






Contrasting a semiotic conceptualization of translation with AI text production: The case of audio captioning

Riku Haapaniemi , Annamaria Mesaros , Manu Harju ,
Irene Martín-Morató , Maija Hirvonen 

Tampere University, Finland



ABSTRACT

Using a semiotically-informed material approach to the study of translation, this paper analyses an artificial intelligence (AI) system developed for automatic audio captioning (AAC), which is the automated production of written descriptions for non-lingual environmental sounds. Comparing human and AI text production processes against a semiotic framework suggests that AI uses computational methods to reach textual outcomes which humans arrive at through semiotic means. Our analysis of sound description examples produced by an AAC system makes it apparent that this distinction is useful in articulating the complex relationship between human and AI translation processes. Acknowledging the central role of semiotic meaning-construction in human text production and its arguable absence in AI computational processes allows for AI processes to be discussed under a translational framework, while still recognizing their fundamental differences from comparable human translation processes. Further, audio captioning provides a clear example of a translation task where non-lingual content must be considered on equal terms with lingual text, and our discussions illustrate how this can be achieved in computational and semiotic processes alike. Overall, this paper promotes a nuanced understanding of meaning in text production and suggests multiple fruitful points of convergence and divergence between translation theory and AI research.

Keywords: artificial intelligence, audio captioning, intersemiotic translation, natural language processing, semiotics

Primerjava semiotične konceptualizacije prevoda z besedilom, ki ga tvori UI: primer avdiopodnaslavljanja

IZVLEČEK

V prispevku je s stališča semiotične materialne smeri v prevodoslovju predstavljena analiza sistema umetne inteligence (UI), ki je bil razvit za pripravo avtomatskih avdiopodnapisov (automatic audio captioning oziroma AAC), tj. avtomatsko tvorjenih pisnih opisov nejezikovnih zvokov okolja. Na podlagi semiotične primerjave človeškega in umetno-inteligenčnega procesa tvorjenja besedila je mogoče sklepati, da UI uporablja računalniške metode za dosego besedilnega cilja, medtem ko človek besedilni cilj doseže s pomočjo semiotičnih sredstev. Analiza opisov zvokov, ki jih tvori

sistem AAC, pokaže, da je to razlikovanje koristno, če želimo ubesediti kompleksni odnos med človeškim in umetnointeligenčnim procesom prevajanja. Hkrati predstavlja avdiopodnaslavljanje očiten primer prevodne naloge, pri kateri je treba nejezikovno vsebino obravnavati na enak način kot jezikovno besedilo. Razprava pokaže, kako je mogoče to doseči tako v računalniškem kot v semiotičnem procesu. V prispevku je poudarjen pomen podrobnega razumevanja pomena pri tvorjenju besedila, hkrati pa je izpostavljenih več plodnih področji, kjer se teorija prevajanja in raziskovanje UI stikata in razhajata.

Ključne besede: umetna inteligenca, avdiopodnaslavljanje, intersemiotični prevod, obdelava naravnega jezika, semiotika

1. Introduction

Two major conceptual issues are currently inciting debate and discussion across the field of translation studies (TS). The first is related to the advent of neural machine translation, large language models, and other artificial intelligence (AI) solutions for natural language processing (NLP). These developments challenge TS to account for the increasingly high-quality and widespread production of translations by machines, and in doing so raise conceptual questions about what kinds of phenomena should be understood as translation – “is machine translation translation?,” as Kenny, do Carmo, and Nurminen (2022, 396–417; see also do Carmo, Kenny, and Nurminen 2022) put it. The second issue focuses on the question of whether the conceptualization of translation as a phenomenon should be based on linguistic information processing, or rather be developed within more generalized hermeneutic or semiotic frameworks (e.g. Venuti 2019; Bennett 2022; Zheng, Tyulenev, and Marais 2023) that would better account for different kinds of meaning and acknowledge the centrality of personal experience in the construction of meaning. Therefore, if TS is to take both of these issues seriously, it will need the ability to do two somewhat opposing things simultaneously: engage with instances of machine language generation as translation, and conceptualize translation as a decidedly non-mechanic and non-language-centred meaning-construction process.

In this conceptual paper, we seek to find some common ground between these two perspectives by applying a meaning-focused conceptualization of translation to the analysis of an AI text production process. Instead of asking whether machine translation is translation, or what translation is, we reorganize those questions into a discussion of how human and AI translation processes compare with each other against a semiotic understanding of translation. In order to articulate where human and AI text production processes diverge from each other and where they converge, we discuss the basic principles of AI text generation in terms of translation theory and contrast that discussion with a semiotic conceptualization of human translational processes. It

is argued here that while human translation is understood predominantly as a process predicated on the construction of *meaning* through semiotic means (following e.g. Haapaniemi 2024), AI text production appears as a computational process of identifying and replicating aspects of lingual *form*, not of meaning-construction (as also argued by e.g. Bender and Koller 2020) – at least, not as meaning is understood in the semiotic sense. To illustrate what this fundamental difference means in practice and in terms of translation, we take the process of *audio captioning* – the production of verbal descriptions for non-lingual environmental sounds – as an example of a translation process which invites a non-language-centred, semiotic conceptualization of translation when studied as a human process, but which AI can nevertheless conduct computationally.

2. Translational perspectives on the convergences and divergences between human and AI text production processes

2.1 Audio captioning as an example of AI text generation

AI text generation today comprises of a wealth of techniques which produce texts of many kinds (e.g. summarizations, analyses, and translations between languages and distinct modalities, such as image descriptions), commonly employing algorithmic NLP models which analyse and/or produce strings of natural language based on identified relationships between language fragments. Generative AI systems can utilize language as input or as output, and use architecture similar to NLP models to work in non-lingual modalities at both ends of the process. One example of an AI text generation process that involves both lingual and non-lingual modalities is automatic audio captioning (AAC). NLP models are developed for the purpose of AAC in order to create computational systems that recognize the most important sounds in acoustic environments and produce written descriptions of those sounds. These written *captions* produced by the NLP model are formulated as descriptive sentences similar to how a human would describe their perception of the sounds they hear. For example, an audio caption for a scene recorded at a busy airport could be “people are talking and somebody whistles in the distance” (Figure 1).



Figure 1. Scheme of an encoder-decoder system for AAC.

In contrast to many conventional methods of computational audio analysis, NLP-based AAC systems do not simply place sounds into pre-existing categories (as in e.g. Virtanen, Plumbley, and Ellis 2018) but describe them verbally in order to convey an acoustic scene. Further, NLP-based systems do not map auditive features to existing linguistic features (as e.g. spoken phonemes are mapped to written language strings in automatic speech recognition, see Mei et al. 2022) but generate original descriptions for each audio clip. The specific AAC system discussed in this paper is developed by the cross-disciplinary GUIDE research project of Tampere University, Finland (for more on the development of GUIDE's AAC system, see Martín-Morató, Harju, and Mesáros 2022).

As it involves the production of verbal content from non-verbal content, AAC can be treated as an example of *intersemiotic* machine translation (after Jakobson 1959). GUIDE's NLP model is one of many such AI systems developed for AAC, and these models often employ the same fundamental principles as interlingual machine translation systems (Martín-Morató, Harju, and Mesáros 2022). As a process, audio captioning combines non-lingual, multimodal and intersemiotic aspects of translation, raising the question of how different forms of expression relate to the meaning derived from them during translation; and AAC conducted by AI complicates this further by raising the question of how these complex translational phenomena relate to the computational processes of NLP systems. These aspects make audio captioning an especially interesting example for the discussion of the differences between how humans and AI systems translate.

