings "to keep going." Finally the resulting string of the target tokens (i.e. sub-characters, characters, or subwords) is post-processed into an output sequence of the actual target words. Table 1 below illustrates various (real and hypothetical) segmentations of 'periodontists'.

<table>
<tr><th>Segmentation method</th><th>Segmentation output</th></tr>
<tr><td>Morphological parser</td><td>perio | dont | ist | s</td></tr>
<tr><td>Subword (e.g. SentencePiece)</td><td>period | on | t | ist| s</td></tr>
<tr><td>Character</td><td>p | e | r | i | o | d | o | n | t | i | s | t | s</td></tr>
</table>

Table 1: Alternative segmentations of 'periodontists'

Figure 4 illustrates various (real and hypothetical) segment alignments for the translation of 'periodontists' as *parodontistes*, which could be gleaned from the attention or cross-attention weights:


perio | dont | ist | s
paro | dont | iste | s
period | on | t | is | t | s
par | od | ont | is | t | es
p | e | r | i | o | d | o | n | t | i | s | t | s
p | a | r | o | d | o | n | t | i | s | t | e | s

(a) (b) (c)

Figure 4: Alternative alignments for 'periodontists' → *parodontistes*

When everything is said and done, the difference between 'perio|dont|ist|s' and 'period|on|t|ist|s' is still staring us in the face. And making semantic sense of alignments such as those in Figures 4(b) and (c) still looks to be an (almost) impossible task.[41] But one general lesson from the above discussion is that most of the segmentations and translation alignments similar to those shown in Table 1 and Figure 4

[41] As noted earlier, the occurrence of 'period' in 'period|on|t|ist|s' is quite similar to the occurrence of 'nine' in 'ca|nine'; Quine (1960, §30) and Kaplan (1969) would consider both of them *orthographic accidents*. Approaching the matter from a completely different, computational angle, Chung and colleagues seem to agree with Quine and Kaplan on this point: "Because of this view of words as basic units of meaning (either in the form of lexemes or derived form) from linguistics, much of previous work in natural language processing has focused on using words as basic units of which a sentence is encoded as a sequence. Also, the potential difficulty in finding a mapping between a word's character sequence and meaning [for instance, 'quit', 'quite' and 'quiet' are one edit distance

can be operative in highly successful end-to-end NMT systems, despite the fact that many of them cut across the natural “meaning joints” typically corresponding to intra-word morpheme boundaries.

Perhaps no semantic sense can be made of such phenomena, and we simply have to accept the “unreasonable effectiveness” of subword and character segmentation in NMT as another unfathomable fact of deep learning. I want, however, to end by briefly reviewing two unrelated linguistic proposals that might help make non-morphemic segmentation a bit more reasonable.

5 Can non-morphemic subword segments have a semantic oomph?

At some approximation,[42] all the segmentation methods considered in Section 3 can be said to learn a function f from the representations (i.e. the embeddings) of subword units e_{sub} to the representations (the embeddings) of words e_w:

$$e_w = f\,(e_{sub}, \sigma(w)) \tag{11}$$

Here σ is a chosen segmentation algorithm, which takes a word token as input and returns a sequence of segments. For example, a morphologically aware algorithm such as Morfessor may be expected to yield

$$\sigma(`\mathit{periodontists}') = (`\mathit{perio}', `\mathit{dont}', `\mathit{ist}', `\mathit{s}') \tag{12}$$

while BPE or SentencePiece may return

$$\sigma(`\mathit{periodontists}') = (`\mathit{period}', `\mathit{on}', `\mathit{t}', `\mathit{ist}', `\mathit{s}') \tag{13}$$

and a pure character segmentation will produce

$$\sigma(`\mathit{periodontists}') = (`\mathit{p}', `\mathit{e}', `\mathit{r}', `\mathit{i}', `\mathit{o}', `\mathit{d}', `\mathit{o}', `\mathit{n}', `\mathit{t}', `\mathit{i}', `\mathit{s}', `\mathit{t}', `\mathit{s}') \tag{14}$$

Some MT researchers explicitly refer to f as a *composition function*,

> which can establish a mapping between combinations of orthographic units and lexical meaning, that is learned using the bilingual context so that it can produce representations that are optimized for machine translation (Ataman and Federico 2018, 306).[43]

away from each other but have distinct meanings] has likely contributed to this trend toward word-level modelling” (Chung, Cho, and Bengio 2016, 1695).

[42] Adapted from Vania (2020, §3.2).

[43] See also Ling et al. (2015).

This makes sense in view of the fact that the geometric relations among the word and subword embeddings are expected to model semantic and other linguistic relations among the corresponding tokens.[44] Thus the meaning of ‘perio|dont|ist|s’ may be learned as a function of the meanings of the morphemes ‘perio-’, ‘-dont-’, ‘-ist-’ and ‘-s’ and the way they are put together (in this case, a simple concatenation). The question, however, is whether such a function is genuinely compositional in a strict semantic sense.[45] It becomes particularly pressing in cases of non-morphemic segmentations such as (13) and (14). Can any sense be made of them outside machine learning?

