Language and Cognition (2026), 18, e51, 1–26
doi:10.1017/langcog.2026.10100

ARTICLE

Why AI is not (quite) the same as artificial intelligence: reconciling isomorphism and communicative efficiency in formal reduction

Natalia Levshina

Radboud University, Netherlands
Email: natalevs@gmail.com

(Received 12 January 2026 UTC; Revised 25 May 2026 UTC; Accepted 12 June 2026 UTC)

Abstract

In line with the principle of isomorphism, formally reduced expressions such as clippings or abbreviations develop distinct social and/or semantic functions. This article argues that such differentiation follows the principles of communicative efficiency, with accessibility playing a central role. The claim is supported by a corpus-based comparison of the full form a*rtificial intelligence* and the abbreviated form *AI* in online news and Reddit comments. The short form is used far more frequently in the informal Reddit comments, which rely on a rich common ground, than in the news data. The two forms also display distinct semantic and grammatical profiles. Using distributional semantics (word and sentence embeddings) and Universal Dependencies, I show that the division of semantic and grammatical 'labor' between the forms is efficient. The abbreviated form tends to express more accessible meanings related to individual user experience, whereas the full form is more strongly linked to less accessible, abstract meanings of AI as a scientific field, industrial sector, technology or machine capability. In addition, the abbreviated form more often serves as the first element of compounds, whereas the full form often functions as a prepositional modifier, in line with principles of efficient word order.

Keywords: artificial intelligence; communicative efficiency; distributional semantics; principle of isomorphism

1. Theoretical background

This study examines the interplay between the principles of isomorphism and communicative efficiency. The principle of isomorphism, which is closely related to the Principle of Contrast (Clark, 1987) in language acquisition and the Principle of No Synonymy (Goldberg, 1995) in Construction Grammar, holds that two formally distinct forms should also differ in function (Haiman, 1980). These differences are

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

not limited to semantics in a narrow sense, but may also be pragmatic, stylistic and sociolinguistic in nature (cf. Goldberg, 1995; Leclercq & Morin, 2023).

Communicative efficiency, in turn, means minimization of the cost-to-benefit ratio in communication. A key principle of efficient language use is a negative correlation between accessibility and costs (Levshina, 2022): the more accessible a meaning is – due to immediate context, encyclopedic knowledge, entrenchment, perceptual salience or other factors – the lower its production cost should be. Lower costs often translate into reduced articulatory effort through formal reduction (Bybee, 2010; Levshina, 2022; Zipf, 1965[1935]), which can result in situations where both the full and the reduced variants become conventionalized and entrenched within a speech community. This typically occurs in the following cases:

- lexical replacement: an existing longer expression is replaced with a shorter one that is already available, as in *automobile* > *car* (Zipf, 1965[1935]: 34);
- lexical reduction, including clippings (e.g., *mathematics* > *maths*, *telephone* > *phone*), acronyms (e.g., *NATO*) and initialisms (*Artificial Intelligence* > *AI*);
- phonological reduction: shortening, loss of articulation detail and chunking of words and phrases, for example, *I do not know* > *dunno, going to* > *gonna* (Bybee, 2010);
- omission of words or morphemes, leading to the emergence of constructional variants. For instance, fitness enthusiasts often omit the object of the verb *lift* (weights) in Reddit interactions (Glass, 2020), resulting in variation between *lift weights* and simply *lift*.

Such situations appear to violate the principle of isomorphism: two distinct forms (the full and reduced form) are used to express the same meaning. Language has several ways of resolving this apparent conflict. One possibility is that the full form is completely ousted, as in *brandy* (< *brandywine*). Another is semantic differentiation, whereby the forms develop distinct meanings (e.g., *miss* vs. *mistress* or *cab* vs. *cabriolet*) (Marchand, 1960: 363). A further solution involves stylistic and sociolinguistic specialization, such that the variants are no longer interchangeable in particular social contexts. For example, the contracted form *wanna* is preferred by younger speakers and in informal speech, whereas the full variant *want to* is more common in more formal contexts (Levshina & Lorenz, 2022). Similarly, Jacobs and MacDonald (2023) show that language users prefer the original and clipped forms in pairs like *chimp* – *chimpanzee* in systematically different contexts, which can be interpreted as reflecting register or style differences.

Although semantic differences, social (pragmatic, stylistic, sociolinguistic) factors and efficiency are usually considered separate determinants of the choice between full and reduced forms (cf. Hilpert et al., 2023), this article argues for a unified theoretical account. Specifically, it proposes that both social and semantic differentiation are driven by efficiency. Central to this process is the same factor that triggers formal reduction in the first place, namely, the accessibility of meaning.

Consider the pair *TV* – *television* as an example. The full form is more likely to occur in formal contexts, whereas *TV* is more informal. This is a very common pattern. Hilpert et al. (2023), in an analysis of 50 pairs of clippings and original forms like *admin* and *administration*, found that the clipped forms tend to co-occur with

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

linguistic features characteristic of involved text production, such as privative verbs, contractions and first-person pronouns (Biber, 1988). This finding suggests that reduced variants are preferred in contexts characterized by a rich common ground. A natural explanation is that common ground increases the accessibility of referents, making the use of reduced forms communicatively efficient. In addition, informal interaction can lower the threshold for the creation and uptake of nonconventional forms. They often emerge as a sign of familiarity when the meaning is highly accessible to members of a speech community (Plag, 2003).

Moreover, *TV* and *television* also differ semantically. The full form *television* typically denotes the medium (1a) or the industry (1b), whereas *TV* often refers to a physical device (1c). All examples are taken from the blog subcorpus of the COCA (Davies, 2008), with emphasis added:

(1) a. *Books will never replace* ***television****, and* ***television*** *will never replace books.*
b. ***Television*** *pundits have called her the most popular woman in America.*
c. *I can stream the PS3 to the Vita… but the* ***TV*** *goes dark when I do that.*

The meaning of TV as a physical device is more accessible to many language users because such devices are present in most households. In contrast, the medium and industry meanings are typically relevant only in professional or institutional contexts and therefore less immediately accessible. In addition, the short form *TV* also frequently occurs in compounds: *TV show*, *TV program*, *TV host*, *TV series*, which are aspects of everyday viewing experience. It is efficient to encode the more accessible meanings related to such routine, viewer-oriented experiences with the more economical expression.

Reduced forms are thus more likely to be used in contexts with rich common ground (most prominently, in informal interactions) and to express more accessible meanings than full forms. In the case of widely used technical devices like television, informal communication tends to focus on appliances and gadgets we use regularly and on our daily experiences with them. At the same time, common ground or informality does not preclude reference to technical concepts accessible mostly to specialists. On the contrary: speakers in expert communities routinely use reduced forms in their in-group communication (Zipf, 1965[1935]: 32). For instance, software developers use informal clippings such as *dev* < *developer* or *mod* < *moderator*, which refer to accessible referents within the field.

The present paper focuses on the full form *Artificial Intelligence* and the initialism *AI*. Despite its importance for our society (cf. Petricini, 2025), the term *Artificial Intelligence* has been notoriously vague since its origin in 1955, when John McCarthy wrote a funding proposal together with Marvin Minsky, Claude Shannon and Nathaniel Rochester for the now-famous Dartmouth workshop held in 1956. Rather than offering a definition, the proposal described several tasks that artificial intelligence was expected to accomplish: 'An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves' (McCarthy et al., 2006[1955]: 12). This vagueness has been interpreted in different ways: some view it as a productive feature that enabled the field's development (Stone et al., 2016), while others emphasize its instrumental function, arguing that 'since the start imprecise jargon was used to make exaggerated promises with the goal of pleasing investors, claims that fundamentally remain promises to this day' (Guest et al., 2025: 3).

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

To further complicate matters, our idea of what counts as Artificial Intelligence is constantly shifting. As Schank observed (Schank, 1991: 40), 'for a particular task, if no machine ever did it before, it must be AI'. Thus, technologies initially framed as AI often lose that status once they become familiar and a new AI milestone appears – 'an odd paradox,' as McCorduck (2004: 423) terms it. For example, optical character recognition was once regarded as an AI task but is no longer perceived as such. In recent years, ChatGPT and the likes have probably become the prototype of AI, combining several salient technologies and applications: large language models, generative AI, artificial neural nets and chatbots (cf. Guest et al., 2025; Rosch & Mervis, 1975). Yet, as users become accustomed to machines producing coherent text in response to prompts, a new AI milestone may emerge.