NLP models require *training* to be able to produce captions similar to human outputs. AAC systems are trained on human-produced texts that correspond to its input audio sequences, which allows the NLP model to analyse and identify the dependencies between its inputs and expected output captions. In recent years, a few AAC datasets have been collected through various crowdsourcing solutions for training and testing purposes (e.g. AudioCaps, in Kim et al. 2019; Clotho, in Drossos, Lipping, and Virtanen 2020; MACS, in Martín-Morató and Mesáros 2021; see also Hodosh, Young, and Hockenmaier 2013 on the drawbacks of using crowdsourcing for gathering AI training data). When gathering training data for an AAC dataset, human annotators are asked to listen to a short audio clip and describe what they hear in one complete sentence (see the examples discussed in section 3.3; for more details on how data was gathered for specific datasets, see the sources mentioned above; for more details on the dataset analysed in this paper, see section 3.1). From these audio clips and their corresponding captions, the AAC system learns the mappings between the statistical features of the recordings' acoustic content and the language used to describe the scenes human listeners have identified from the recordings.

In AAC, the NLP model treats both the audio clip and its written caption as a sequence of statistical *tokens*, each of which has their own probable relationships to each other as well as to the specific sequences they make up in the system's inputs and outputs. The most common contemporary model for these sequence-to-sequence tasks is the *transformer* (Vaswani et al. 2017). Transformers involve an *encoder*, which transforms the input content (in this case, an audio file) into a sequence of tokens, and a *decoder*, which transforms the token sequence into corresponding output content (written language). Due to their ability to assess the relevance of individual tokens and their connections not just between the input and output sequence, but also within the same sequence – in translational terms, not just target language in relation to the source text, but also target text elements in relation to each other – transformer models have been deemed especially useful for machine translation tasks (as discussed in e.g. Vardasbi et al. 2023) and other NLP tasks that involve the production of a language sequence based on an input sequence, such as AAC output.

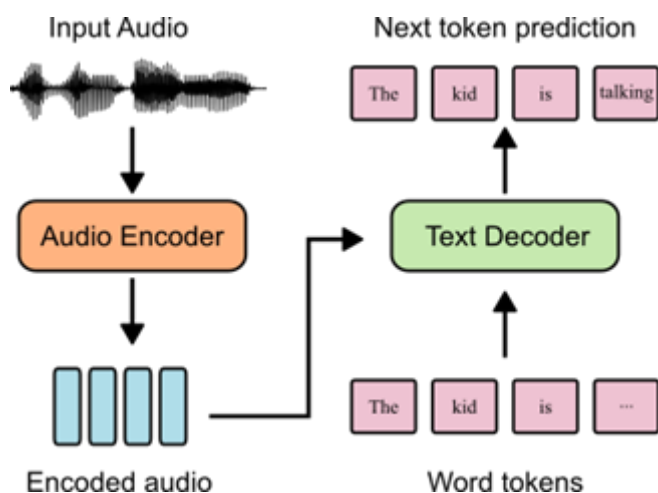


Figure 2. Encoder-decoder framework overview: the decoder predicts the next word token from the partly generated sentence and the encoded audio.

In processing its training data, GUIDE's AAC system therefore seeks to find the best alignment between the sequence of statistical features representing the audio input and the string of natural language which comprises human-produced captions (Figure 2). Over the course of its training, the NLP model's probability calculations are constantly weighed towards making more accurate predictions in order to consistently produce language-strings that are reasonably close to human-produced captions. One

part of this is the *evaluation* of the system's outputs in comparison to corresponding human text products. Evaluation metrics tend to focus on measuring the overlap between AI-generated captions and human-produced reference captions. There have been recent attempts to develop metrics specifically designed for evaluating AAC (e.g. CB-score, in Martín-Morató, Harju, and Mesaros 2022; FENSE, in Zhou et al. 2022; SPICE+, in Gontier, Serizel, and Cerisara 2023), but the most commonly used metrics are the ones that were originally designed for machine translation (e.g. BLEU, in Papineni et al. 2002) or image captioning (e.g. SPIDER, in Liu et al. 2017). What these metrics do not reveal, however, is whether the AI caption actually relates to the sounds heard in the recording in a meaningful way. For example, the BLEU or SPIDER metrics might give high ratings to sentences that just repeat the most common sentence fragments in the reference data, or ones that are structurally valid but semantically nonsensical. In other words, these metrics evaluate the NLP model's *formal* linguistic competence (i.e., its mastery of linguistic rules and patterns, Mahowald et al. 2023, 3) but not its *functional* linguistic competence (i.e., the ability to understand and use language to accomplish communicative goals, *ibid.*; see also the distinction made between the formal “deep syntax” of language and “communicative intent” in Bender and Koller 2020, 5192).

This difficulty in devising reliable evaluation metrics beyond assessing just the formal aspects of language reflects how computational systems relate to language and other human forms of expression. Transformer-based NLP models, like the one employed by GUIDE's AAC system, do not treat their inputs and outputs as meaningful expressions serving a communicative function, but as sequences of statistical relationships between tokens, regardless of what forms of expression or textual modes their inputs and outputs contain. Indeed, and as noted above, besides language processing transformers can be used for many kinds of data, from images (Radford et al. 2021) and video (Iashin and Rahtu 2020) to audio (Gontier, Serizel, and Cerisara 2021; Elizalde et al. 2023) – the same sequence-to-sequence methodology can be applied to any kind of digital input or output, since all digital data is reducible to sequences of values. In AAC and in other text production tasks, then, the fundamental operating principles of transformer-based AI systems are the same regardless of the modalities utilized: in all cases, the model maps the probable correspondences between sequences of input tokens and sequences of output tokens from training data. When the model is presented with a new input sequence, it predicts the most probable sequence of output tokens in response. At its core, AI text production is a process of *statistical modelling* and *probability calculation*; or, in other words, a process of comparing instances of lingual *form* in its training data to the formal aspects of its input and compiling a sequence of lingual form as output. This indicates a fundamental difference from the

semiotic conceptualization of human text production processes, including translation, which, as we will discuss below, are seen as being predicated on the construction of *meaning* – although some interesting parallels between the two processes can also be identified.

2.2 Form, meaning, and AI: Fundamental points of divergence

On the level of generalizations, it is easy enough to find common denominators between TS and NLP research. Various AI systems, developed with the ambition of creating more human-like machine intelligence (Bisk et al. 2020; Lake and Murphy 2021), are indeed capable of different kinds of translation tasks. These range from machine translation between natural languages (e.g. Sutskever, Vinyals, and Le 2014) to lingual descriptions of images, audio and video (e.g. Drossos, Adavanne, and Virtanen 2017; Mogadala, Kalimuthu, and Klakow 2019; Aafaq et al. 2020). Speaking in very general terms, it could be said that TS research and NLP development share an interest in questions of how the relevant features of an initial textual entity (source text or input) are retained or replicated in a subsequent textual entity (target text or output) with some recognizable relationship of correspondence to the initial entity. As noted, in both fields this interest extends from interlingual exchanges to instances of intersemiotic translation (Jakobson 1959) between different kinds of semiotic systems, such as between verbal and non-verbal forms of expression – as in audio captioning.

