Machine Translation: Early Criticisms Revisited

Stefania Centrone*
Technical University of Munich
Munich, Bavaria, Germany
stefania.centrone@tum.de

Cosimo Perini Brogi*
IMT School for Advanced Studies Lucca
Lucca, Tuscany, Italy
cosimo.perinibrogi@imtlucca.it

Stefan Reifberger*
Technical University of Munich
Munich, Bavaria, Germany
stefan.reifberger@tum.de

Abstract

Hubert Dreyfus questioned the foundational ideas of early artificial intelligence research. Challenging the prevailing orthodoxy, he posited that the failure to manifest advancements in areas like language translation and problem-solving stems from a foundational misalignment with the intricacies of human "information processing". This paper restates Dreyfus' challenge. Based on case studies it argues that contemporary neural network systems have taken up the challenge by implicitly addressing three distinct philosophical problems, posed by Ludwig Wittgenstein, George Pólya, and Edmund Husserl.

CCS Concepts

• **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; **Natural language processing**; **Machine translation**; *Natural language generation*; *Language resources*; *Cognitive science.*

Keywords

Machine translation, Philosophical foundations, Limits of artificial intelligence, LLMs, Syntax/semantics interface, Meaning-as-use, Heuristics.

ACM Reference Format:
Stefania Centrone, Cosimo Perini Brogi, and Stefan Reifberger. 2024. Machine Translation: Early Criticisms Revisited. In *Human Centred Artificial Intelligence - Education and Practice (HCAIep '24), December 02–03, 2024, Naples, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701268.3701286

1 Introduction

In *What Computers Can't Do: The Limits of Artificial Intelligence*, Dreyfus critiques the early years of artificial intelligence research, arguing against the belief that machines can emulate human intelligence [12]. The pioneers of the 1950s aimed to construct intelligent machines to deepen our understanding of the mind/brain.[1] They relied not on exhaustive "solution-finding" but on "search" algorithms and symbolic reasoning, a method termed "cognitive simulation".

Dreyfus asserts that the shortcomings in language translation, problem-solving, and pattern recognition are due to these tasks necessitating a unique human form of "information processing" that resists artificial replication [12, p. 13]. Consider the sentences:

He saw her duck. She was afraid of the football hitting her. (*EX*)

When translating the sentences to other languages, ambiguity arises because "duck" can be a noun or a verb. Any reliable translation method must solve the problem of selecting the correct option. Humans, according to Dreyfus, process information holistically. This means they grasp not only of single words but sentences and groups of sentences to reduce ambiguities as the above. Call the task of replicating this contextual grasping *Dreyfus' Challenge.*

This paper examines notable solutions proposed for machine translation (MT) over the last seventy years and questions whether exhaustive algorithms can replicate this human processing. Focusing on automated translation, we highlight significant challenges in machine translation literature,[2] often dismissed by computer scientists as philosophical concerns.

We argue that the history of machine learning has progressively approached solutions to these philosophical concerns, though a definite and uniform solution is lacking. Wittgenstein's *meaning-as-use* focuses on practical and contextual language use relevant to developing MT systems that can handle ambiguity and context-dependence effectively. Husserl's *logical grammar* offers insights into syntactic coherence, emphasizing logical structures in language. Pólya's *heuristic processes* highlight the importance of holistic problem understanding and solution planning, informing the design of MT systems that better mimic human problem-solving.

We first present important approaches to machine learning and how they made progress in solving Dreyfus' Challenge. We will exemplify our work with the above example sentences as a case study. Then, we introduce the above philosophical problems and how the new machine-learning approaches tackled these.

2 Main approaches to machine translation

In his 1960 paper [2, p. 92-93], Bar-Hillel recalls that in June 1952, he was likely the sole individual working full-time on machine translation. By 1958, this number had grown to around 250, marking the emergence of a multi-million dollar industry. Bar-Hillel critiques the unrealistic ambition of achieving fully automatic high-quality translation, arguing that early advancements fostered an illusion of imminent success. In contrast, the disparity between initial outputs and quality translation remained significant.