Since early publications, the initialism *AI* has served as a compact form, often introduced in parentheses after the full form and then used independently, as in the example below:

(2) *Nvidia's historic run was fueled by the rapid growth of the **artificial intelligence (AI)** market, which sparked brisk sales of its high-end data center GPUs for processing **AI** tasks.* [Leipzig Corpora Collection, English News corpus 2024]

This use is similar to that of anaphoric pronouns and other short expressions referring to an accessible referent previously introduced in the discourse (Ariel, 2001). However, as the corpus data below will show, the short form *AI* is also frequently used independently and has, so to speak, developed a life of its own. In some cases, speakers even explicitly contrast *AI* with *Artificial Intelligence*:

(3) *If you think **'AI'** (not actually **artificial intelligence**, just a marketing term for LLMs) is going to discover anything, you severely misunderstand what this software is.* [Reddit]

How different, then, are the two forms in contemporary usage? To answer this question, I use a distributional approach to semantics, which has been fruitfully applied to the study of shortened lexical forms (Hilpert et al., 2023; Jacobs & MacDonald, 2023; Zheng et al., 2019). The analysis is based on word and sentence embeddings derived from two corpora: online news and thematic subreddits dedicated to Artificial Intelligence. These corpora differ substantially in terms of formality and common ground. In addition, I compare the grammatical properties of *AI* and *Artificial Intelligence* using tools from the Universal Dependencies framework.

This article investigates two hypotheses based on the principle of a negative correlation between accessibility and communicative costs (expressed here as formal length). Here, I distinguish between two types of accessibility: accessibility arising from common ground, and accessibility stemming from the greater familiarity and relevance of specific senses and referents of *Artificial Intelligence* and *AI* in general. The hypotheses are as follows:

(1) Data from the subreddits dedicated to AI-related topics, which provide rich common ground, making the meanings of *Artificial Intelligence* and *AI* highly accessible, contain a higher proportion of the short form *AI* than the online news corpus;

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

(2) The short form *AI* is associated with more accessible meanings than the full form *Artificial Intelligence.*

Before turning to the corpus analysis, it is necessary to outline the main senses of *Artificial Intelligence/AI*, as they have been used in the professional literature. This is the aim of Section 2. Section 3 introduces the corpora. Section 4 presents the results of the analyses of the word and sentence embeddings, followed by an examination of the grammatical differences in Section 5. Finally, Section 6 summarizes the findings and discusses them from the perspective of communicative efficiency and the principle of isomorphism.

2. What is artificial intelligence/AI?

Although Artificial Intelligence (AI) is a very vague concept, several recurrent senses can nevertheless be identified. One prominent sense is the capability of a machine to mimic some aspects of natural (human or animal) intelligence, for example, the ability to learn:

(4) *I describe a program exhibiting **AI** as one that can change as a result of interactions with the user* (Schank, 1991: 38).

Given that intelligence itself is a highly underspecified notion, this sense has a range of interpretations, including the display of human-like behavior, the presence of structures analogous to the human brain, or the implementation of cognitive functions and principles associated with the human mind (Wang, 2019).

Another use of Artificial Intelligence/AI concerns the realization of this capability through computational methods, algorithms and supporting infrastructure. This usage has at least two aspects. The first is technology in a broad and abstract sense, referring to general approaches and methods for achieving intelligent behavior (e.g., *Deep learning is a core type of artificial intelligence*). The second concerns specific artifacts, such as software, tools, models and other products that instantiate these methods (e.g., *Students use AI for writing assignments; AI-generated images*). The meaning of AI referring to software products and computational technologies can also undergo a metonymic shift to denote AI companies or the industry as a whole, for example, *invest in AI*, *AI bubble.*

In addition, artificial intelligence (AI) is a field of science and engineering that aims at developing such technologies and understanding the nature of machine 'intelligence':

(5) *Artificial Intelligence is a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behavior, and with the creation of artifacts that exhibit such behavior* (Shapiro, 1992: 54).

The motif of human-like intelligent machines acting as independent agents, sometimes as antagonists of humans, has been a recurring theme in science fiction, exemplified by Skynet in the *Terminator* films or the manipulative android in *Ex Machina.* More recently, however, rapid technological advances have brought

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

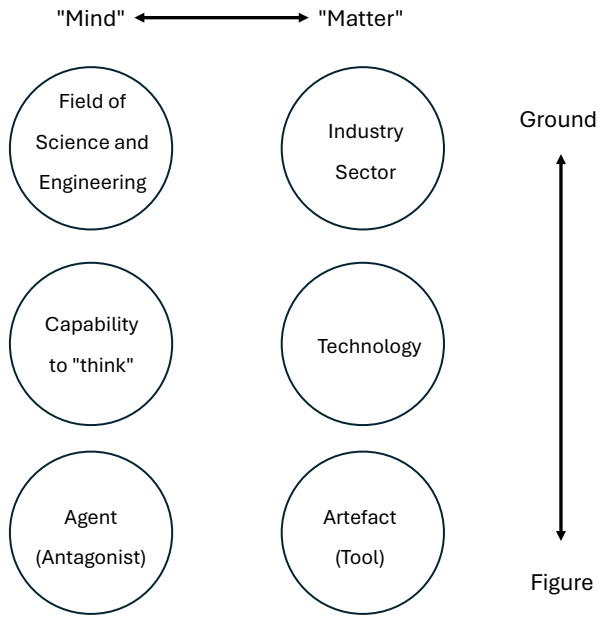


Figure 1. The main senses of Artificial Intelligence/AI.

concerns about AI outsmarting and potentially harming humans into mainstream public and academic discourse, as in the following example from Reddit:

(6) *Imagine if the AIs had to prioritize saving themselves versus other beings.*

All these senses can be organized along two conceptual dimensions, as illustrated in Figure 1: 'Mind' versus 'Matter,' and 'Figure' versus 'Ground'. The horizontal dimension contrasts academic approaches to AI, AI as 'intelligence' or a sentient being ('Mind') with technical implementations and artifacts that have a more direct impact on the material world ('Matter'). The vertical dimension, which is similar to the Figure–Ground distinction in Cognitive Semantics (Janda, 1996; Talmy, 2000), goes from less salient and individuated senses of AI as a field or a sector (often construed as metaphorical locations), to AI as a capability or technology (something one can possess, use or implement), and finally to the most salient and fully individuated referents (AI as a software artifact or an Agent).

In the sections that follow, I examine how these semantic dimensions and functions are reflected in language use as represented in the corpora.

3. Corpus data

3.1. Online news corpus

The online news corpus is part of the English component of the Leipzig Corpora Collection[1] (Goldhahn et al., 2012). The corpus consists of sentences extracted from

[1]Downloadable from https://wortschatz-leipzig.de/en/download/eng (last access 24.05.2026).

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

texts collected through daily web crawling of news portals and arranged in alphabetical order. For the present study, I used one million sentences from texts published in 2024, the most recent corpus available at the time of writing. The sentences contain approximately 20 million words.

The language in the corpus is quite formal. The sentences often contain complex noun phrases and nonfinite clauses with participles and infinitives, as below:

(7) *Advanced technologies in artificial intelligence continue to evolve, characterized by improvements in machine learning algorithms, increased data processing capabilities and enhanced computational power.*

An automatic, case-insensitive search using regular expressions yielded many instances of both the full and the short forms. References to company and product names (e.g., *Meta AI* and *Galaxy AI*), URLs and websites containing '.ai' were excluded. Manual inspection of a subset of the results confirmed that the string 'AI' nearly always referred to Artificial Intelligence. A single instance in which the abbreviation denoted artificial insemination of cattle was identified and removed. In total, 4,054 instances of the short form and 584 instances of the full form were retained for subsequent analyses. Despite the formal nature of the text type, the full form is remarkably infrequent, representing 12.6% of all instances.

3.2. Comments on Reddit

The second source of corpus data was Reddit. It is a platform for discussing different topics in so-called subreddits, which consist of discussion threads initiated by a post, followed by user comments. For this study, I analyzed two AI-related subreddits: *r/ArtificialIntelligence* and *r/artificial*. A total of 300 threads initiated in 2024 or 2025 were extracted. Automatically generated comments produced by the AutoModerator and other bots were removed. The resulting corpus contains approximately 31,000 comments, totaling approximately 1,444,500 words.

The language in this corpus is predominantly informal. Many comments contain emojis and emoticons, profanities, contractions, ellipsis and other features characteristic of informal computer-mediated communication, as illustrated in the examples below.