However, these parallels become more ambiguous when we move from generalizations to details and start to question terms such as “semiotic system” and their implications for the translation processes studied by TS. The *material approach* to translation (Haapaniemi 2024) is an example of a semiotically-informed theoretical framework that explores these implications. Based on the concept of materiality (Littau 2016) and a Peircean understanding of semiotics (Peirce 1994; Short 2007; see also e.g. Robinson 2016; Marais 2019; Sealey 2019), the material approach sees translation as the construction of meaning from a material source text, consisting of different kinds of semiotic *signs* and their material forms, and the subsequent compilation of the target text, which includes a new semiotic sign-complex to be communicated in new material forms to a new set of recipients in a new context of reception (Haapaniemi 2024, 30–33). The role of translators is therefore that of mediators between their interpretation of the source text in the source context and their perception of the target context, compiling a complex of signs that allows for the target text’s recipients to construct the desired kinds of meaning from it (Haapaniemi 2024, 30–33). Meaning-construction from texts is seen as a relational process that involves the recipient’s sensory perception of material forms, interpreting those forms as signs, and relating

the signs to their knowledge and experience of the world and of any communicational conventions concerning those signs (Haapaniemi 2024, 14–17; compare, for example, with the notion of “sense” discussed in Risku and Pircher 2008, 158–157, or the concept of “functional linguistic competence” in Mahowald et al. 2023). In contrast, as discussed in the previous section, AI text generation is fundamentally a process of calculating probabilities of correspondence between statistical patterns in its input and output (also discussed from a TS perspective in e.g. Asscher 2022, 4–5). This means that, in semiotic terms, AI operates on form where humans operate on meaning (as also proposed in Bender and Koller 2020, 5186–5188). The intermodal or interlinguistic computations in AAC – or in image captioning, video description, or machine translation – can find and produce correspondences between two or more instances of tokenized form (Bender and Koller 2020, 5192), but the meanings that humans derive from those forms (what is termed “communicative intents” in Bender and Koller 2020) are, arguably, beyond the current grasp of NLP (see also De Deyne et al. 2021; for arguments to the contrary, see e.g. Søgaaard 2023, 39–45).

If we accept that AI does not engage with form as signs, then AI does not construct meaning from that form, and the mode in which that form is presented is not a semiotic system to the AI – it is a dataset comprised of statistical relations between tokens, between aspects of form. In Peircean semiotics and in material translation theory, meaning is constructed through a holistic sign-process involving the perspective and experience of a specific semiotic actor. Meaning is therefore not just the established relationship between two forms of expression, or even the relationship between an expression and whatever that expression is considered to refer to; rather, in the semiotic conceptualization, meaning is ultimately specific to the individual semiotic actor constructing it, although still informed by and dependent on community, context, and convention (as discussed e.g. in terms of “icosis” from a Peircean perspective in Robinson 2016; for an overview of meaning as both a personal and social phenomenon in TS beyond explicitly semiotic theories, see e.g. Muñoz Martín and Rojo López 2018). It appears to be possible to computationally identify some aspects of the conventionalized relationship between an instance of language and the way in which language users are likely to utilize that instance of language for the purposes of meaning-construction (veering from the general sign-processes of semiotics to the language-focused area of semantics; for more on what kinds of semantic relations NLP can arguably learn from form alone, see e.g. Søgaaard 2023; for more on the specific quantitative kind of distributional semantics employed by NLP models, see e.g. Lenci and Sahlgren 2023). Nevertheless, meaning itself – how lingual forms are treated as signs and engaged with semiotically – is grounded in an individual language user’s situated experience of material reality and socio-cultural context (as argued

in e.g. Haapaniemi 2024; see also the similar distinction made between “sense” and “meaning” e.g. in Risku and Pircher 2008, or between “formal” and “functional” linguistic competence in Mahowald et al. 2023). Ultimately, this means that the meaning different communicators express through and derive from language is not inscribed or deposited in linguistic form, and therefore cannot be learnt from form alone by computational models. This view is supported by, among others, Bender and Koller (2020), who argue that the semantic similarities identifiable from linguistic form are “only a weak reflection of actual meaning” (Bender and Koller 2020, 5193).

2.3 Parallels between TS and AI studies: Points of convergence beyond fundamental differences

Using the semiotic terms of the material approach to the study of translation, the process of audio captioning when conducted by a human is as follows:

- The translator (annotator) receives a material text (audio recording).
- The translator identifies the relevant sign-complex in it (noteworthy sounds separated from noise and other textual or medial aspects deemed irrelevant).
- The translator constructs meaning from these signs (interprets what is making the sounds or conceives of other ways to describe them).
- The translator produces a new complex of signs based on the whole process (verbal language describing what was heard is used in a way the translator assumes will be understandable to the reader of the caption).

There is a precedent for the use of the material approach for the study of translational processes such as audio captioning: this framework has been previously employed in case studies focused on the meaningfulness of non-lingual auditory elements (Haapaniemi and Laakkonen 2019 on song translation) and on the production of informative texts (Haapaniemi 2023 on institutional communication). In other subfields of TS, translation from non-verbal signs to verbal signs is of a particular concern to studies in the multimodality of translation (see e.g. Kaindl 2013) and in audiovisual translation and media accessibility (see e.g. Remael, Reviere, and Vandekerckhove 2016), for example, in cases like audio description, when visual content is translated or interpreted into verbal descriptions for people with visual disability (e.g. Maszerowska, Matamala, and Orero 2014; Hirvonen and Saari 2024) or when written descriptions of non-verbal sounds are provided in subtitling for d/Deaf and hard of hearing audiences (e.g. Zárte 2021). Furthermore, a number of studies on

translations of natural sounds have utilized similar semiotic perspectives in recent years to highlight the role of non-lingual signs and non-human actors in different semiotic processes (e.g. Vihelmaa 2018; Sealey 2019; Taivalkoski-Shilov and Poncharal 2020). Analysing audio captioning in TS terms, and specifically from the semiotic perspective on translation as articulated by the material approach, is therefore a logical development.

The combination of material, non-lingual, and intersemiotic concerns with AI positions AAC at an interesting nexus point between semiotic conceptualizations of translation and the arguably non-semiotic mechanics of NLP models used in accomplishing AAC tasks. This, in turn, draws out parallels between TS and AI studies that go beyond the fundamental differences that a semiotic perspective sees between the human translation processes generally studied by TS and the computational processes discussed in AI studies. To start with, it is worth pointing out that some of the most foundational texts in TS speak of human translation processes in terms very similar to how computational language processing works. For example, Nida (1964, 146) considers translation a process of *decoding* and *encoding*, anticipating the structure of modern transformer models (Figures 1 and 2). In this sense, both fields share some ancestors in the information processing models of the 1940s (e.g. Shannon 1948). At the same time, it should be noted that while these classical definitions are still widespread within and especially outside TS (see e.g. Pym 2010, 19–20), in contemporary TS scholarship translation is usually seen not as transfer between codes, but as a more complex process predicated on factors like the translator's linguistic expertise, cultural considerations, and the different purposes of use for which translations are created (for a general overview, see e.g. Pym 2010; for how different TS approaches relate to NLP, see e.g. Asscher 2022; Asscher 2023). However, despite the nominal acknowledgement of other forms of expression in the concept of intersemiotic translation, classical understandings are sometimes argued to be fundamentally language-focused (Marais 2019, 19), and as such ill-suited for exploring translational processes that involve non-lingual content. In contrast, approaches that focus on materiality (e.g. Littau 2016), multimodality (e.g. Kaindl 2013; Ketola 2018; Tuominen, Jiménez Hurtado, and Ketola 2018) or experientiality (e.g. Campbell and Vidal 2024) incorporate various kinds of non-lingual meanings and meaningful relationships between lingual and non-lingual signs into the range of phenomena that can and must be considered in translation. On a more conceptual level, translation has been identified as a fundamental part of all forms of meaning-construction (Marais 2019) and as an aspect of various kinds of transformational processes (Robinson 2017; Blumczynski 2023).