Below I consider two extant linguistic proposals approaching this question from the opposite sides of the semantics spectrum.

5.1 Zadrozny on the “triviality” of compositional semantics

On one side, there is a tradition of arguing that semantic compositionality is vacuous or trivial, culminating in Zadrozny’s proof (Zadrozny 1994) that for any function $m(s)$ from a set of strings S to meanings $m \in M$, there is a new meaning function μ such that for any $s, t \in S$,

(15) $$\mu(s \cdot t) = \mu(s)(\mu(t))$$

(16) $$\mu(s)(s) = m(s)$$

where ‘·’ is string concatenation. For example, on this proposal, μ maps ‘chases’, ‘mice’, ‘rainbows’, ‘chases mice’, and ‘chases rainbows’ to *functions* from themselves to their ordinary meanings:

$$\begin{aligned}
&\mu(\text{‘}\textit{chases}\text{’}) = f_1 : \text{‘}\textit{chases}\text{’} \mapsto m(\text{‘}\textit{chases}\text{’})\\
&\mu(\text{‘}\textit{mice}\text{’}) = f_2 : \text{‘}\textit{mice}\text{’} \mapsto m(\text{‘}\textit{mice}\text{’})\\
&\mu(\text{‘}\textit{rainbows}\text{’}) = f_3 : \text{‘}\textit{rainbows}\text{’} \mapsto m(\text{‘}\textit{rainbows}\text{’})\\
&\mu(\text{‘}\textit{chases mice}\text{’}) = f_4 : \text{‘}\textit{chases mice}\text{’} \mapsto m(\text{‘}\textit{chases mice}\text{’})\\
&\mu(\text{‘}\textit{chases rainbows}\text{’}) = f_5 : \text{‘}\textit{chases rainbows}\text{’} \mapsto m(\text{‘}\textit{chases rainbows}\text{’})
\end{aligned}$$

But due to the type raising exhibited in (15), μ also maps some elements from its own range (i.e. *functions* such as above) to other such elements; specifically:

[44] See Section 2.2 above.

[45] The issue of semantic compositionality in NLP is by no means new. Studies exploring ways of modeling morphology in neural network-based language modeling predate NMT and BPE (Luong, Socher, and Manning (2013), Botha and Blunsom (2014)). And important recent work probing the ability of neural networks to learn long-distance syntactic and semantic relations was already mentioned above. For an illuminating summary, see Baroni (2019). For interesting recent attempts to connect broader linguistic and philosophical concerns about compositionality with cutting-edge work in NLP, see Hupkes et al. (2020); Nefdt (2020); Dankers et al. (2022).

$$\mu(\text{‘}\textit{chases}\text{’}) = f_1 : \mu(\text{‘}\textit{mice}\text{’}) \mapsto \mu(\text{‘}\textit{chases mice}\text{’})$$
$$\mu(\text{‘}\textit{chases}\text{’}) = f_1 : \mu(\text{‘}\textit{rainbows}\text{’}) \mapsto \mu(\text{‘}\textit{chases rainbows}\text{’})$$

thus ensuring compositionality:

$$\mu(\text{‘}\textit{chases mice}\text{’}) = \mu(\text{‘}\textit{chases}\text{’})(\mu(\text{‘}\textit{mice}\text{’}))$$
$$\mu(\text{‘}\textit{chases rainbows}\text{’}) = \mu(\text{‘}\textit{chases}\text{’})(\mu(\text{‘}\textit{rainbows}\text{’}))$$

while still recovering the original meanings via (16):

$$\mu(\text{‘}\textit{chases}\text{’})(\text{‘}\textit{chases}\text{’}) = m(\text{‘}\textit{chases}\text{’})$$
$$\mu(\text{‘}\textit{mice}\text{’})(\text{‘}\textit{mice}\text{’}) = m(\text{‘}\textit{mice}\text{’})$$
Etc.

This allows one to embrace $m(\text{‘}\textit{chases mice}\text{’}) = m(\text{‘}\textit{chases}\text{’})(m(\text{‘}\textit{mice}\text{’}))$ and reject $m(\text{‘}\textit{chases rainbows}\text{’}) = m(\text{‘}\textit{chases}\text{’})(m(\text{‘}\textit{rainbows}\text{’}))$ — an intuitively correct outcome in light of the latter's idiomaticity. Semantic compositionality is thus trivial.

In view of the central role of string concatenation in subword and character segmentation methods in NLP, Zadrozny's result might be used to argue that a "composition function" f (see (11) above) learned by such methods could be said to be compositional in a semantic sense, making the meaning of ‘canine’ computable from the meanings of ‘ca’ and ‘nine’ — or even of ‘c’, ‘a’, etc. — and the orthography of a given language. Nothing else is needed.