(8) a. *when ai needs therapy:-p*
b. *For an artificial intelligence sub you guys sure want AI to be a bubble lmao*
c. *How did you manage to give the AI OCD??* 😂
d. *Why is every post here written by ai goddamn.*

An automatic case-insensitive search using regular expressions yielded many relevant expressions. As in the online news corpus, occurrences of the string 'AI' in URLs and website names, as well as in company and product names (e.g., *Perplexity.ai*, *Meta AI* and misspelled *OpenAI*) were excluded. In total, 17,178 occurrences of the short form were identified. By contrast, the full form was extremely rare, with only 117 occurrences, accounting for less than 0.7% of all instances. Table 1 summarizes the frequencies of the full and short forms in both corpora.

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

Table 1. Occurrences of each form in the examined corpora

Corpus	Frequency of the short form	Frequency of the full form	Total frequency
Online news	4,054 (87.4%)	584 (12.6%)	4,638 (100%)
Reddit	17,178 (99.3%)	117 (0.7%)	17,295 (100%)

Overall, *AI* is the dominant form in the online news corpus, and the near-exclusive form on Reddit. These findings support the first hypothesis of the study: the high accessibility of the meanings represented by *Artificial Intelligence/AI*, resulting from the common ground shared by the users of the AI-related subreddits, leads to a much higher proportion of the short form in the Reddit corpus than in the online news.

4. A distributional semantic analysis of the short and full forms

4.1. Using sentence embeddings to identify distinctive contexts

Manual inspection of samples from the corpora suggests that the most prominent senses of *Artificial Intelligence/AI* are technology (especially generative AI) and software artifacts, especially chatbots based on large language models like ChatGPT and models for generating images and videos like Midjourney and SORA. However, assigning all instances to the sense categories outlined in Section 2 (Figure 1) proved difficult, even when rich contextual information from the original online news articles and the Reddit threads was available. This difficulty motivated the adoption of a bottom-up approach based on distributional semantics to identify the differences between the forms.

This section examines the distinctive role of the contexts in which the two forms occur. If two forms serve different functions, they should appear in different contexts – paraphrasing John Firth's (1957) famous dictum, they should 'keep company' with different linguistic units. If we can predict the occurrence of one form rather than the other from contextual features with greater accuracy than chance, then we can consider the contexts distinctive, and the forms can be said to have sufficiently differentiated functions.

The context was defined as the sentence in which a full or short form occurred. To obtain the context vectors, I used pre-trained sentence embeddings from the SBERT Python module (Reimers & Gurevych, 2019).

The approach is different from the traditional static type-based word embeddings (such as Word2Vec embeddings, which are discussed in Section 4.2) in that it also adds information about the position and context of every word, as Transformer-based models do. When an SBERT model is trained, the token-level vectors produced by BERT/RoBERTa are pooled (e.g., by averaging) to obtain fixed-size sentence embeddings. The Transformer weights are then fine-tuned so that the resulting sentence embeddings are semantically meaningful, with more similar sentences exhibiting higher cosine similarity. Once the model has been trained, it can be used to obtain an embedding for a new sentence by pooling the contextualized token vectors generated for that sentence. This approach combines context sensitivity with computational efficiency.

For the analysis presented below, I used the most up-to-date model all-mpnet-base-v2, which is based on Microsoft's MPNet-base model and fine-tuned on more

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

Table 2. Performance of conditional random forests in predicting the full and short forms. OOB: based on out-of-bag observations; AUC: Area Under the ROC Curve

Data source	*N* sentences	Masking method	OOB accuracy	OOB AUC
Online news	1160 (Full and short can co-occur)	Empty string	65.7%	0.700
		Nonce string	66.3%	0.713
	844 (Full and short cannot co-occur)	Empty string	62.9%	0.677
		Nonce string	63.4%	0.681
Reddit	220 (Full and short can co-occur)	Empty string	72.3%	0.769
		Nonce string	69.1%	0.773
	188 (Full and short cannot co-occur)	Empty string	64.9%	0.693
		Nonce string	66.5%	0.740

than one billion sentence pairs from various web sources to predict which sentences are semantically similar.[2] The model encodes a sentence or a short paragraph as a 768-dimensional vector.

An important methodological step was to ensure that information about the use of the full or short form was not encoded in the embeddings, in order to avoid circularity. This was done using two masking strategies: replacing the target forms with an empty string (i.e., removing them altogether), and replacing them with a nonsensical character string. As shown in Table 2, both strategies led to similar results.

Given the strong skew toward the short form, especially in the Reddit data, I included all available sentences with the full form and drew an equally sized random sample from the instances of the short form in each dataset. In addition, instances in which the target form occurred more than twice were excluded. As a result, the training data from the online news corpus consisted of 580 sentences with the full form and 580 randomly sampled sentences with the short form (1160 in total). For the Reddit corpus, 110 examples of the full form were matched by 110 randomly sampled sentences with the short form (220 in total).

To ensure that the presence of one variant did not facilitate the prediction of the other, I additionally selected sentences in which only the target form was present and the alternative form was absent. This step excluded, among others, appositional constructions introducing or clarifying the abbreviation, such as 'AI (Artificial Intelligence)' and 'Artificial Intelligence, or AI'. This procedure yielded an online news dataset containing 422 sentences with the full form and 422 randomly sampled sentences with the short form (844 in total). For the Reddit data, a parallel dataset was created with 188 sentences, consisting of 94 examples of each form.

In the next step, the sentence embeddings were created and used as input to a machine learning algorithm trained to predict whether a sentence originally contained the full or the short form. I employed conditional random forests implemented in the R package *party* (Hothorn et al., 2006), an ensemble machine learning algorithm for classification and regression which aggregates predictions from many conditional inference trees. In these trees, splits are chosen via statistical hypothesis tests, which typically obviates the need for pruning. Conditional inference trees and random forests have proven useful in a wide range of linguistic applications and are well suited to data with many predictors and relatively few observations (Levshina, 2020).

[2]See more details at https://www.sbert.net/ and https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (last access 24.05.2026).

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

An alternative approach would be to fine-tune BERT or SBERT rather than using pretrained embeddings. However, this would require substantially larger amounts of data and could increase the risk of overfitting, given the relatively small size of the datasets.

Each random forest consisted of 2000 conditional inference trees. For each split, 30 dimensions were sampled randomly from the 768 SBERT dimensions as candidate predictors. Other reasonable hyperparameters were tested and yielded results very similar to the ones reported below. All remaining hyperparameters were left at their default values (see the *cforest()* function).

The machine learning results are displayed in Table 2, separately for each sampling and masking method. Accuracy is defined as the proportion of correct predictions made by the random forests. AUC stands for the area under the ROC curve, which is identical to the *C*-index used in logistic regression. Unlike simple accuracy based on predicted labels, AUC takes into account the predicted probabilities assigned to each class. The estimation of accuracy and AUC was performed using the out-of-bag (OOB) samples, which had been left out during the training.

All classifiers performed well above the baseline levels of 50% accuracy and 0.5 AUC. While the performance was modest, the results nevertheless indicate that the sentence embeddings contain some information associated with the choice of form. This, in turn, suggests systematic contextual differences between the full and short forms. The choice of masking strategy did not play a major role. The models trained on the sentences in which both forms were allowed to co-occur achieved higher accuracy, likely due to the larger amount of training data.

An additional advantage of this approach is interpretability: we can identify the most distinctive contexts by examining sentences assigned the highest predicted probabilities for each form. The sentence below received the highest predicted probability of the full form in the model trained on the online news data with the nonce-string masking condition and excluding the sentences where the full and short form co-occurred. The target form is in bold.

(9) *Specifically, the stocks of microchips and semiconductor companies tied to* ***artificial intelligence*** *have been leading the market higher.* [Probability of the full form 74.8%]

This sentence, like many similar high-probability examples, concerns the market performance of companies producing microchips, such as Nvidia, and other technologies that are both essential to and stimulated by the recent developments in AI as a technology and industrial sector.

As for the short form, two sentences with the highest predicted probabilities are shown below. Note that the context in (10a) is misclassified as typical of the short form, although the original sentence contains the full form. These and other high-probability contexts focus on the benefits and opportunities of AI for users and entrepreneurs.

(10) a. *The application of* ***artificial intelligence*** *has a range of benefits it can offer to people.* [Probability of the short form 72.1%]
b. *It aims to empower and enable entrepreneurs, even those without a technical background, to take advantage of* ***AI*** *as a partner in building their businesses.* [Probability of the short form 71.7%]

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

In the Reddit data, the full form received high predicted probabilities in contrastive contexts that explicitly include the word 'intelligence,' as in the example below. The full form is often used in metadiscussions that seek to clarify what (artificial) intelligence is, and how AI differs from human intelligence.