Among these approaches seeking to redefine the concept of translation and widen the scope of its applicability, those informed by semiotics should be highlighted (for an overview, see e.g. Marais 2019, 28–30, 39–82). These approaches provide an alternative to language-focused conceptualizations by focusing on the basic principles of meaning-construction identifiable in all kinds of meaningful exchanges, lingual or otherwise – including translation from non-lingual to lingual systems, as in audio captioning. And yet, despite translation theories identifying semiotic processes as playing a central role in human translation while also identifying completely different kinds of processes as being the basis for AI text production, there are also significant similarities between the operating principles of NLP models and the proposed general semiotic mechanics underlying all different kinds of translational phenomena. The material approach, for one, sees translation as a generic process whose semiotic mechanics are fundamentally the same, regardless of the specific sign-system applied on the source or target side (Haapaniemi 2024, 24–27, 30–33); what varies is the way in which each text’s modalities and forms of expression enable the text’s recipient to construct meaning from them. On the level of mechanics, this is not completely different from how NLP models get from their input to their output. For one, AI is similarly agnostic in its approach to different sign systems or modalities: regardless of whether its input is linguistic, visual, auditive, or multimodal, the NLP model goes through the same process of identifying statistical patterns and calculating probabilities, and these patterns can just as well be converted into linguistic, visual or auditive output (see e.g. Geng et al. 2022 on unified multimodal encoders for vision and text).

By conducting their translation tasks through token patterns and their probable relationships – the “interlingua” (Raley 2022, 35) which machines use to pivot between input and output sequences – NLP models create generic rules of dependency applicable to all kinds of input and output, similar to how semiotics seeks to identify the universal mechanics of sign-based activities undertaken by semiotic actors. Further, transformer models employ attention mechanisms (Bahdanau, Cho, and Bengio 2016) which assess the relevance of individual tokens in relation to the entire input and output sequences, enabling an effect that resembles contextual awareness: in a mass of language data containing multiple instances of lingual communication, the material realities and socio-cultural conventions affecting human text production are reflected as probabilistic relations between expressions and the communicative purposes for which they tend to be used (Søgaard 2023, 41–43). As suggested in the previous section, these relations can be identified and replicated by NLP models (Søgaard 2023, 41–43), but again, these relations are not in themselves meaning in the semiotic sense.

It appears, then, that the text production phenomena which TS and AI studies are interested in are outwardly similar, but the processes behind these phenomena are based on completely different principles. One is based on the construction of meaning by semiotic actors, whereas the other is computational and probabilistic. There are functional or mechanical similarities between these processes, but there are also fundamental differences in what kind of processes they are. And yet there are also further connections between the two processes despite these differences. For example, as illustrated by the discussions of the development of GUIDE's AAC system in section 2.1, NLP models are trained on data produced by humans and therefore by semiotic means; the evaluation of NLP output involves assessing whether it is meaningful in the expected way, i.e. whether it resembles corresponding textual products of human meaning-construction; and, finally, NLP systems and their outputs are utilized as part of human semiotic processes, in the interpretation and production of texts. As discussed earlier, GUIDE's AAC system is a great example of how interwoven and complicated the practical and the conceptual relationship between meaning-based human translation processes and form-based computational translation processes is. Next, we will analyse some captions produced by GUIDE's AAC system to illustrate how the similarities and differences between the two processes appear in practice, and to see if the combination of a semiotically-informed TS approach with an AI-based subject of analysis can help disentangle the knotty conceptual issues discussed above.

3. Analysis: Comparing GUIDE's AAC system's processes and products with semiotic text production

3.1 How this study utilizes AAC data to contrast semiotic and computational translation processes

In this conceptual paper, our main aim is to discuss the differences between how semiotic and computational text production processes relate to meaning. We therefore use GUIDE's AAC system's captions only as illuminating examples. As such, the following should not be taken as a systematic evaluation of the AAC system's output quality: the analysis is not evaluative in a quantitative sense, and the examples discussed are not necessarily representative of how the system operates in general. Rather, in this section our goal is to illustrate how the theoretical juxtapositions discussed in the previous section are manifested in real texts. To achieve this, we have picked examples that clearly exhibit the distinctions between human-produced and AI-generated texts which are in the focus of this study. As such, the following analysis makes no claims or

judgments about the quality or value of GUIDE's AAC system compared to other NLP models, or of AI language generation processes compared to human text production processes. More systematic analyses which contribute to the development, training, and evaluation of GUIDE's AAC system and other NLP models have been carried out elsewhere (Martín-Morató, Harju, and Mesáros 2022), and the analysis conducted here is intended purely as an extrapolation of the theoretical discussions above.

For the purposes of the analysis, we used a section of the Clotho NLP training dataset (Drossos, Lipping, and Virtanen 2020) that is specifically intended for testing AAC models. The full dataset from which our examples are picked consists of 1,045 audio files and a set of 1,045 caption predictions produced by GUIDE's AAC system, including the highest-rated prediction for each file according to the SPIDER caption quality evaluation metric (Liu et al. 2017), as well as several corresponding human-produced captions for each file, totalling over 5,000 captions with which the AI-generated captions can be compared. Both the human and AI captions were compiled based on the acoustic contents of the audio files alone, and neither human annotators or the AAC system had access to the file name or other contextual clues as to the scene or event described. The file names and captions are presented in Tables 1–3 exactly as they appear in the dataset.

It should also be noted that, as mentioned in section 2.1, collecting training data for NLP models for the purposes of AAC is difficult and relatively little of it is available. As a result, the outputs of AAC tasks tend to vary in quality more than the textual output of NLP models used in machine translation and other more conventional sequence-to-sequence tasks; the fundamental transformer architecture is the same, but the quality of the output is highly dependent on the volume and quality of the training data that is available (Wu et al. 2023). In the context of this study, however, this variance in quality is beneficial: as the examples discussed below illustrate, the difference between human and AI text production processes can be seen most clearly in the unintuitive, strange, or otherwise unsuccessful captions – not necessarily in formal terms like grammar or syntax, but in terms of what elements of the audio clip in the captions stand out as relevant or meaningful.

We begin the analysis by examining example captions that function well as descriptions of auditive scenes, i.e. captions where GUIDE's AAC system has successfully imitated the products of human meaning-construction processes through computational means. We then examine examples of captions which are not convincing imitations of the products of meaning-construction, and discuss how they reflect the non-semiotic nature of their production process. Finally, we compare human-produced and AI-produced captions for the same audio file and summarize how the differences

between the two processes are reflected in these examples and what that implies about the relationship of each process to form and meaning in texts.