While those working on explainable AI may welcome this formal rapprochement, the real question is whether Zadrozny's function μ, mapping strings to *other functions*, satisfies the needs of compositional semantics. Early discussions raised serious doubts about it. Whereas $\mu(s)$ may itself be compositional in some sense, its relation to genuine concerns about meaning and semantic compositionality may be rather distant. Kazmi and Pelletier (1998) note that the values of $\mu(s)$ are at best "pointers to meanings." Dever (1999) demonstrates that $\mu(s)$ violates a major constraint on meaning by failing to preserve synonymy: it maps two distinct expressions having the same intuitive meaning to *different functions*. As noted above, this has to do with type raising resulting in two distinct components of μ, one mapping strings to "meanings" and the other mapping them to other outputs of μ. With all the heavy lifting done by the latter, Westerståhl concludes that "Zadrozny's theorem . . . makes the meaning assignment one-one in an unmotivated way, thereby side-stepping the compositionality issue" (1998, 641–2). Szabó concurs making a more general point on all similar proposals: "It is trivial that we can compositionally assign something to each expression of a language (for example, if expressions serve as their own meanings, semantics is certainly compositional!) but it does not follow that it is trivial to adequately assign meanings to them" (2020, §1.2).

Zadrozny himself partly concedes the point by distinguishing between "systematic" and "non-systematic" ways of trivializing compositionality. He ends his 1994 paper by noting that one of the more bizarre consequences of his result is that

> we do not have to start building compositional semantics for natural language beginning with assigning of the meanings to words. We can do equally well by assigning meanings to *phonemes* or even *LETTERS*, assuring that, for any sentence, the intuitive meaning we associate with it would be a function of the meanings of the letters from which that sentence is composed. But then the cabalists had always known it. (Zadrozny 1994, 341)

Again, this may strike a positive chord with some recent work on character segmentation. But for the reasons noted above, Zadrozny's approach is unlikely to be of much help in aligning this work with more intuitive ways of thinking of subword meaning. There is, however, a proposal on the other side of the spectrum which explores an intriguing and linguistically motivated way of making strict semantic sense of non-morphemic subword segments.

5.2 Artstein on compositional semantics for prosodic constituents

In a number of works published in 2002–2005, Ron Artstein argues that *phonological decomposition* can be used to assign meaning to arbitrary subword segments, such as 'mite' in 'stalagmite', and 'ortho' and 'perio' in 'ortho and periodontists',[46] in a strictly compositional way that should satisfy the needs of both common sense and rigorous linguistic semantics. The main support for the proposal comes from the analysis of intonational *focus* and *coordination* at the subword level.

Focus is a familiar grammatical phenomenon used to indicate which part of the sentence contributes new or contrastive information. Association with focus is widely regarded as a compositional semantic process,[47] as witnessed in sentences such as:

(17) John only introduced TED to Mary.

where the domain of 'only' is restricted in a predictable way so that the *focused meaning* of the VP 'introduced Ted to Mary' is not the property of introducing Ted to Mary but the *set* of properties of the form 'introduced x to Mary' where x ranges over individuals. Suppose John did not introduce anyone other than Ted to Mary,

[46] Although 'mite' has a meaning on its own it does not contribute the latter to the meaning of 'stalagmite'. In this sense, the occurrence of 'mite' in 'stalagmite' is similar to the occurrence of 'nine' in 'canine' (Section 1). Furthermore, 'stalagmite' derives from *stalagma* (Greek for 'dropping'), so 'stalag|mite' cuts across natural morpheme boundaries. While 'ortho' and 'perio' are morphemes they do not occur on their own, and their commonly accepted lexical and etymological profiles do not explain what goes on in subword coordination such as (21) below. For details, see Artstein (2002, Chapters 2 and 4).

[47] Artstein (2002, 11ff) adapts the semantics of syntactic focus due to Rooth (1992).

although he introduced Ted to Ann, Bob, and any number of other people. Then (17) is *true* on its intended meaning supplied by intonational focus syntactically marked with $[\,]_F$ and by the relevant context variable C_i co-indexed with *only*:[48]

(18) John only$_i$ [$_{\text{VP}}$ [$_{\text{VP}}$ introduced TED$_{\text{F}}$ to Mary] ∼C$_i$].

Artstein proposes to extend this strictly compositional semantics to focus cases such as:

(19) John only brought home a stalagMITE from the cave.