(11) *In reality there is no actual 'intelligence' in* ***artificial intelligence*** *and the nomenclature is a misnomer entirely.* [Probability of the full form 75.6%]

The short form is predicted in the contexts in which AI is construed as a useful tool and an efficient substitute for human labor (especially in software development) – something that many contributors discuss with concern. Note that example (12b) is misclassified, as the original version contained the full form instead of the predicted short form. Nevertheless, there are plenty of instances in which the short form is used when discussing the risks of AI displacing human jobs.

(12) a. *If there was a way I could use* ***AI*** *instead of developers to create our software, I would do it in a heartbeat – it would mean a huge financial windfall for me.* [Probability of the short form 75.5%]
b. *It's impossible to predict what will happen if* ***artificial intelligence*** *takes over most jobs, and frankly, I'm not very excited about the unknown* [Probability of the short form 73.4%]

While the models trained on the two corpora highlight different semantic aspects, as illustrated by the examples above, a careful manual inspection of the highly distinctive patterns reveals one overarching pattern. The short form is preferentially predicted in contexts that refer to AI as an artifact or an agent (antagonist) that affects an individual directly, as a user, entrepreneur or employee. By contrast, the full form is likely to occur in contexts related to the role of AI in the industry and stock market (especially in the online news), or in broader metadiscussions of (artificial) intelligence as a capability (especially on Reddit). Therefore, the meaning of *AI* is more relevant for the everyday experiences of an average reader than that of *Artificial Intelligence*. In terms of the semantic dimensions introduced in Figure 1, the senses expressed by *AI* also tend to be closer to the 'Figure' pole of the Figure–Ground continuum.

Although the contextual information allows us to distinguish between the full and short form consistently well above the chance level, the predicted probabilities are never close to 100%, with many contexts displaying only a marginal preference for one or the other form. A manual inspection of contexts with nearly equal probabilities reveals that such sentences are often short and semantically generic:

(13) a. ***AI*** *especially is gonna be a potential horror show.*
b. ***AI*** *is here to stay, IMO.*
c. ***Artificial Intelligence*** *is not some mass that expands on its own and needs to be expanded.*

This lack of specific semantic information makes these sentences difficult to classify. Another group of contexts is associated with specialized domains and applications of AI, including the arts, computer vision, energy, military use and other technical fields.

Overall, these findings suggest an association between contextual information and the choice between the full and short forms. Although the strength of this association

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

appears to be modest and varies across discourse contexts, the analysis nevertheless makes it possible to identify semantic tendencies associated with each form.

4.2. Identification of semantic neighbors with the help of word embeddings

The previous section employed SBERT embeddings for the classification task, relying on contextualized sentence-level representations suitable for measuring the distinctiveness of contexts in which the two forms occur. By contrast, the present section examines the differences between *AI* and *Artificial Intelligence* by identifying words that are semantically similar to each form because they occur in similar contexts. This approach requires word embeddings, which assign a single vector to each word type (Firth, 1957; Turney & Pantel, 2010) and are particularly well suited for exploring semantic fields and lexical relationships.

The results reported in this section are based on Word2Vec models (Mikolov et al., 2013) trained on the corpus data using the Python package *Gensim*. During training, the model (a shallow neural network) learns to predict words from their surrounding context. The main idea is that words that predict other words in similar ways tend to have similar meanings. In the resulting model, words are represented by vectors in a high-dimensional semantic space. Words that occur in similar contexts will be located close to one another in this space. This relationship can be quantified using different similarity measures, such as cosine similarity.

Before training the models, several preprocessing steps were taken. All tokens were lowercased. Importantly, the phrase *Artificial Intelligence* was coded as a single token: *artificial_intelligence*, to ensure that it receives one vector representation, comparable to that of *ai*. In addition, a number of frequently mentioned technologies and subtypes of AI were similarly treated as single token: *machine_learning*, *deep_learning*, *generative_ai*, *generative_artificial_intelligence*, *(artificial_)neural_network*, *artificial_general_intelligence*, in order to investigate potential sense specialization and relationships between closely related technical terms. The same was done with product names such as *Meta AI*, *Galaxy AI* and *Ai Pin*.

The models included all words occurring at least five times in the corpus and were trained using the skip-gram method, which has been shown to perform better on relatively small datasets. In this approach, the model learns to predict the vectors of surrounding context words based on the vector representation of a target word. Prior to training, both the sentences from the online news corpus and the Reddit posts were randomly reshuffled. The context window size was limited to five words and constrained by sentence boundaries. Although larger context windows may improve Word2Vec embedding accuracy (Di Gennaro et al., 2021), the structure of the news corpus (consisting of isolated sentences) motivated the use of more local contextual information. Every word was represented by a 100-dimensional vector (cf. Zheng et al., 2019, who found this dimensionality to yield the best results in their study of Chinese abbreviations).

An important issue is the substantial frequency imbalance between the two forms. Although frequency does not directly determine distributional similarity, it may influence it indirectly because more frequent words participate in more training examples, triggering more updates to the neural network weights than infrequent words. To investigate the role of frequency imbalance, I split all instances of *AI* in each corpus into two groups, creating two independent vectors. In the first condition, I applied a balanced 50/50 split, randomly replacing each occurrence of 'ai' with either 'ai_index_a' or 'ai_index_b'. In the second condition, the split probabilities reflected the

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

distribution of the short and full forms in the news corpus: 87.4% versus 12.6% (see Table 1), again with labels assigned randomly. Finally, I applied the extreme imbalance observed in the Reddit corpus, with probabilities of 99.3% and 0.7%.

Word2Vec models were then trained on the modified corpora, and cosine similarity was computed between the vectors 'ai_index_a' and 'ai_index_b'. This procedure was repeated five times for each corpus and probability bias, in order to account for variability introduced by the random assignment of labels and the stochastic nature of Word2Vec training, and the results were averaged. Cosine similarity values range from −1 to 1, where 1 means perfect similarity (as in a word compared to itself), 0 orthogonal relationships (i.e., no shared semantic dimensions, as in *cucumber* vs. *justice*), and −1 a perfectly opposite meaning, which is very rare in word embedding spaces. Since the semantic content is identical, the cosine similarities between the resulting vectors should in principle be close to 1. As shown in Table 3, this expectation is borne out for the evenly split data in both corpora. The cosine similarities also remain very high for the 87.4% versus 12.6% split.

In the condition with extreme imbalance (99.3% versus 0.7%), the similarities decrease, although they remain relatively strong: 0.81 in the online news corpus and 0.78 in the Reddit corpus. This decline is likely driven less by the relative frequency imbalance itself than by the low absolute frequency of the label 'ai_index_b' (averaging only 28 instances in the news corpus and 120 in the Reddit corpus). Other factors, such as corpus size and contextual diversity, may also contribute to the observed reduction in similarity. Still, the similarities remain sufficiently strong to support the application of distributional semantic methods.

Table 4 displays the top 20 nearest neighbors of the short and full forms in the online news corpus, based on cosine similarities. The similarity between the full and

Table 3. Cosine similarity between the vectors representing differently labelled AI tokens, averaged across five trained Word2Vec models

	Average cosine similarity		
Corpus	AI_{50} vs. AI_{50}	$AI_{87.4}$ vs. $AI_{12.6}$	$AI_{99.3}$ vs. $AI_{0.7}$
Online news	0.97	0.95	0.81
Reddit	0.97	0.95	0.78

Table 4. Top ten closest semantic neighbours of the short and full forms, based on Word2Vec and cosine distances. Data: online news corpus

Short form		Full form	
Neighbor	Cosine	Neighbor	Cosine
artificial_intelligence	0.90	ai	0.90
generative_ai	0.89	generative_ai	0.86
genai	0.86	machine_learning	0.82
machine_learning	0.83	technology	0.79
gpt	0.83	computing	0.78
chatbots	0.82	generative_artificial_intelligence	0.78
chatgpt	0.81	genai	0.77
generative_artificial_intelligence	0.81	chatbots	0.76
llm	0.80	llm	0.76
algorithms	0.80	innovations	0.75

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

short forms is very high (0.90). For comparison, the words *television* and *tv* have a similarity of 0.87, while *telephone* and *phone* reach only 0.74 according to the same model. The two neighbor lists overlap substantially, both including the variants of *generative AI*, *machine_learning*, *chatbots* and *llm*. At the same time, the full form shows closer associations with the words *technology, computing* and *innovations*, whereas the short form is more similar to *gpt*, *chatgpt* and *algorithms*.