3.2 Contrasting convincing and unconvincing predictions by AAC system

As already mentioned, NLP models generate verbal contents through statistical analysis and probability calculation. The strings of language that the AAC system's NLP transformer model produces as output in response to the acoustic contents of audio files are compiled as a prediction of the most probable response to the input sequence, with the prediction process being based on statistical patterns found in human-produced language data containing similar pairs of input and output sequences. In this way, the AAC system is often able to produce perfectly understandable captions that resemble those produced by human annotators (Table 1).¹

Table 1. Examples 1–4 of the AAC system's captions that resemble human-produced captions.

No.	Name of audio file	Automatic caption predicted by AAC system
1	Car Driving Interior.wav	a car is driving down the road with the wind blowing in the background.
2	Diesel Engine Rattle.wav	a machine is running at a constant speed.
3	Door Creaking 01.wav	a door creaks as it is opened and closed.
4	young artists.wav	people are talking in a large room with each other.

However, despite its ability to produce structurally convincing language, some of the AAC system's predictions look very foreign to a human reader (Table 2):

Table 2. Examples 5–10 of AAC system's captions that fail to resemble human-produced captions.

No.	Name of audio file	Caption predicted by AAC system
5	C Minor Chords Musical Soundscape.wav	a synthesizer is playing a synthesizer with a synthesizer.
6	cookieSheetWiping.wav	a door opens and closes, closes, and closes, and closes a door.
7	20160506_sharpening.02.wav	a person is using a cater to make a cat.
8	Garden chimes.wav	a person is playing a strimy sound in a comppppompompppppppppppppppppppppppppppppppppppppppass.
9	coffee.wav	a person is flipping a bag of wood with a knife.
10	Blade sharpening.wav	a person is hitting a metal container with a clock.

1 In Tables 1–3, any unconventional language, spelling or formatting errors, etc. are replicated from the dataset.

In Examples 5 and 6, the described events overlap and repeat in unintuitive ways, and the subjects and objects of actions are mixed up: the synthesizer plays another synthesizer with a third synthesizer (Example 5); the door is opened only once but closed multiple times, and it is another door doing it (Example 6). In Examples 7 and 8, the system has produced unconventional, even straightforwardly erroneous language that is of little help in describing the scene, such as the word “cater” (Example 7) or the words “strimy” and “compppompompppass” (Example 8). Examples 9 and 10 are perhaps most interesting in that they use understandable vocabulary and convey an image of a scene, but the scene they describe is bizarre – not something that a human would probably imagine in an honest attempt to describe a real auditive scene. Further, these scenes are not ones that a human would be able to derive from sound alone: how does the act of “flipping a bag of wood” with “a knife” sound distinct from flipping it with any other instrument (Example 9), and how is it possible to tell by ear that it is “a clock” that is hitting the “metal container” (Example 10)?

What makes Examples 1–4 (Table 1) convincing captions and Examples 5–10 (Table 2) less so is how they relate to human meaning-construction processes. Examples 1–4 successfully predict a textual composition that a human could conceivably produce in a similar communicative situation, because the scene they describe and the way they describe it are comparable to what could be arrived at through constructing meaning from signs perceived through sound. Examples 5–10, on the other hand, do not resemble the results of human meaning-construction – either the scene they describe or the way they describe it is strange. In these instances, the results of the AI’s computational process fail to convincingly imitate the results of a human meaning-construction process. And indeed, as noted in section 2, the goal in training, evaluating, and developing the AI system is to coach it to get better at predicting which kinds of results would be closest to human-produced texts, and thus to derive from form through statistical analysis the kind of response that would otherwise be arrived at through meaning-construction.

3.3 Comparing AAC system’s predictions with human-produced descriptions

Since comparison to human textual products is such an essential part of how NLP models are developed, it is worth taking a closer look at a specific example caption produced by the AAC system and comparing it to the human-produced captions used for evaluation (Table 3). This will allow for a more detailed discussion of how AI processes and produces language as form, in contrast to how humans interpret and

express meanings from and through linguistic form. Again, the difference between Example 10 and its evaluation pairs 11–14 is most simply explained as the consequence of the different processes that have preceded their creation.

Table 3. Human-produced descriptions (Examples 11–14) compared to AAC system’s caption prediction for the same file.

No.	Human annotators’ descriptions for audio file “Blade sharpening.wav”	AAC system’s prediction (Example 10)
11	A knife is scraped a dozen times across a sharpener	a person is hitting a metal container with a clock .
12	A knife is scraped against a sharpener over a dozen times.	
13	A person sharpens a knife, the blade is clicking on the sharpening surface of the tool repeatedly.	
14	A type of knife is being sharpened continuously.	

As discussed in section 2.2, the material approach provides one framework for describing the translation process that humans and other semiotic actors engage in, as the annotators who produced Examples 11–14 have done. According to this framework, translators interpret the source text by relating the signs they perceive in it to their experience of the source context and produce the target text by relating their interpretation to their experience of the target context. It is conceivable how this process of relating source-side experience to target-side experience results in the caption “A knife is scraped a dozen times across a sharpener” (Example 11) for the sound of metal coming repeatedly into contact with a rough surface, but how such a process would result in the caption “a person is hitting a metal container with a clock” (Example 10) is less conceivable. Based on human experience of living in a society and being familiar with how humans derive meaning from acoustic sounds and from verbal language, describing a sound as that of a knife being scraped across a sharpener makes sense; describing that sound as that of a clock hitting a metal container makes less sense, because the act being described is improbable in human society and because it requires making distinctions humans are unlikely to make based on sound alone. However, to an AI that is disconnected from that experience, the latter option is no less sensible than the first, as long as the language fragments the description is comprised of fit together into a probable sequence.

As shown in Examples 1–4, through probability calculation and statistical modelling the AAC system is able to produce passages of verbal text that could conceivably convey the desired message to a human reader. The flipside of this, as shown in Examples 5–10, is that the meanings that are derivable from these forms may just as well make no sense. This is because, from a semiotic perspective, no sense-making – in

other words, using situated experience to relate signs to personal interpretations and communicative conventions – was involved in their production. As an NLP model, the AAC system only deals with aspects of form: it segments the input sequence of acoustic content into smaller units, and calculates the probable correspondences between these units and the linguistic units that make up its output sequence. Through training and evaluation, during which the system's outputs are assessed by humans according to how convincing they are in terms of both form and meaning, these probability calculations may become accurate enough to predict and produce strings of linguistic form that are very similar to human-produced captions. However, as illustrated by Examples 11–14, human annotators' captions are the product of a completely different kind of text-production process: a semiotic process based on meaning-construction, on the reception and interpretation of material forms as signs and the communication of the interpreted meanings through linguistic form. From a semiotic perspective, computational text generation appears only as the analysis and manipulation of form in relation to form, essentially skipping the meaning-construction stage seen to be at the heart of the human translation process and jumping directly from input source sequence to output target sequence. As seen in Examples 5–10, sometimes this leads to target texts that suggest meanings human interpreters would not have derived from the source text; on the other hand, as seen in Examples 1–4, sometimes the end result is perfectly adequate for the purposes of human meaning-construction, despite the fact that no semiotic construction of meaning took place in the text's production.

4. Discussion

The theoretical discussions and analyses conducted in this paper suggest that a semiotic perspective brings out a stark contrast between human and AI translation processes. According to the semiotically-informed TS framework utilized above, human translators produce texts through a semiotic process of meaning-construction by perceiving material forms as signs and constructing meaning from them. Conversely, AI systems – such as GUIDE's AAC system – produce texts by identifying and reproducing patterns they find within form through statistical computation, not by engaging with form as signs, and as such may not be considered to engage in meaning-construction in the semiotic sense. This pair of statements has a number of implications for the fields of TS and AI studies in general, and suggests that TS may be able to offer some new perspectives to the study and development of AI systems.