(19) is intended to be true in case John also brought home other objects from the cave (rocks, insects, etc.), as long as he didn't bring home a *stalactite*.[49] By analogy with (17), the focused meaning of 'brought home a stalagmite from the cave' is not the ordinary property of bringing home a stalagmite from the cave but the set {'x brought home a stalagmite from the cave', 'x brought home a stalactite from the cave'}, with the relevant focus induced by the ungrammatical shift of stress in 'stalagmite' in (19). An analysis parallel to (18) then yields:

(20) John only$_i$ [$_{\text{VP}}$ [$_{\text{VP}}$ brought home a stalag[MITE]$_{\text{F}}$ from the cave] ∼C$_i$].

An obvious problem is that unlike in (17), it is not clear what focus operates on in (19), semantically speaking. As noted above, although 'mite' happens — by *orthographic accident* similar to the occurrence of 'nine' in 'canine' — to denote something, its denotation is irrelevant to the meaning of 'stalagmite'.

A similar phenomenon is exhibited in *subword coordination*:

(21) Bill and Martha are ortho and periodontists.

which is expected to be true if Bill is an orthodontist and Martha is a periodontist, with 'and' operating on the apparently opaque word parts 'ortho' and 'perio' (Artstein 2002, Chapter 4 and Artstein 2005).

With the hope of recovering the standard meaning of 'stalagmite' from 'stalag' and 'mite', and of 'orthodontist' and 'periodontist' from 'ortho', 'perio' and 'dontist' in a strictly compositional way while fully respecting linguistic intuitions, Artstein develops his phonological decomposition approach as follows:

> The denotation of a focused or coordinated part is the sound of that part itself, so the word parts *mite* in [19] and *ortho* and *perio* in [21] denote their own

[48] For details, see Artstein 2002, §2.2.

[49] Artstein (2004, 1) begins with an even more graphic example of intonational focus in 'stalagMITE' from a New Yorker cartoon, which features a psychiatric patient standing upside down, with his feet on the ceiling. The psychiatrist tells his wife that the first order of business is "to persuade the patient that he is a stalagmite," thus implying that the patient thinks he is a stalactite.

> sounds. Sounds are objects in the model (entities of type e). The rest of the word — the unfocused part, or the part outside the coordinate structure — denotes a function from sounds to word meanings, which retrieves the original meaning of the word. Thus, *stalag* denotes a function that for each sound α yields the meaning of the word stalagα, if such a word exists; similarly, *dontist* maps a sound β to the meaning of the word β*dontist*. The meanings of two parts of a single word combine through the composition rule of function application to yield the meaning of the word they form; focus and coordination have access to the individual word parts, and they manipulate them to arrive at the meanings of focus constituents and coordinate structures (Artstein 2002, 3).

Thus, instead of being opaque, 'mite' and 'tite' in 'stalagmite' and 'stalactite' denote their own sounds:

$[\![\text{MITE}_{\text{F}}]\!] = [\text{ma}^{\text{j}}\text{t}]$
$[\![\text{TITE}_{\text{F}}]\!] = [\text{ta}^{\text{j}}\text{t}]$

while $[\![\text{stalag}]\!]$ is a partial function f such that:

$f([\![\text{MITE}_{\text{F}}]\!]) = [\![\text{stalagmite}]\!]$
$f([\![\text{TITE}_{\text{F}}]\!]) = [\![\text{stalactite}]\!]$
$f(\alpha)$ is undefined for any other α.[50]

Subword coordination exhibited in (21) can be decomposed in a similar way.[51]

The phonological nature of non-morphemic subword semantics also takes center stage in imposing distinctly *prosodic constraints* on the material that can be focused or coordinated, as illustrated in the following contrast (Artstein 2004, 13):

(22) This is a morphological problem that gets a (ˈPHONO)(ˌlogi)cal solution.

(23) I have trouble with morphology, but he will only discuss pho(ˈnolo)gy.
*(ˈPHONO)(ˌlogy)

Specifically, focus cannot be marked on 'phono' in (23), even though it is a morpheme. As Artstein notes (2002, i, 28ff), only prosodic units the size of at least a *metrical foot* can be focused or coordinated, and the existing prosodic structure must be preserved.[52]

In one important respect, Artstein's semantics of focus and coordination is similar to Frege's original treatment of quotation, which was extended by Kaplan to modal and propositional attitude contexts (see Section 1): according to the latter,

[50] Artstein 2004, 7.

[51] Artstein 2002, 55–7. Artstein also shows (2002, Chapter 5) that his phonological-decomposition analysis of focus can be applied to *echo questions* such as 'This is a stalag-WHAT?'

[52] The constraints at work here are quite similar to those in the famous examples of expletive infixation (McCarthy 1982). Cf. 'psychobloodylogical' vs. *'psychobloodylogy'.

in some contexts expressions refer to *themselves*; while in Artstein's theory, they refer to their *phonological form.* Both approaches are designed to deal with the alleged opacity of the respective contexts. One important difference, of course, is that Frege's and Kaplan's proposals operate at the level of words, whereas Artstein's phonological decomposition is designed to handle non-morphemic subword segmentation. Bringing such a seemingly wild phenomenon to the forefront of semantics is notable.