To explore the differences between the vectors of the full and short forms, I exploited one well-known property of word embeddings. Word vectors can be added and subtracted, and such operations often correspond to meaningful semantic relations. A classic example is *king – man + woman ≈ queen*: if one takes the vector of *king*, subtracts from it the vector of *man*, and adds the vector of *woman*, one can get an approximation of the vector of *queen* (Mikolov et al., 2013). It should be noted, however, that not all semantic relations are captured by such vector operations equally well in all languages (Köper et al., 2015).

If we want to explore the difference between two words, we can subtract the vector of one word from the vector of the other word, and see which words are the most closely aligned with the resulting difference vector. Accordingly, I constructed a difference vector by subtracting the vector of *artificial_intelligence* from the vector of *ai*: *diff_vector = ai – artificial_intelligence.* To interpret this vector, I considered as candidates the 1,000 closest neighbors of *ai* and the 1,000 neighbors of *artificial_intelligence* with a minimum frequency of 20 as candidates, and computed their cosine similarity to the difference vector. The words with positive cosines can be interpreted as capturing what one might call 'AI without Artificial Intelligence,' whereas the words with negative cosines are interpreted as approximating the meaning 'Artificial Intelligence without AI'. The top 20 candidates on each side for the online news data are presented in Table 5.

Table 5. Top words closest and furthest to the difference vector between ai and artificial_intelligence in the online news corpus

'AI without artificial intelligence'		'Artificial intelligence without AI'	
Word	Cosine	Word	Cosine
interface	0.209	science	−0.333
edits	0.205	climate	−0.265
android	0.205	proliferation	−0.264
multitasking	0.2	geopolitics	−0.256
login	0.199	pioneered	−0.236
sandbox	0.198	globalization	−0.22
subscriber	0.194	advances	−0.21
ui [User Interface]	0.19	robotics	−0.206
editing	0.188	emerging	−0.206
tweak	0.188	hydrogen	−0.206
bing	0.187	global	−0.204
siri	0.187	technological	−0.204
downloading	0.187	semiconductors	−0.2
functionality	0.182	semiconductor	−0.198
captions	0.176	microscope	−0.195
prompts	0.174	lasers	−0.195
macos [MacOS]	0.174	industry	−0.194
toggle	0.173	transnational	−0.191
disable	0.17	misinformation	−0.19
intuitive	0.17	accelerating	−0.19

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

The absolute cosine values are modest, indicating that the difference vector is only weakly aligned with the words.[3] Nevertheless, some systematic patterns emerge. The words in the leftmost column of Table 5, which are the most similar to the differential vector (i.e., the closest to 'AI without Artificial Intelligence'), are related to the practical use of software by individuals, for example:

- types of software and platforms (*android*, *bing, siri, macos*);
- interaction with software (*interface, ui, multitasking, subscriber, login, downloading, prompts*, *edits, editing, tweak, toggle, disable, intuitive);*
- specific purposes for using software (*functionality, captions, sandbox*).

These patterns suggest that the short form AI is closely associated with AI construed as a software tool that users interact with directly and employ for specific tasks. A few illustrative examples of this usage are provided below.

(14) a. *This leaves the clinician with the easier task of simply editing the prompt that the **AI** created and adding any missing information.*
b. *Whether you are researching complex topics, creating social media posts or summarizing long reports, Merlin is on call — no more juggling multiple tabs or switching between all the different **AI** tools.*
c. *I have seen this demonstrated by setting **AI** applications as browser bookmarks or homepages for easy access.*

By contrast, the meaning 'Artificial Intelligence without AI,' which is captured by the words with negative cosine values in the right-hand part of Table 5, is most closely associated with the word *science*, followed by expressions related to cutting-edge technologies (*robotics, lasers, hydrogen, semiconductors, emerging, advanced, pioneered*), as well as global and societal challenges (*climate, globalization, geopolitics, misinformation*). Therefore, 'Artificial Intelligence without AI' is largely a scientific and technological domain embedded in broader processes of innovation and global change, rather than a directly manipulable tool. Examples of this use of the full form, framing *Artificial Intelligence* as a research field and a site of large-scale challenges, are provided below.

(15) a. *Armed with a team of world-class mathematics, statistics and coding experts, Woo set out to tackle one of the most challenging frontiers in **artificial intelligence**: autonomous software engineering.*
b. *Automation and **artificial intelligence** are transforming global sectors, with McKinsey Global Institute predicting that 375 million jobs may be obsolete by 2030.*

A comparable model was trained on the Reddit data in order to assess whether the patterns observed in the news corpus generalize across genres. Because the Reddit corpus is small and noisy, and the full form occurs only 117 times, the resulting

[3]Additional comparisons of other pairs, e.g., *king – queen* or *television – tv*, show that cosine values of the nearest neighbours are frequently below 0.5, which suggests that the semantic difference between two words often does not correspond closely to any existing words.

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

Table 6. Top words closest and furthest to the difference vector between ai and artificial_intelligence in the Reddit corpus

'AI without artificial intelligence'		'Artificial intelligence without AI'	
Word	Cosine	Word	Cosine
web	0.265	alignment	−0.297
llm	0.249	sentience	−0.283
competing	0.236	intelligence	−0.280
hack	0.232	consciousness	−0.275
somebody	0.223	emergent	−0.250
software	0.215	universe	−0.249
mostly	0.209	depth	−0.278
chatgpt	0.197	artificial	−0.243
writers	0.195	adaptability	−0.240
worthless	0.193	creativity	−0.236

patterns should be interpreted with caution. In this model, the cosine similarity between the full and short form was considerably lower (only 0.68). The nearest neighbor or *ai* was *generative_ai* (0.71), indicating that the full and short forms are less similar in the subreddit data than in the online news corpus. Despite the small corpus size and the low frequency of the full form, the difference vector yields surprisingly consistent patterns. Table 6 displays its 10 nearest and 10 most distant neighbors.

As indicated by the results in Table 6, the meaning 'AI without Artificial Intelligence' is most similar to the meaning of the World Wide Web. This interpretation is supported by an inspection of the contexts, which frequently draw comparisons between the developments in AI and the emergence of the World Wide Web, as in the context below.

(16) *I think the benefits of ai are more subtle than what the web was. The web allowed access to information but you still had to think for yourself. This was a problem for people without education or mathematical ability. Ai takes the web to the next step.*

The term *AI* appears to be primarily understood as referring to large language models, which often compete with human professionals (e.g., with content *writers*), whose labor is construed as easily replaceable by anyone (*somebody*), as illustrated in the following example:

(17) *But as AI gets better, the barrier to entry gets lower and lower. We are already at the point where somebody with zero coding experience can make simple programs. Soon we are going to get to the point where somebody with zero coding experience can make sophisticated programs.*

As for 'Artificial Intelligence without AI,' it is associated with the sense 'Capability': half of the top 10 words relate to cognition: *intelligence*, *sentience*, *consciousness*, *adaptability*, *creativity* and (emotional and intellectual) *depth*:

(18) *Understanding language and vision are the *first* steps of artificial intelligence.*

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

These properties are sometimes viewed as emergent properties of Artificial Intelligence, and at other times as qualities that AI lacks and cannot obtain (see example in [11]).

Note that *alignment*, which is the first on the list, usually means a requirement for designing artificial intelligence systems that should reliably pursue human goals and values, be helpful and safe:

(19) *Now, models have been tuned to prioritize 'alignment' — meaning they are trained to be more agreeable, polite and less critical. Good for avoiding lawsuits. Bad for honest feedback.*

To summarize, the analysis of the word embeddings suggests that the short form *AI* is more closely associated with the 'Artifact' sense, where it resembles other tools, platforms and models, such as the Web and large language models employed for editing, writing and text-related tasks. In this function, *AI* is represented as capable of acting as an Agent or even Antagonist, because it can threaten the demand for human labor. Accordingly, the distinctive uses of the short form are closer to the 'Figure' pole of the Figure–Ground continuum in Figure 1.

By contrast, the full form *Artificial Intelligence* appears to be more strongly associated with the senses of an academic field or technological and industrial domain in the news, and with a cognitive capability in the Reddit data, placing it closer to the 'Ground' pole of the semantic schema.

Taken together, these patterns tentatively suggest that the meanings associated with AI may be more cognitively accessible than those associated with Artificial Intelligence. Although the cosine similarities between the difference vectors and their nearest neighbors are modest, the patterns appear to be systematic and interpretable. Notably, the Reddit data appear to place greater emphasis on the more abstract 'Mind' side of the Mind–Matter axis, whereas the news corpus more strongly foregrounds distinctions related to the applied, material 'Matter' side.