The differences revealed between human and AI text production processes are certainly noteworthy in themselves. Recognizing the fundamental difference between

semiotic and computational processes helps identify and analyse the similarities and dissimilarities between how translation tasks are conceptualized in TS theory and how AI systems conduct those tasks. Acknowledging this difference allows TS and AI studies both to remain cognizant of the fact that even when AI is able to produce texts that are conducive to meaning-construction in much the same ways as human-produced texts are, the actual routes by which the two processes arrive at similar endpoints are wildly different. It seems that, in favourable conditions, NLP models can indeed emulate the products of semiotic processes without actually engaging in semiotic processes, which is itself a testament to their efficiency and complexity. However, the fact that the end products of AI and human text production processes can be very similar, or even identical, does not mean that the processes themselves are necessarily at all alike. Using semiotic theory as the lynchpin for the distinction between human and AI translation processes could provide a fruitful framework for a number of crossover studies between TS and AI studies. This distinction would allow either delving deeper into the material approach or other modern-day semiotic theories from an AI perspective, or going back to reassess early translation theories (e.g. Nida 1964) to see if some of those frameworks subsequently deemed too mechanical for the complexity of human translation processes might still have something to provide to the study of computational translation processes.

As important as it is to acknowledge the fundamental differences between meaning-focused semiotic text production and form-focused computational text generation, it is just as important to recognize the myriad connections between human and AI translation processes. For instance, it is worth noting that the computational process undertaken by AI and the semiotic process undertaken by a human translator inhabit a similar mediatory role between source input and target output. Further, even if AI is considered to arrive at its outputs in a fundamentally different manner than human translators, it is undeniable that human semiotic processes intertwine with AI development, training, and use, and in doing so affect its calculations. The NLP model of a generative AI system may not be a semiotic actor, but it can play the part of one, and in order to function in that role it is trained so that the results of its computational processes approximate the results of human semiotic processes as closely as possible. Producing similar results does not make the two processes the same, but it may make them analogous enough to make some established TS concepts useful for AI studies (as discussed in e.g. Krüger 2022; Asscher 2022; Jiménez-Crespo 2023). This crossover appeal would benefit TS, too: transformer models are the current state-of-the-art in generative AI, and if the operation of transformer-based models can be treated as analogous to the kinds of human translation tasks already studied by TS and if NLP tasks can therefore be incorporated within the

scope of TS inquiry, then TS is in a position to contribute to some of the timeliest discussions in all academic research, and perhaps even influence the development of future AI systems.

Finally, the semiotic conceptualization of the relationship between form and meaning utilized in this paper provides both TS and AI studies with novel perspectives on how the initial and subsequent textual units studied in both fields relate to one another. For TS scholars, the analysis conducted in this paper – concerning a subject that, as noted in the introduction, stands at an interesting nexus between two timely topics of academic discussion – suggests that a semiotic perspective may be fruitful in studying AI translation processes, perhaps allowing for some common ground in ongoing conceptual debates. For AI scholars, this combination of semiotics and TS concepts offers an alternative perspective to what meaning is in human translational text production processes (in contrast to the semantics- or linguistics-based understandings, for example, Lenci and Sahlgren 2023; Søgaaard 2023; Mahowald et al. 2023) and why that kind of meaning might be what makes those processes so different from comparable AI processes. Having such a perspective is important because the interaction between humans and AI is a fundamental component of AI development, and this interaction will likely continue to gain prominence in both translation research and practical translation work. A clear conceptualization of how language and meaning relate to each other, and how AI and humans both relate to them, will contribute to a realistic picture of how humans and AI systems conduct translational processes and identify the similarities and dissimilarities between these processes. As human and AI tasks continue to converge, it is essential to construct a deep understanding not only of how humans and AI systems relate, but also of how scholarship on human processes like translation relates to scholarship on comparable AI processes.

References

- Aafaq, Nayyer, Ajmal Mian, Wei Liu, Syed Gilani, and Mubarak Shah. 2020. "Video Description: A Survey of Methods, Datasets, and Evaluation Metrics." *ACM computing surveys* 52 (6): 1–37. <https://doi.org/10.1145/3355390>.
- Asscher, Omri. 2022. "The Explanatory Power of Descriptive Translation Studies in the Machine Translation Era." *Perspectives* (e-publication ahead of print): 1–17. <https://doi.org/10.1080/0907676X.2022.2136005>.
- Asscher, Omri. 2023. "The Position of Machine Translation in Translation Studies: A Definitional Approach." *Translation Spaces* 12 (2): 1–20. <https://doi.org/10.1075/ts.22035.ass>.

- Bahdanau, Dzmitry, KyungHyun Cho, and Yoshua Bengio. 2016. "Neural Machine Translation by Jointly Learning to Align and Translate." *arXiv* 1409.0473: 1–15. Accessed May 16, 2024. <https://arxiv.org/pdf/1409.0473>.
- Bender, Emily M., and Alexander Koller. 2020. "Climbing towards NLU: On meaning, form, and understanding in the age of data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5185–5198. Stroudsburg: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.463>.
- Bennett, Karen. 2022. "The unsustainable lightness of meaning: Reflections on the material turn in Translation Studies and its intradisciplinary implications." In *Recharting Territories: Intradisciplinarity in Translation Studies*, edited by Gisele Dionísio da Silva and Maura Radicioni, 49–73. Leuven: Leuven University Press. <https://doi.org/10.2307/j.ctv2q4b064.6>.
- Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, et al. 2020. "Experience Grounds Language." In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 8718–735. Association for Computational Linguistics (ACL). <https://doi.org/10.18653/v1/2020.emnlp-main.703>.
- Blumczynski, Piotr. 2023. *Experiencing Translationality. Material and Metaphorical Journeys*. London and New York: Routledge. <https://doi.org/10.4324/9781003382201>.
- Campbell, Madeleine, and Ricarda Vidal, eds. 2024. *The Experience of Translation: Materiality and Play in Experiential Translation*. London and New York: Routledge. <https://doi.org/10.4324/9781003462538>.
- De Deyne, Simon, Danielle J. Navarro, Guillem Collell, and Andrew Perfors. 2021. "Visual and Affective Multimodal Models of Word Meaning in Language and Mind." *Cognitive Science* 45 (1): 1–44. <https://doi.org/10.1111/cogs.12922>.
- Do Carmo, Félix, Dorothy Kenny, and Mary Nurminen. 2022. "Is machine translation translation? Exploring conceptualizations of translation in a digitally saturated world." Call for abstracts, special issue of *Translation Spaces*. Accessed April 12, 2024. https://benjamins.com/series/ts/ts_cfp.pdf.
- Drossos, Konstantinos, Sharath Adavanne, and Tuomas Virtanen. 2017. "Automated Audio Captioning with Recurrent Neural Networks." In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 374–78. The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/WASPAA.2017.8170058>.
- Drossos, Konstantinos, Samuel Lipping, and Tuomas Virtanen. 2020. "Clotho: An Audio Captioning Dataset." In *2020 IEEE International Conference on Acoustics,*