Artstein's approach also stands in contrast with influential views relating the availability of focus to the semantic transparency of word parts. He quotes Chomsky's footnote to the effect that "the focus must be composed of full lexical items" adding that "this amounts to the claim that the semantics of focus can only apply to units that have an independent lexical meaning" (quoted in Artstein 2004, 16). The irrelevance of the independent meaning of 'mite' to the meaning of 'stalagmite', and the total lack of independent meaning in 'perio' should put the advocates of the conventional wisdom on the alert.

Artstein's account is certainly controversial and brings with it quite a bit of theoretical pain, such as the need to deal with numerous adverse cases.[53] It is not my goal to defend it here. I brought it up as an interesting example of a linguistically-motivated approach to non-morphemic subword meaning and a useful contrast to Zadrozny's trivialization result. Together, they mark the opposite boundaries of the logical space available to those seeking to align the astounding success of segmentation methods in NMT with intuitive demands on meaning. Any attempt to throw light on their "unreasonable efficiency," even by analogy with rare but cognitively plausible linguistic phenomena, should, I think, be welcomed.

5.3 Cutting across boundaries: phonology and non-morphemic segmentation in NLP and human language processing

It might seem that phonological decomposition is at a remove from written translation, human or machine. Translation, however, is one of a family of interrelated NLP tasks which can be used in combination, for example in automatic speech translation. Crossing the boundaries between orthography, phonology, and morphology is quite appropriate from this broader perspective and may in fact be fruitful. Indeed, the foregoing discussion should make it clear that any human-drawn boundaries between the levels of grammatical organization are at best relative. I end this section with a brief mention of some roles phonetic and phonological parameters were found to play in subword processing tasks, both in machines and humans.

As regards the former, Kim, Hirasawa, and Komachi (2020) report improvements in the *zero-shot* North Korean → English NMT, based on character segmentation enhanced with *phoneme decomposition.* The idea is motivated by the grammatical differences (in word segmentation, initial sound rule, and compounding) between

[53] Artstein is by no means unaware of this.

the two Korean languages, and the virtual absence of North Korean-English parallel data. Building on previous work on similar low-resource settings, Kim and colleagues demonstrate the potential of their "character-phoneme BPE" segmentation model.

As already mentioned (note 38), an intricate interplay of phonology and semantics at the level of sub-character segmentation in logographic languages such as Chinese and Japanese was put to work in NMT (Zhang and Komachi 2021). The composite nature of Chinese *hanzi* raises intriguing questions about the encoding and processing of their phonological and semantic components in both NLP and human language processing. Sub-character segmentation methods adopted in the former are initially blind to the distinctions between these two aspects of grammar but may learn some of them during training. Yet the way the sub-character elements — the ideographs (radicals) and the strokes — end up dividing the combined morpho-phonetic task between themselves may be rather unusual, depending on the numerous details of the training corpora, training regiments, and the domains.

What about human processing of the Chinese characters? For example, can phonetic radicals whose main job in complex characters is to supply pronunciation cues, activate *semantic* access, especially when such radicals can also stand on their own, being in this respect similar to 'nine' in 'canine' but operating *below* not above the character level? In a recent psycholinguistic study, Yeh, Chou, and Ho (2017) investigated this question with a version of the *Stroop test*.[54] Their main finding is the semantic activation precipitated by radicals such as 青 ('cyan') in cases where they perform their phonetic function (i.e. provide pronunciation cues) in a semantically unrelated composite character such as 猜 ('guess'), thereby interfering with the latter's processing. This seems to support the view that (i) *hanzi* are recognized by activating access to radicals first, and that (ii) the radicals thus activated need not be semantically transparent and may, in fact, have a 'wrong' meaning when occurring in isolation.

Non-morphemic subword, character (and sub-character!) segmentation thus may have *some* non-trivial cognitive plausibility to it.[55]

6 Concluding remarks

The main lessons from the case study of the subword and character segmentation methods in neural machine translation undertaken in this paper are as follows: (i)

[54] Widely used to measure cognitive interference or facilitation between color perception and text processing (see, e.g., Scarpina and Tagini 2017). The original Stroop effect is usually associated with the delay in processing 'red' printed in green, say, compared with processing a non-color word such as 'barn' printed in green.

[55] Artstein (2002, Section 6.3) mentions earlier psycholinguistic studies that seem to be in line with his claim that the semantics of non-morphemic subword segments tracks their phonological form. For example the lexical processing of a word like 'candle' begins as soon as [kænd] is heard and activates the meaning of the semantically unrelated 'candy'. A similar effect can happen at the end of a word, with the second syllable in 'beaker' activating access to 'speaker'.

these methods are many and diverse, forming almost a continuous spectrum, with explicit morphologically-informed approaches on one side and sub-character methods on the other; (ii) nearly all these methods have demonstrated highly successful performance in select settings, often without recognizing any *a priori*, human-imposed boundaries of subword structure; (iii) this success calls for an explanation that must be sensitive to the details of a given application; (iv) such an explanatory project could be aligned with important topics in theoretical linguistics and philosophy of language, and (v) might suggest new ways of thinking about some traditional problems by extending the boundaries of their logical space.