5. Grammatical profiles

Across both corpora, plural forms were attested only for the short form *AI*. In the online news corpus, eight out of 4,054 instances of the short form occurred in the plural. In the Reddit corpus, 222 instances of *AIs/ais* and 28 cases of *AI's/ai's* were identified (after a manual disambiguation with the genitive form). The fact that only the short form pluralizes is compatible with the previous findings that it more favors Figure-like senses, because plurality is associated with individualization and boundedness, both characteristic of Figures (Janda, 1996). The plural forms usually refer to different large language models available on the market and express the senses 'Artifact' or 'Agent':

(20) *Chat gpt and all **ais** just agree with you, they are sycophants. Be very careful.* [Reddit]

To analyze the syntactic functions of the forms, I parsed all sentences in the online news corpus using a neural model trained to produce Universal Dependencies annotation (Qi et al., 2020; Zeman et al., 2023). As for the Reddit data, it was quite noisy, with frequent missing punctuation and typos, which is why I performed a

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

Table 7. XXX

Determiner	Short form	Full form
Definite (*the*, possessive, demonstrative pronouns)	100 (2.5%)	5 (0.9%)
Indefinite (*a(n)*, *some*)	25 (0.6%)	3 (0.5%)
Adjectival modifier	263 (6.5%)	21 (3.6%)
Total	*4,054 (100%)*	*584 (100%)*

manual annotation of all 117 instances of the full forms and a randomly selected, size-matched sample of contexts with the short form.

First, I zoomed in on the use of different determiners and adjectival modifiers with the full and short forms. This information was extracted automatically from the corpus data annotated with Universal Dependencies and then checked manually. The frequencies in the online news corpus are shown in Table 7. The counts in the manually annotated Reddit data were very low, which is why they are not reported here.

Although the determiners and adjectives are generally rare with both forms, the short form shows a higher frequency of occurrence with definite determiners, including the definite article, and with adjectives, as in the examples below. Using the short form *AI* to denote definite referents is efficient because such referents are accessible in discourse.

(21) a. In the case of generative AI, your willingness to wait for a response gives more time for **the AI** to do a more in-depth analysis.
b. *Who in their right mind thought pairing a snarky writer with* ***a megalomaniacal AI*** *was a good idea?*

Finally, I examined the syntactic functions of both forms by analyzing the Universal Dependency relations, which are as follows:

- 'appos': appositions, for example, *Artificial Intelligence (**AI**);*
- 'compound': part of a nominal or other compound, for example, ***AI** tools*, ***AI**-generated;*
- 'conj': a conjunct in a coordinate structure, for example, *virtual reality and **Artificial Intelligence***;
- 'nmod': nominal modifier of another noun, marked with a preposition or the genitive suffix, for example, ***AI**'s impact*, *the impact of **Artificial Intelligence***;
- 'nsubj': nominal (non-clausal) subject, for example, ***AI** is sucking up too much electricity*;
- 'obj': object, for example, *train **AI***;
- 'obl': oblique, or any prepositional phrase serving as an argument or adjunct of a verb, for example, *talk to the **AI***;
- 'other': all other infrequent categories, such as nominal predicates and heads of non-clausal units.

Figure 2 displays the percentages of the dependency relations for each form in the online news corpus. The results show that the short form is used predominantly as the nonhead element in compounds. In many cases, *AI* serves to narrow down the referent denoted by the head noun by indicating its membership in a certain class (*AI model, AI agent, AI system*), in line with Croft's (2022: 141) 'typifying' function of nominal

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

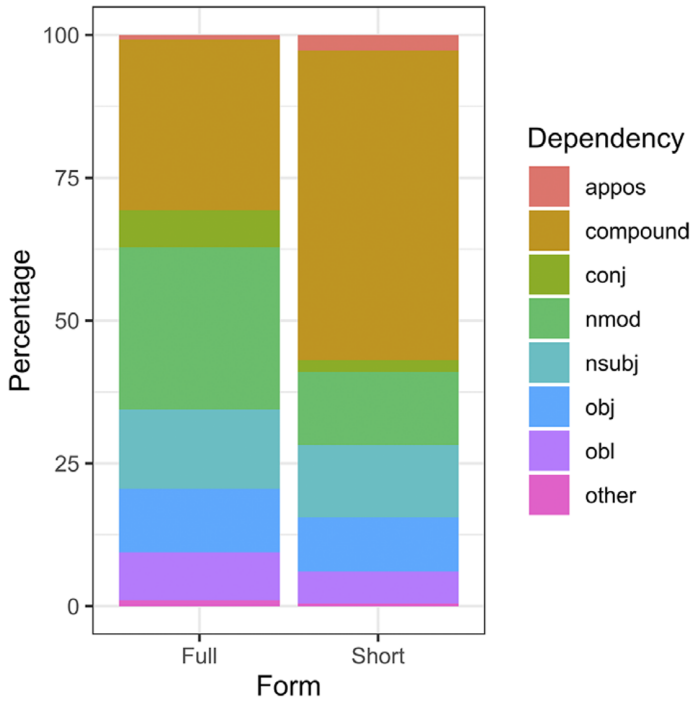


Figure 2. Proportions of Universal Dependencies in the online news corpus for the full form (N = 584) and the short form (N = 4046).

modifiers. It is well known that the meanings of noun–noun compounds are notoriously multifaceted and context-dependent (Downing, 1977). Accordingly, compounds with *AI* exhibit very diverse interpretations, including 'made with the help of AI' (as in *AI image*), 'being about AI' (*AI hype*) or 'having AI as the object of action' (*AI misuse*). A second major group consists of compounds with past participles (e.g., *AI-driven*, *AI-powered*, *AI-generated*), in which *AI* functions as an agent or an instrument.

The full form occurs more frequently as a nominal modifier, often with a preposition *of*, but a range of other prepositions is also attested, as in *advances in artificial intelligence.* It also more frequently serves as a coordinate conjunct and an oblique phrase. The coordinate conjunct function is often observed when *Artificial Intelligence* appears in lists of different technologies or research fields, as in the following example:

(22) *Currently staffed with 200 experts, Sony's research and development division, focused on agriculture, climate and* ***artificial intelligence****, is set to fortify its operations in Bengaluru.*

The functions of obliques are diverse, but some of them are semantically similar to the participial compounds with the short form (e.g., *AI-generated*), introducing the agent or instrument:

(23) *The singer was a victim of deepfakes generated by* ***artificial intelligence*** *that resulted in pornographic images of her all over social media.*

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

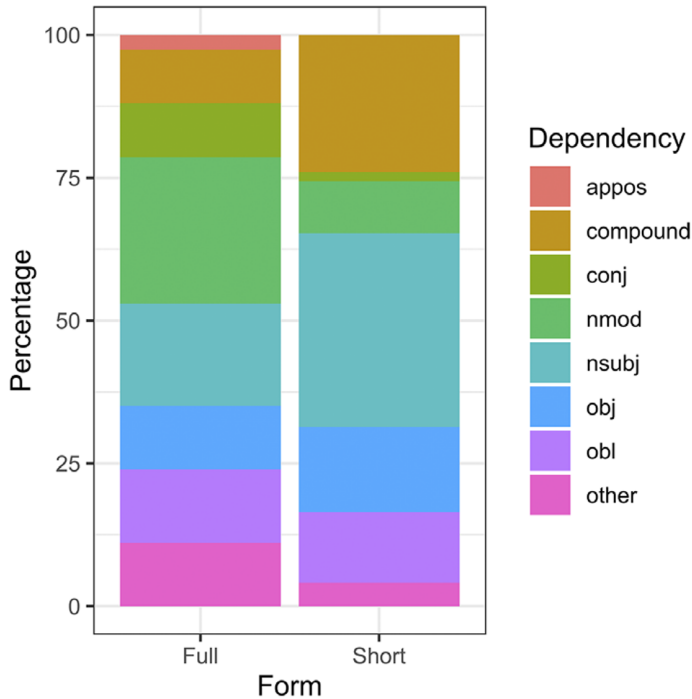


Figure 3. Proportions of Universal Dependencies in the Redditcorpus for the full form (N = 117) and the short form (N = 117, arandom sample).