*All authors contributed equally to this research.
[1]We borrow the term 'mind/brain' from contemporary bio-linguistics [8–10].

[2]Various formal approaches to resolving lexical ambiguity are examined in [29]. For a seminal overview of statistical machine translation techniques, see [20]. Additionally, [13] propose an efficient transformer-based method for automated translation across 100 languages without requiring English intermediation.

HCAIep '24, December 02–03, 2024, Naples, Italy

ACM ISBN 979-8-4007-1159-6/24/12
https://doi.org/10.1145/3701268.3701286

The past seventy years, methodologies have emerged to tackle human language translation. These can be broadly classified into rule-based machine translation (RBMT), statistical machine translation (SMT), example-based machine translation (EBMT), and large language model (LLM) based machine translation.

Rule-based Machine Translation. RBMT employs manually constructed linguistic rules and dictionaries to map features from the source to the target language. It includes two sub-approaches: *interlingua-based* and *transfer-based.*

The *interlingua* approach aims for language independence, allowing translation between any two languages once a robust interlingua representation is established.[3] However, creating an effective interlingua representation that captures meaning across languages is challenging.

The *transfer-based* approach relies on language-specific knowledge, allowing for tailored translation strategies. While it offers flexibility, it can be complex to develop and maintain due to the need for extensive linguistic rules and resources. RBMTs often reduce the holistic "grasping the context" to discrete rule-based operations. A classic example is SYSTRAN, extensively used since the 1960s.

SYSTRAN analyzes the grammatical structure of our example sentences (*EX*) based on predefined rules: it identifies "he" as the subject recognizes "saw" as the past tense of "to see," and encounters "her duck," which can be interpreted as either:

1 Possessive pronoun "her" with the noun "duck" (her duck—the bird she owns).
2 Object pronoun "her" with the verb "duck" (her action of ducking).

Without additional context, SYSTRAN might default to the more common or straightforward option based on its rules. Suppose it selects the noun interpretation and produces the French translation: "Il a vu son canard." (He saw her duck—the bird.) If the intended meaning was the action, the correct translation should be: "Il l'a vue se baisser." (He saw her duck—she lowers herself.)

The example identifies context insensitivity as a strong limitation of RBMT in using context beyond individual sentences. Henceforth, to disambiguate meaning dependent on braoder textual cues is instrinsically challenging for RMBT systems as SYSTRAN. Another limitation is their rigid dependence on predefined linguistic rules; effectiveness is tied to the quality and exhaustiveness of the rule set. When faced with unanticipated language uses, idiomatic expressions, or exceptions not anticipated during rule development, the system may produce incorrect or awkward translations.

Example-based Machine Translation. EBMT identifies patterns in previously translated sentences to generate translations, potentially producing smoother results than RBMT. However, its effectiveness is limited by the availability of relevant examples and idiomatic expressions in databases [19]. Consider again, our example sentence. When processing it, an EBMT system searches its example database for sentences containing similar phrases. Assume the database includes the following examples:

1 Source: "He saw her dog."
Target: "Il a vu son chien."
2 Source: "She was ducking because she was afraid of the football hitting her."
Target: "Elle se baissait parce qu'elle avait peur que le ballon la frappe."

Using these examples, the EBMT system can piece together the translation. It identifies that "He saw her ..." matches the structure of the first example. It also recognizes that "... duck. She was afraid of the football hitting her" corresponds to "baissait parce qu'elle avait peur que le ballon la frappe." and generates the translation: "Il l'a vue baissait parce qu'elle avait peur que le ballon la frappe." While the translation is not completely grammatically correct, it effectively conveys the intended meaning.

However, the system's ability to translate the idiom correctly depends on the presence of that idiom in the example database. If the second example were not in the database – say only the first – the translation would be incorrect. Also, if the idiom appears in a slightly different form, the system may struggle to adapt the example to the new tense or form. Thirdly, the contextual usage may invalidate the translation: without understanding the context, the system might incorrectly apply an idiomatic translation to a literal situation.