While exploring these avenues in detail is not possible here, I want to briefly revisit the relationship between translation and meaning mentioned in Section 1.[56] In light of the foregoing discussion, this relationship may give rise to a number of theoretical options. According to the traditional view, the connection between translation and meaning is very tight: good translation involves, and perhaps boils down to, meaning preservation. Some go further and argue, in effect, that meaning *is* what is preserved by good translation (Jakobson 1959). If that is the case then, given that good translation may be learned by a neural network as direct mapping between semantically opaque subword segments devoid of any "standard meaning," the empirical success of these methods could perhaps constitute a *reductio* of the attempts to ground meaning in translation (rather than in reference or truth conditions).

Alternatively, and more controversially, one may refuse to call what neural networks do *translation*, perhaps by analogy with refusal to characterize natural language "understanding" performed by neural networks trained on any number of chosen objectives, as any kind of *understanding.* Or even with the well-motivated refusal to call LaMDA or other large language models *sentient.*[57] On this view, real translation requires genuine grasp of meaning grounded in world knowledge, an unlimited number of contextual parameters, and other anchoring points unavailable to neural models. This is a plausible view and it is, in fact, popular in some circles. But defending it vis-à-vis recent developments which are rapidly erasing the remaining boundaries between human and machine translation is getting progressively difficult and may, in the end, be an uphill battle. Despite all its very real limitations, MT has become amazingly, even scaringly, good. Simply refusing to characterize it as *translation* may not be the best way to deal with the current situation. And the analogies with the hype associated with large language models may be strained: general enthusiasm about machine translation has no science fiction flavor to it, and no claim of "sentience" of any sort is made by the stakeholders. Indeed, MT is widely (but wrongly) perceived as being "solved" and, therefore, relatively uninteresting, compared to giant language models and other recent machine learning sensations.[58]

[56] I thank a referee for encouraging me to do so, and for a very helpful suggestion.

[57] In reaction to Google's engineer's recent provocative claim (`https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine`).

[58] One of the luminaries of the field of machine translation describes the flow of intellectual resources through MT as follows: "It is only a slightly exaggerated characterization to say that

The two theoretical options sketched above are based on severing the connection between neural MT and meaning, either by denying that translation, human or machine, is constitutive of meaning or, more controversially, by denying that MT is genuine translation. This leaves room for a third and, perhaps, the most controversial option that must be mentioned. Traditional semantics starts by drawing a binary divide between meaningful and meaningless strings. And much of traditional philosophy of language starts with a reflection on Meaning with a capital 'M'. But recent progress in natural language processing, as manifested in the stunning success of NMT, suggests that one could also start by thinking of linguistic meaning as something that can be *processed* by natural and artificial neural network-based systems alongside other features (morphosyntactic, phonetic, or orthographic) and then let the chips fall where they may. To borrow Kaplan's expression, this way of thinking may be "ripe with insight." In the end, there may be many dimensions of meaning that are specific to particular tasks, language pairs, domains, corpora, and processing methods. And the relative significance of such dimensions may turn on a complicated balance of factors including overall performance, computational efficiency and cost, as well as explanatory transparency.[59] On this approach, rather than being a discrete parameter or feature, meaningfulness may be a continuous and multidimensional phenomenon. But we must leave the matter here.

References

Ron Artstein. *Parts of Words: Compositional Semantics for Prosodic Constituents*. PhD thesis, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, 2002.

Ron Artstein. Focus Below the Word Level. *Natural Language Semantics*, 12(1): 1–22, 2004. doi: 10.1023/B:NALS.0000011145.76954.0b.

Ron Artstein. Coordination of Parts of Words. *Lingua*, 115(4):359–393, 2005. doi: 10.1016/j.lingua.2003.09.007.

Duygu Ataman. *Learning Morphology for Open-Vocabulary Neural Machine Translation*. PhD thesis, Università degli Studi di Trento, Trento, 2019.

the recent wave of young and excitable deep-learning researchers moved in, showed improvements through their methods, declared success, and moved to bigger and better things" (Koehn 2020, 29).

[59] I note a convergence with broadly similar lessons Baroni draws from his recent review of generalization and compositionality in deep learning-based NLP: "Language ... is ... host to plenty of productive phenomena that obey less systematic, fuzzier laws, ranging from phonologically driven generalizations of irregular inflections [...], to partial semantic transparency in derivational morphology [...], to semi-lexicalized constraints in syntax [...], to the early stages of grammaticalization in language change [...]. Progress in understanding the linguistic capabilities of neural networks might help us to make precise predictions about the origin, scope and mechanics of these phenomena, and ultimately to develop a more encompassing account of the amazing productivity and malleability of human language" (Baroni 2019, 6).