- Speech, and Signal Processing Proceedings, 736–40. The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICASSP40776.2020.9052990>.
- Elizalde, Benjamin, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. “CLAP Learning Audio Concepts from Natural Language Supervision.” In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICASSP49357.2023.10095889>.
- Geng, Xinyang, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. 2022. “Multimodal Masked Autoencoders Learn Transferable Representations.” *arXiv 2205.14204* (preprint): 1–15. <https://doi.org/10.48550/arxiv.2205.14204>.
- Gontier, Félix, Romain Serizel, and Christophe Cerisara. 2021. “Automated audio captioning by fine-tuning BART with audioset tags.” In *DCASE 2021 - 6th Workshop on Detection and Classification of Acoustic Scenes and Events*, 170–74. Accessed May 14, 2024. <https://inria.hal.science/hal-03522488/document>.
- Gontier, Félix, Romain Serizel, and Christophe Cerisara. 2023. “Spice+: Evaluation of Automatic Audio Captioning Systems with Pre-Trained Language Models.” In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICASSP49357.2023.10097021>.
- Haapaniemi, Riku. 2023. “How production and distribution processes shape translations in organisations: A material perspective.” *Translation Spaces* 12 (1): 74–96. <https://doi.org/10.1075/ts.22038.haa>.
- Haapaniemi, Riku. 2024. “Translation as meaning-construction under co-textual and contextual constraints: A model for a material approach to translation.” *Translation Studies* 17 (1): 20–36. <https://doi.org/10.1080/14781700.2022.2147988>.
- Haapaniemi, Riku, and Emma Laakkonen. 2019. “The materiality of music: Interplay of lyrics and melody in song translation.” *Translation Matters* 1 (2): 62–75. https://doi.org/10.21747/21844585/tm1_2a4.
- Hirvonen, Maija, and Betta Saari. 2024. “Scripted or spontaneous? Two approaches to audio describing visual art in museums.” *Perspectives* 32 (1): 76–99. <https://doi.org/10.1080/0907676X.2022.2046816>.
- Hodosh, Micah, Peter Young, and Julia Hockenmaier. 2013. “Framing image description as a ranking task: data, models and evaluation metrics.” *The Journal of Artificial Intelligence Research* 47 (1): 853–99. <https://doi.org/10.1613/jair.3994>.
- Iashin, Vladimir, and Esa Rahtu. 2020. “A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-Modal Transformer.” In *31st British Machine Vision Conference 2020, BMVC 2020*, 1–16. BMVA Press. <https://doi.org/10.1109/CVPRW50498.2020.00487>.

- Jakobson, Roman. 1959. "On Linguistic Aspects of Translation." In *On Translation*, edited by Reuben Arthur Brower, 232–39. New York: Oxford University Press. <https://doi.org/10.4159/harvard.9780674731615.c18>.
- Jiménez-Crespo, Miguel A. 2023. "'Translationese' (and 'post-editese'?): no more: on importing fuzzy conceptual tools from Translation Studies in MT research." In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, edited by Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, et al., 261–68. Tampere: European Association for Machine Translation.
- Kaindl, Klaus. 2013. "Multimodality and Translation." In *The Routledge Handbook of Translation Studies*, edited by Carmen Millán and Francesca Bartrina, 257–69. London and New York: Routledge.
- Kenny, Dorothy, Félix do Carmo, and Mary Nurminen. 2022. "Is Machine Translation Translation?" In *EST Congress 2022, Abstracts*: 396–417. Accessed May 14, 2024. <https://biblio.ugent.be/publication/01GQEW3CA4FBPJ63GNGQ9JTW0>.
- Ketola, Anne. 2018. *Word-Image Interaction in Technical Translation: Students Translating an Illustrated Text*. Tampere: Tampere University Press.
- Kim, Chris Dongjoo, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. "Audiocaps: Generating Captions for Audios in the Wild." 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)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 119–132, Minneapolis, Minnesota. Association for Computational Linguistics. Accessed May 14, 2024. <https://aclanthology.org/N19-1000.pdf>.
- Krüger, Ralph. 2022. "Some Translation Studies informed suggestions for further balancing methodologies for machine translation quality evaluation." *Translation Spaces* 11 (2): 213–33. <https://doi.org/10.1075/ts.21026.kru>.
- Lake, Brenden M., and Gregory L. Murphy. 2021. "Word meanings in minds and machines." *Psychological Review* 130 (2): 1–31. <https://doi.org/10.1037/rev0000297>.
- Lenci, Alessandro, and Magnus Sahlgren. 2023. *Distributional Semantics*. Cambridge University Press. <https://doi.org/10.1017/9780511783692>.
- Littau, Karin. 2016. "Translation and the Materialities of Communication." *Translation Studies* 9 (1): 82–96. <https://doi.org/10.1080/14781700.2015.1063449>.
- Liu, Siqi, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. "Improved Image Captioning via Policy Gradient Optimization of SPIDeR." In *Proceedings 2017 IEEE International Conference on Computer Vision*, 873–81. The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICCV.2017.100>.

- Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. "Dissociating language and thought in large language models: a cognitive perspective." *arXiv* 2301.06627 (preprint): 1–45. <https://doi.org/10.1016/j.tics.2024.01.011>.
- Marais, Kobus. 2019. *A (Bio)Semiotic Theory of Translation: The Emergence of Social-Cultural Reality*. New York and London: Routledge. <https://doi.org/10.4324/9781315142319>.
- Martín-Morató, Irene, and Annamaria Mesaros. 2021. "Diversity and bias in audio captioning datasets." In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, edited by Frederic Font, Annamaria Mesaros, Daniel P.W. Ellis, Eduardo Fonseca, Magdalena Fuentes, and Benjamin Elizalde, 90–94. <https://doi.org/10.5281/zenodo.5770113>.
- Martín-Morató, Irene, Manu Harju, and Annamaria Mesaros. 2022. "A summarization approach to evaluating audio captioning." In *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, edited by Mathieu Lagrange, Annamaria Mesaros, Thomas Pellegrini, Gaël Richard, Romain Serizel, and Dan Stowell, 116–120. Accessed May 14, 2024. https://dcase.community/documents/workshop2022/proceedings/DCASE2022Workshop_Martin-Morato_35.pdf.
- Maszerowska, Anna, Anna Matamala, and Pilar Orero, eds. 2014. *Audio Description. New Perspectives Illustrated*. Amsterdam and Philadelphia: John Benjamins. <https://doi.org/10.1075/btl.112>.
- Mei, Xinhao, Xubo Liu, Mark D. Plumbley, and Wenwu Wang. 2022. "Automated audio captioning: an overview of recent progress and new challenges." *EURASIP Journal on Audio, Speech and Music Processing* 2022 (1): 1–18. <https://doi.org/10.1186/s13636-022-00259-2>.
- Mogadala, Aditya, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. "Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods." *The Journal of Artificial Intelligence Research* 71: 1183–317. <https://doi.org/10.1613/jair.1.11688>.
- Muñoz Martín, Ricardo, and Ana María Rojo López. 2018. "Meaning." In *The Routledge Handbook of Translation and Culture*, edited by Sue-Ann Harding and Ovidi Carbonell Cortés, 61–78. London and New York: Routledge. <https://doi.org/10.4324/9781315670898-4>.
- Nida, Eugene. 1964. *Toward a Science of Translating: with Special Reference to Principles and Procedures Involved in Bible Translating*. Leiden: Brill. <https://doi.org/10.1163/9789004495746>.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the Annual*