Figure 3 displays the same dependencies in the Reddit data. Although the functions related to nominal phrases (compound and nominal modifier) are overall less prominent than in the online news corpus, most likely due to the less nominal style of Reddit comments, the short form is nevertheless more frequently used as a compound and less frequently as a nominal modifier than the full form. In addition, the full form is more frequently used as a coordinate conjunct, typically following a reference to another technology. While the two forms no longer differ in the frequency of occurrence as obliques, *AI* is now more frequently attested in the function of a subject. In these cases, it often has the meaning of an 'Agent,' as in the example below:

(24) *The AI was literally wrapping my entire app in try-catch blocks by the end.*

The preference of *AI* in the compound function and *Artificial Intelligence* in the nominal modifier role can be explained in terms of efficient word order. According to the principle 'Easy First,' short, syntactically simple, frequent and previously mentioned words and phrases tend to be produced before long, complex, infrequent and new forms (MacDonald, 2013). When *AI* is produced early in noun phrases as a compound element and *Artificial Intelligence* appears late as a prepositional modifier, this helps save time and processing costs in language production. Also, it can help avoid syntactic ambiguity. For example, *Artificial Intelligence bubble* can be parsed in two ways: [[Artificial Intelligence] bubble] or [Artificial [Intelligence bubble]]. By contrast, the compound *AI bubble* is unambiguous.

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

6. Summary and discussion

The corpus analysis was based on two corpora: sentences from online news texts, which cover a broad range of topics and offer few, if any, opportunities for interaction, and comments from subreddits dedicated specifically to interaction about AI. These differences in the accessibility of the referents due to common ground are mirrored in the relative frequencies of the full and short forms, in full accordance with the expectations based on the principle of negative correlation between accessibility and costs – one of the main principles of efficient communication. Although the short form is predominant in both corpora, the proportion of full forms is higher in the online news corpus (about 13%) than in the subreddit comments, in which the short form is used nearly exclusively (in more than 99% of all instances).

The analysis of sentence and word embeddings suggests that the short form is associated with more accessible meanings than the full form. This contrast can be summarized very briefly as follows: *AI* is relevant to individual users, whereas *Artificial Intelligence* is relevant to society, economy and humanity at large.

More specifically, *AI* is discussed more frequently as a practical tool useful for the immediate needs of individual users or businesses, especially in connection with writing and editing tasks. It also shows stronger similarity with terms that reflect users' practical experience with AI as a tool. In this sense, *AI* typically refers to large language models, along with the interfaces through which the user interacts with them (e.g., in chatbots). Therefore, *AI* is more often used in the senses closer to the 'Figure' pole in the semantic schema in Figure 1, namely, an artifact (a program, tool, model) and an agent, or even an antagonist, for instance, as a competitor on the labor market. For most users, practical experience with a technology (whether positive or negative) is part of their daily life, and thus more accessible than abstract discussions of scientific progress or global challenges.

By contrast, the full form *Artificial Intelligence*, compared with *AI*, is primarily seen as a branch of science and technology, a key driver of industry and a source of global challenges. In these senses, it is closer to the 'Ground' pole of the Figure–Ground continuum. In the Reddit data, the full form is also more frequently used in metadiscussions concerning the (apparent or contested) capacity of machines to 'think'.

The differences between the forms vary somewhat across the corpora. In the online news corpus, the contrast is primarily located on the 'Matter' side of the semantic schema in Figure 1. In the Reddit data, the distinction also involves the 'Mind' side, contrasting Artificial Intelligence as an academic field or abstract capability with AI as an agent. This pattern aligns well with the communicative goals and discourse properties of the two text types: the primary function of news is to report verifiable events anchored in time and space and affecting people, markets or institutions, while Reddit discussions are more oriented toward argumentation, explanation and negotiation of complex concepts.

The semantic differences between the forms are mirrored in their grammatical profiles. The short form, which more often denotes individuated referents, can occur in the plural, unlike the full form. It is also more frequently used as a definite noun, which is efficient, as definite referents are highly accessible (Ariel, 2001). In addition, the two forms differ in their syntactic roles. The full form acts more frequently as a prepositional nominal modifier (e.g., *use of Artificial Intelligence*), while the short form more often appears as a nonhead element of compounds (e.g., *AI tools*).

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

This asymmetry can also be explained by efficiency: the shorter and less complex form *AI* is better suited to appear early in the noun phrase than the heavier expression *Artificial Intelligence*. These patterns are consistent with the principles of efficient word order, more exactly, the principle 'Easy First' (MacDonald, 2013). They also reduce the potential for syntactic ambiguity.

Notably, in the Reddit corpus the short form is used more frequently as the subject, which aligns with its more prominent role as an Agent (Antagonist). In addition, AI commonly occurs in participial compounds (e.g., *AI-generated*), where it functions as an agent or instrument that enables or assists the end user.

In summary, the semantic and grammatical patterns observed in the data support the principles of communicative efficiency: the less costly shorter form is preferred in contexts where the relevant meanings are more accessible. It is also mentioned earlier, which facilitates language production. What about the principle of isomorphism? The distributional semantic analyses show that *AI* and *Artificial Intelligence* display a high degree of similarity, as reflected in the high cosine similarity between their vectors, especially in the online newspaper corpus. This result is consistent with the results in Zheng et al. (2019), based on Chinese, who report high similarity scores between word embeddings of full forms and abbreviations.

At the same time, the contexts in which the forms occur are distinctive enough for us to predict the form based on the sentence-level distributional semantics with accuracy at a rate well above chance. Moreover, the short form is used almost exclusively in the Reddit corpus. Even in the online news corpus, instances of the full form are relatively infrequent. This means that all three strategies of resolving the tension between formal reduction and the principle of isomorphism are at work: semantic differentiation, register specialization and the partial ousting of the full form.

We can therefore conclude that both principles – communicative efficiency and isomorphism – are supported by the data. Crucially, these principles are not in conflict. The semantic differentiation and register specialization of the forms are driven by the principle of negative correlation between accessibility and costs, whereby the shorter form is favored in contexts with higher accessibility. Note, however, that the results should be interpreted with caution because both the classification accuracy and the cosine similarities are modest and therefore provide only tentative support for the broader theoretical conclusions proposed here.

The use of reduced forms is often anchored in communicative settings characterized by rich common ground, which favor formal reduction due to higher referential accessibility and weaker constraints on creative and nonstandard language use. At the same time, such settings can also lead to an association between reduced forms and meanings that are closely tied to speakers' everyday experiences. This helps explain why colloquial style and personal relevance can go together in formally reduced forms, as observed for *chemo* in Hilpert et al. (2023), and for *TV* and *phone*, as discussed in Section 1.

These findings imply that the semantic and register-based specialization substantially limits the extent to which full and reduced variants can be used for maintaining uniform information density (*pace* Levy & Jaeger, 2007; Mahowald et al., 2013; see also Levshina & Lorenz, 2022; Jacobs & MacDonald, 2023). The predictability effects of local co-text on the choice between full and clipped forms found by Mahowald et al. (2013) (but not detected in Jacobs & MacDonald, 2023) may tentatively be attributed to register-specific differences in predictability of words based on their co-text (Bentum et al., 2019). Such differences, in turn, may reflect the conditions under

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

which language is produced. For example, in spoken interaction with shared context, speakers have little need for highly specific (and less predictable) expressions, given the availability of rich common ground. In addition, time pressure in spontaneous face-to-face interactions can result in a limited and repetitive linguistic repertoire, further increasing predictability (Bentum et al., 2019).

This discussion is relevant for the broader theoretical question of the relationship between isomorphism and optionality in language (cf. Leclercq & Morin, 2026). Language users abbreviate forms for reasons of brevity (Zipf, 1965[1935]). However, this also increases the number of available forms from which speakers can choose, potentially incurring additional cognitive costs in language production (Kemp et al., 2018; Zipf, 1949). In particular, the presence of close semantic neighbors can slow down language production in lexical selection tasks (e.g., Fieder et al., 2019; but see Ma et al., 2026 on grammatical variation). Still, the articulatory benefits of abbreviated forms, given that articulation constitutes a major bottleneck in language production (Levinson, 1995), appear to outweigh the costs associated with managing additional variants, if such costs arise at all. When alternating forms differentiate their functions through language users' inferences based on associations between forms and different usage contexts (e.g., varying levels of time pressure or formality), the system becomes more 'isomorphic'. This can provide additional communicative benefits for the addressee (cf. Zipf's, 1949 Force of Diversification) but may also reduce the potential for efficient articulation in speech production (Levshina & Lorenz, 2022). To summarize, both isomorphism and optionality are associated with their own specific costs and benefits.

An important question not addressed in this study is whether the semantic differences may confound the register effect. More exactly, can the extremely high proportion of the short form on Reddit be explained by a predominance of the most accessible senses in that corpus? An informal inspection suggests that this explanation is unlikely, as the semantic functions of *AI* on Reddit are very diverse. Further research is required, however, to disentangle the causal relationships between semantics, register and formal variation.