Duygu Ataman and Marcello Federico. An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–110, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL https://aclanthology.org/W18-1810.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473*, 2014. URL https://arxiv.org/abs/1409.0473.

Nikolay Banar, Walter Daelemans, and Mike Kestemont. Character-Level Transformer-Based Neural Machine Translation. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, NLPIR 2020, pages 149–156, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3443279.3443310. URL https://doi.org/10.1145/3443279.3443310.

Tamali Banerjee and Pushpak Bhattacharyya. Meaningless yet Meaningful: Morphology Grounded Subword-Level NMT. In *Proceedings of the Second Workshop on Subword/Character Level Models*, pages 55–60, New Orleans, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1207. URL https://aclanthology.org/W18-1207.

Marco Baroni. Linguistic Generalization and Compositionality in Modern Artificial Neural Networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190307, 2019. doi: 10.1098/rstb.2019.0307.

Jan A. Botha and Phil Blunsom. Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1899–II–1907. JMLR.org, 2014.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. Revisiting Character-Based Neural Machine Translation with Capacity and Compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1461. URL https://aclanthology.org/D18-1461.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. A Character-Level Decoder without Explicit Segmentation for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1160. URL https://aclanthology.org/P16-1160.

Marta R. Costa-jussà and José A. R. Fonollosa. Character-Based Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2058. URL `https://aclanthology.org/P16-2058`.

Marta R. Costa-jussà, Carlos Escolano, and José A. R. Fonollosa. Byte-Based Neural Machine Translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 154–158, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4123. URL `https://aclanthology.org/W17-4123`.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.286. URL `https://aclanthology.org/2022.acl-long.286`.

Hope Dawson and Michael Phelan, editors. *Language Files: Materials for an Introduction to Language and Linguistics.* The Ohio State University Press, Columbus, 12th edition, 2016.

Josh Dever. Compositionality as Methodology. *Linguistics and Philosophy*, 22(3): 311–326, 1999. doi: 10.1023/A:1005410301126.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. One Size Does Not Fit All: Comparing NMT Representations of Different Granularities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1504–1516, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1154. URL `https://aclanthology.org/N19-1154`.

Gottlob Frege. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50, 1892.

Yingqiang Gao, Nikola I. Nikolov, Yuhuang Hu, and Richard H.R. Hahnloser. Character-Level Translation with Self-Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1591–1604, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.145. URL `https://aclanthology.org/2020.acl-main.145`.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL `https://aclanthology.org/C14-1111`.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1108. URL `https://aclanthology.org/N18-1108`.

Rohit Gupta, Laurent Besacier, Marc Dymetman, and Matthias Gallé. Character-Based NMT with Transformer. *arXiv:1911.04997 [cs]*, 2019. URL `http://arxiv.org/abs/1911.04997`.

Martin Haspelmath. The Indeterminacy of Word Segmentation and the Nature of Morphology and Syntax. *Folia Linguistica*, 45(1):31–80, 2011. doi: 10.1515/flin.2011.002.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research*, 67:757–795, April 2020. doi: 10.1613/jair.1.11674. URL `https://jair.org/index.php/jair/article/view/11674`.

Roman Jakobson. On Linguistic Aspects of Translation. In Reuben Brower, editor, *On Translation*, pages 232–239. MIT Press, Cambridge, MA, 1959.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1001. URL `https://aclanthology.org/P15-1001`.

Dan Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, Upper Saddle River, N.J, 3rd ed. draft edition, 2022. URL `https://web.stanford.edu/~jurafsky/slp3/`.

David Kaplan. Quantifying in. *Synthese*, 19(1-2):178–214, 1969.

Ali Kazmi and Francis Jeffry Pelletier. Is Compositionality Formally Vacuous? *Linguistics and Philosophy*, 21(6):629–633, 1998. doi: 10.1023/A:1005388721969.

Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi. Zero-Shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 72–78. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-srw.11. URL `https://aclanthology.org/2020.acl-srw.11`.

Philipp Koehn. *Neural Machine Translation*. Cambridge University Press, 2020.

Julia Kreutzer and Artem Sokolov. Learning to Segment Inputs for NMT Favors Character-Level Processing. *CoRR*, abs/1810.01480, 2018. URL `http://arxiv.org/abs/1810.01480`.

Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL `https://aclanthology.org/P18-1007`.

Taku Kudo and John Richardson. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL `https://aclanthology.org/D18-2012`.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017. doi: 10.1162/tacl_a_00067. URL `https://aclanthology.org/Q17-1026`.