- Meeting of the Association for Computational Linguistics (ACL)*, 311–18. Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073135>.
- Peirce, Charles Sanders. 1994. *The Collected Papers of Charles Sanders Peirce*. Charlottesville: InteleX.
- Pym, Anthony. 2010. *Exploring Translation Theories*. London and New York: Routledge. <https://doi.org/10.4324/9780203869291>.
- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. “Learning Transferable Visual Models From Natural Language Supervision.” In *Proceedings of Machine Learning Research* 139, edited by Marina Meila and Tong Zhang, 8748–763. Accessed May 16, 2024. <https://proceedings.mlr.press/v139/radford21a/radford21a.pdf>.
- Raley, Rita. 2022. “Translation ‘degree zero.’” In *Time, Space, Matter in Translation*, edited by Pamela Beattie, Simona Bertacco, and Tatjana Soldat-Jaffe, 33–38. London and New York: Routledge. <https://doi.org/10.4324/9781003259732-4>.
- Remael, Aline, Nina Reviere, and Reinhold Vandkerckhove. 2016. “From Translation Studies and Audiovisual Translation to Media Accessibility: Some Research Trends.” *Target* 28 (2): 248–60. <https://doi.org/10.1075/target.28.2.06rem>.
- Risku, Hanna, and Richard Pircher. 2008. “Visual Aspects of Intercultural Technical Communication: A Cognitive Scientific and Semiotic Point of View.” *Meta* 53 (1): 154–66. <https://doi.org/10.7202/017980ar>.
- Robinson, Douglas. 2016. *Semiotranslating Peirce*. Tartu: University of Tartu Press.
- Robinson, Douglas. 2017. *Translationality: Essays in the Translational-Medical Humanities*. London and New York: Routledge. <https://doi.org/10.4324/9781315191034>.
- Sealey, Allison. 2019. “Translation: A Biosemiotic/more-Than-Human Perspective.” *Target* 31 (3): 305–27. <https://doi.org/10.1075/target.18099.sea>.
- Shannon, Claude E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27: 379–423, 623–656. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- Short, Thomas L. 2007. *Peirce’s Theory of Signs*. Cambridge: Cambridge University Press.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. “Sequence to Sequence Learning with Neural Networks.” In *Advances in Neural Information Processing Systems 27*, edited by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, 3104–112. Neural Information Processing Systems Foundation (NeurIPS). Accessed May 16, 2024. <https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf>.
- Søgaard, Anders. 2023. “Grounding the Vector Space of an Octopus: Word Meaning from Raw Text.” *Minds and Machines* 33 (1): 33–54. <https://doi.org/10.1007/s11023-023-09622-4>.

- Taivalkoski-Shilov, Kristiina, and Bruno Poncharal. 2020. *Translating the Voices of Nature/Traduire Les Voix de La Nature*. Montreal: Éditions québécoises de l'oeuvre.
- Tuominen, Tiina, Catalina Jiménez Hurtado, and Anne Ketola. 2018. "Why Methods Matter: Approaching Multimodality in Translation Research." *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 17: 1–21. <https://doi.org/10.52034/lanstts.v17i0.522>.
- Vardasbi, Ali, Telmo Pessoa Pires, Robin M. Schmidt, and Stephan Peitz. 2023. "State Spaces Aren't Enough: Machine Translation Needs Attention. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, edited by Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, et al., 205–16. Tampere: European Association for Machine Translation. Accessed May 16, 2024. <https://aclanthology.org/2023.eamt-1.20.pdf>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems* 30, edited by Ulrike von Luxburg, Samy Bengio, Rob Fergus, Roman Garnett, Isabelle Guyon, Hanna Wallach, and S.V.N. Vishwanathan, 5999–6009. Neural Information Processing Systems Foundation, Inc. (NeurIPS). Accessed May 16, 2024. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c-1c4a845aa-Paper.pdf.
- Venuti, Lawrence. 2019. *Contra Instrumentalism: A Translation Polemic*. Boston: University of Nebraska Press. <https://doi.org/10.2307/j.ctvgc62bf>.
- Vihelmaa, Ella. 2018. "Kielen kääntöpuolella. Kuinka tutkia toislaajisten merkkien kääntymistä ihmiskielelle?" [On the animal side of language. How to study the translation of nonhuman signs into human language?]. Licentiate thesis. Joensuu: University of Eastern Finland. Accessed May 16, 2024. https://erepo.uef.fi/bitstream/handle/123456789/21484/urn_nbn_fi_uef-20190920.pdf?sequence=1.
- Virtanen, Tuomas, Mark D. Plumbley, and Dan Ellis, eds. 2018. *Computational analysis of sound scenes and events*. Cham: Springer International Publishing.
- Wu, Ho-Hsiang, Oriol Nieto, Juan Pablo Bello, and Justin Salamon. 2023. "Audio-Text Models Do Not Yet Leverage Natural Language." In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICASSP49357.2023.10097117>.
- Zárate, Soledad. 2021. *Captioning and Subtitling for d/Deaf and Hard of Hearing Audiences*. London: UCL Press. <https://doi.org/10.2307/j.ctv14t478b>.

Zheng, Bingham, Sergey Tyulenev, and Kobus Marais. 2023. “Introduction: (re-)conceptualizing translation in translation studies.” *Translation Studies* 16 (2): 167–177. <https://doi.org/10.1080/14781700.2023.2207577>.

Zhou, Zelin, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q. Zhu. 2022. “Can Audio Captions Be Evaluated With Image Caption Metrics?.” In *2022 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, 981–85. The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICASSP43922.2022.9746427>.

About the authors

Riku Haapaniemi is a doctoral researcher at Tampere University, Finland. His research concentrates on the concept of materiality in translation studies, its philosophical and ontological implications, its applications in practical translation analysis, and its connections to research in other fields, including semiotics, textual studies, and language technology development. He is also a professional translator and involved in a number of research groups, academic associations, and professional organizations.

Annamaria Mesaros is an Associate Professor at Tampere University. She received her PhD in Signal Processing at Tampere University of Technology in 2012. Her research focuses on computational sound scene analysis, with over 50 scientific publications on this topic and many datasets. She is the coordinator of the evaluation challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), vice-chair of the DCASE Steering Group, member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE Signal Processing Society, and Senior member of the IEEE. She is currently an Academy of Finland Research Fellow for “Teaching Machines to Listen”.

Manu Harju is a PhD student in the Machine listening group at Tampere University. His background is in signal processing and machine learning, and the focus of his current work is on detection and classification of acoustic events.

Irene Martín-Morató received her PhD degree in information technology, communications, and computing in 2019 under the University Faculty Training Programme (FPU) from the Universitat de València, from where she also received her bachelor’s degree (with Honours) and her M.Sc. in telecommunications, in 2014 and 2016 respectively. She is currently a postdoctoral research fellow at Tampere University (Finland). Her research interests lie in the field of acoustic signal processing, machine learning, and audio event detection and classification.

Maija Hirvonen is Professor of German language, culture and translation at the Faculty of Information Technology and Communication Sciences of Tampere University. She co-leads the Tampere Accessibility Unit and the Multimodality in Translation and Interpreting research group at the multidisciplinary research centre for languages and cultures. Hirvonen's research areas include accessibility (especially audio description), multimodality and intermodality in translation and interpreting, blind-sighted interaction, the interface of cognition and interaction, and language-based machine perception.