As AI research is constantly moving toward new frontiers, the semantic differences identified here are likely to become outdated soon. We can also expect to find additional patterns of differentiation in other registers and text types, and in speakers with varying levels of AI literacy and experience. Nevertheless, regardless of which technology comes to the fore next or which corpora are examined, we can expect to find comparable accessibility-driven differences between *Artificial Intelligence* and *AI*. A promising direction for future research would be to examine other clipping pairs, both within and across languages, in order to evaluate the broader generalizability of these theoretical ideas and methodology.

Data availability statement. The datasets and code for the statistical analyses can be found in an OSF project at https://osf.io/u7ek9/overview?view_only=47c0f0b1c00440b8908ba2a031db37cd.

Funding statement. Open access funding provided by Radboud University Nijmegen

Competing interests. The author declares none.

References

Ariel, M. (2001). Accessibility theory. In T. Sanders, J. Schliperoord, & W. Spooren (Eds.), *Accessibility theory* (pp. 29–87). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.8.04ari

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

Bentum, M., Ten Bosch, L., Van den Bosch, A., & Ernestus, M. (2019). Do speech registers differ in the predictability of words? *International Journal of Corpus Linguistics*, 24(1), 98–130. https://doi.org/10.1075/ijcl.17062.ben

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language aquisition* (pp. 1–33). Lawrence Erlbaum Associates.

Croft, W. (2022). *Morphosyntax: Constructions of the world's languages*. Cambridge University Press. https://doi.org/10.1017/9781316145289

Davies, M. (2008). *The corpus of contemporary American English (COCA)*. https://www. english-corpora.org/coca/

Di Gennaro, G., Buonanno, A., & Palmieri, F. A. N. (2021). Considerations about learning Word2Vec. *Journal of Supercomputing*, 77(11), 12320–12335. https://doi.org/10.1007/s11227-021-03743-2

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, 53(4), 810–842. https://doi.org/10.2307/412913

Fieder, N., Wartenburger, I., & Abdel Rahman, R. (2019). A close call: Interference from semantic neighbourhood density and similarity in language production. *Memory and Cognition*, 47(1), 145–168. https://doi.org/10.3758/s13421-018-0856-y

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Blackwell.

Glass, L. (2020). Verbs describing routines facilitate object omission in English. *Proceedings of the Linguistic Society of America*, 5(1), 44–58. https://doi.org/10.3765/plsa.v5i1.4663

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago University Press.

Goldhahn, D., Eckart, Th., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In N. Calzolari, Kh. Choukri, Th. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *PProceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 759–765). ELRA. https://aclanthology.org/L12-1154/

Guest, O., Suarez, M., Müller, B., van Meerkerk, E., Oude Groote Beverborg, A., de Haan, R., Reyes Elizondo, A., Blokpoel, M., Scharfenberg, N., Kleinherenbrink, A., Camerino, I., Woensdregt, M., Monett, D., Brown, J., Avraamidou, L., Alenda-Demoutiez, J., Hermans, F., & van Rooij, I. (2025). *Against the uncritical adoption of 'AI' technologies in academia*. Zenodo. https://doi.org/10.5281/zenodo.17065099.

Haiman, J. (1980). The iconicity of grammar: Isomorphism and motivation. *Language*, 56(3), 515–540. https://doi.org/10.2307/414448

Hilpert, M., Correia Saavedra, D., & Rains, J. (2023). Meaning differences between English clippings and their source words: A corpus-based study. *ICAME Journal*, 47(1), 19–37. https://doi.org/10.2478/icame-2023-0002

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. (2006). Survival ensembles. *Biostatistics*, 7(3), 355–373. https://doi.org/10.1093/biostatistics/kxj011

Jacobs, C. L., & MacDonald, M. C. (2023). A chimpanzee by any other name: The contributions of utterance context and information density on word choice. *Cognition*, 230, 105265. https://doi.org/10.1016/j.cognition.2022.105265

Janda, L. (1996). Figure, ground, and animacy in slavic declension. *Slavic and East European Journal*, 40(2), 325–355. https://doi.org/10.2307/309473

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review Linguistics*, 4(1), 109–128. https://doi.org/10.1146/annurev-linguistics-011817-045406

Köper, M., Scheible, C., & Walde, S. (2015). Multilingual reliability and "semantic" structure of continuous word spaces. In M. Purver, M. Sadrzadeh & M. Stone (Eds.) *Proceedings of the 11th international conference on computational semantics* (pp. 40–45). ACL.

Leclercq, B., & Morin, C. (2023). No equivalence: A new principle of no synonymy. *Constructions*, 15, 1–16.

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

Leclercq, B., & Morin, C. (2026). Attunement: a new principle of optionality [Manuscript submitted for publication].

Levinson, S. C. (1995). Three levels of meaning. In F. Palmer (Ed.), *Grammar and meaning: Essays in honour of Sir John Lyons* (pp. 90–115). Cambridge University Press. https://doi.org/10.1017/CBO9780511620638.006

Levshina, N. (2020). Conditional inference trees and random forests. In M. Paquot & S. T. Gries (Eds.), *Practical handbook of corpus linguistics* (pp. 611–643). Springer. https://doi.org/10.1007/978-3-030-46216-1_25

Levshina, N. (2022). *Communicative efficiency: Language structure and use*. Cambridge University Press. https://doi.org/10.1017/9781108887809

Levshina, N., & Lorenz, D. (2022). Communicative efficiency and the principle of no synonymy: Predictability effects and the variation of want to and wanna. *Language and Cognition*, 14(2), 249–274. https://doi.org/10.1017/langcog.2022.7

Levy, R., & Jaeger, F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS)* (pp. 849–856). The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0111

Ma, R., Van Hoey, T., & Szmrecsanyi, B. (2026). Isomorphism-inspired theorising about optionality and variation: No empirical support from English grammar. *English Language and Linguistics*, 30(2), 239–259. https://doi.org/10.1017/S1360674325000097

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4, 226. https://doi.org/10.3389/fpsyg.2013.00226

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313–318. https://doi.org/10.1016/j.cognition.2012.09.010

Marchand, H. (1960). *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. Otto Harrassowitz.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, 27(4), 12–14.

McCorduck, P., & Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (2nd ed.). A K Peters/CRC Press. https://doi.org/10.1201/9780429258985

Mikolov, T, Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. https://doi.org/10.48550/arXiv.1301.3781

Petricini, T. (2025). The power of language: Framing AI as an assistant, collaborator, or transformative force in cultural discourse. *AI & Society*. https://doi.org/10.1007/s00146-025-02586-2

Plag, I. (2003). *Word-formation in English*. Cambridge University Press. https://doi.org/10.1017/CBO9780511841323

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz & T.-H. Wen (Eds.) *AProceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). ACL. https://doi.org/10.18653/v1/2020.acl-demos.14

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). ACL. https://doi.org/10.18653/v1/D19-1410

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Schank, R. C. (1991). Where is the AI? *AI Magazine*, 12(4), 38–49.

Shapiro, S. C. (1992). Artificial intelligence. In S. C. Shapiro (Ed.), *Encyclopedia of artificial intelligence* (2nd ed., pp. 54–57). Wiley.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, Sh., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *"Artificial intelligence and life in 2030" One hundred year study on artificial intelligence: Report of the 2015–2016 study panel.* Stanford University. http://ai100.stanford.edu/2016-report

Talmy, L. (2000). Figure and ground in language. In L. Talmy, (Ed.) *Towards a cognitive semantics* (pp. 311–344). The MIT Press. https://doi.org/10.7551/mitpress/6847.003.0009

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188. https://doi.org/10.1613/jair.2934

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Zeman, D., Nivre, J., Abrams, M., Ackermann E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Akkurt, S. F., Aleksandravičiut˙ e, G., Alfina, I., Algom, A., Alnajjar, K., Alzetta, C., Andersen, E., Antonsen, L., Aoyama, T. … Ziane, R. (2023). *Universal dependencies 2.12.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). http://hdl.handle.net/11234/1-5150

Zheng, J., Sun, J., Xiao, X., & Yang, L. (2019). Semantics relationship between abbreviations and the original words based on word vectors. In *Proceedings of the 2019 3rd international conference on innovation in artificial intelligence* (pp. 60–64). Association for Computing Machinery. https://doi.org/10.1145/3319921.3319938

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Zipf, G. K. (1965). *The psychobiology of language: An introduction to dynamic philology*. MIT Press.

Cite this article: Levshina, N. (2026). Why AI is not (quite) the same as artificial intelligence: reconciling isomorphism and communicative efficiency in formal reduction, *Language and Cognition*, 18, e51, 1–26. https://doi.org/10.1017/langcog.2026.10100

https://doi.org/10.1017/langcog.2026.10100 Published online by Cambridge University Press