Jiahuan Li, Yutong Shen, Shujian Huang, Xinyu Dai, and Jiajun Chen. When is Char Better Than Subword: A Systematic Study of Segmentation Algorithms for Neural Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 543–549, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.69. URL `https://aclanthology.org/2021.acl-short.69`.

Jindřich Libovický and Alexander Fraser. Towards Reasonably-Sized Character-Level Transformer NMT by Finetuning Subword Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2572–2579, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.203. URL `https://aclanthology.org/2020.emnlp-main.203`.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. Character-Based Neural Machine Translation. *arXiv:1511.04586 [cs]*, 2015. URL `http://arxiv.org/abs/1511.04586`.

Tal Linzen and Marco Baroni. Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7(1):195–212, 2021. doi: 10.1146/annurev-linguistics-032020-051035.

Minh-Thang Luong and Christopher D. Manning. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1100. URL `https://aclanthology.org/P16-1100`.

Thang Luong, Richard Socher, and Christopher Manning. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `https://aclanthology.org/W13-3512`.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1002. URL `https://aclanthology.org/P15-1002`.

John J. McCarthy. Prosodic Structure and Expletive Infixation. *Language*, 58(3):574–590, 1982. doi: 10.2307/413849.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL `http://arxiv.org/abs/1301.3781`.

Ryan M. Nefdt. A Puzzle concerning Compositionality in Machines. *Minds and Machines*, 30(1):47–75, 2020. doi: 10.1007/s11023-020-09519-6.

Michael Nelson. Propositional Attitude Reports. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2022 edition, 2022. URL `https://plato.stanford.edu/archives/spr2022/entries/prop-attitude-reports/`.

Tommi Nieminen. OPUS-CAT: Desktop NMT with CAT Integration and Local Fine-Tuning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 288–294. Association for Computational Linguistics, 2021. URL `https://www.aclweb.org/anthology/2021.eacl-demos.34`.

John E. Ortega, Richard Castro Mamani, and Kyunghyun Cho. Neural Machine Translation with a Polysynthetic Low Resource Language. *Machine Translation*, 34(4):325–346, 2020. doi: 10.1007/s10590-020-09255-9.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

W. V. O. Quine. *Word and Object*. MIT Press, 1960.

Piers Rawling and Philip Wilson, editors. *The Routledge Handbook of Translation and Philosophy*. Routledge, London and New York, 2018.

Mats Rooth. A Theory of Focus Interpretation. *Natural Language Semantics*, 1(1): 75–116, 1992. doi: 10.1007/BF02342617.

Federica Scarpina and Sofia Tagini. The Stroop Color and Word Test. *Frontiers in Psychology*, 8, 2017. doi: 10.3389/fpsyg.2017.00557.

Mike Schuster and Kaisuke Nakajima. Japanese and Korean Voice Search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012. doi: 10.1109/ICASSP.2012.6289079.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://www.aclweb.org/anthology/P16-1162`.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Toolkit for Statistical Morphological Segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden, April 2014.

Association for Computational Linguistics. doi: 10.3115/v1/E14-2006. URL https://aclanthology.org/E14-2006.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Zoltán Gendler Szabó. Compositionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2020 edition, 2020. URL https://plato.stanford.edu/archives/fall2020/entries/compositionality/.

Clara Vania. *On Understanding Character-Level Models for Representing Morphology*. PhD thesis, University of Edinburgh, Edinburgh, 2020.

Clara Vania and Adam Lopez. From Characters to Words to in Between: Do We Capture Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1184. URL https://aclanthology.org/P17-1184.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Ludmila Veselovská. *A Course in English Morpho-Syntax: Syllabi for the Lectures: Examples and Exercises*. Univerzita Palackého, Olomouci, 2009.

Rachel Weissbrod. Meaning. In Piers Rawling and Philip Wilson, editors, *The Routledge Handbook of Translation and Philosophy*, pages 289–304. Routledge, London and New York, 2018.

Dag Westerståhl. On Mathematical Proofs of the Vacuity of Compositionality. *Linguistics and Philosophy*, 21(6):635–643, 1998. doi: 10.1023/A:1005401829598.

Su-Ling Yeh, Wei-Lun Chou, and Pokuan Ho. Lexical Processing of Chinese Sub-Character Components: Semantic Activation of Phonetic Radicals as Revealed by the Stroop Effect. *Scientific Reports*, 7(1):15782, 2017. doi: 10.1038/s41598-017-15536-w. URL http://www.nature.com/articles/s41598-017-15536-w.

Wlodek Zadrozny. From Compositional to Systematic Semantics. *Linguistics and Philosophy*, 17:329–342, 1994.

Longtu Zhang and Mamoru Komachi. Using Sub-Character Level Information for Neural Machine Translation of Logographic Languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(2):1–15, 2021. doi: 10.1145/3431727. URL `https://doi.org/10.1145/3431727`.