Thus, while EBMT offers advantages in handling certain linguistic challenges by leveraging existing translations, its effectiveness is inherently tied to the breadth and depth of its example database. The system's limitations in dealing with idiomatic expressions, especially those that are new, rare, or context-dependent, highlight the ongoing challenges in machine translation.

Statistical Machine Translation. Statistical Machine translation (SMT) uses probabilistic models from large parallel corpora to determine the most probable translations. It has improved upon RBMT and EBMT, by leveraging statistical methods to handle ambiguities and variations in language. Deep learning models, including encoder-decoder architectures, have improved translation quality by better managing long-range dependencies and context. Hybrid systems combining different approaches have also been explored.

Consider an experiment where an SMT system is trained on a large parallel corpus of English and French novels. This rich dataset enables the system to learn statistical relationships between the two languages. An SMT system would analyze our example sentences by breaking them down into words and phrases, computing the probabilities of possible translations based on their frequencies in the corpus. For "duck" in the presence of the word "football" the system considers "canard" (probability: 0.1) and "baisser" (probability: 0.9), selecting "baisser" due to its higher frequency when the word "football" is close to it.

LLM-based Machine Translation. Recent LLM-based machine translation generalizes across various tasks and benefits from self-supervised learning. Notable models like BERT and OpenAI's GPT series have significantly impacted the field, demonstrating impressive performance in tasks like zero-shot machine translation [5]. However, traditional encoder-decoder architectures often outperform ChatGPT for short- and medium-length texts, while the reverse is true for longer texts. Translations from non-English to English generally yield better results across systems, highlighting the complexities of performance evaluation [28].

[3]This method aligns with Leibniz's vision of a universal language. Cp. at least [21, AA VI, 4, p. 263–270].

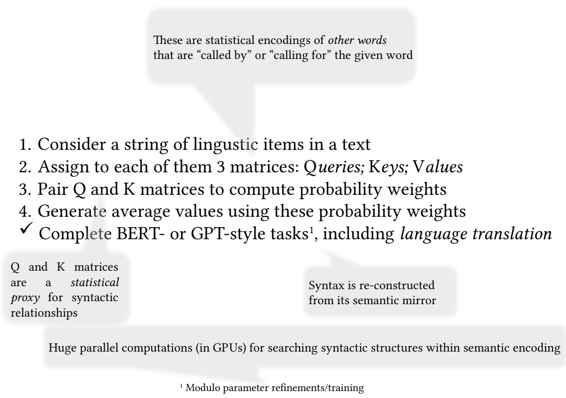


Figure 1: High-level picture of LLMs inverse engineering process for syntax reconstruction

LLMs process text prompts and generate output word by word or token by token, demonstrating a sophisticated ability to manipulate language. They are built on a specific architecture of neural networks using *transformers* and *attention modules*, as introduced by [31].[4] The pre-training phase is resource-intensive, utilising large text corpora and vector encoding to capture semantic relationships. LLMs rely on matrix operations and GPUs for parallel computing, enhancing their processing power.[5] Figure 1 summarises the LLMs task completion at a high-level.

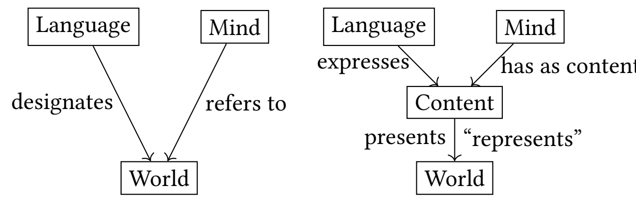
3 Three connections to philosophical problems

Wittgenstein on the problem of ambiguity. Ambiguities related to polysemy in natural language, where a single word or phrase can have multiple meanings, are well recognised, which was the case with the word "duck". Ludwig Wittgenstein's reference to the "Augustinian picture of language" in §1 in his *Philosophical Investigations* questions the traditional theory of meaning, variants of which can be found from Aristotle *De Interpretatione* (16a 3-8) onwards, for example in Boethius [3], Augustine [24], Bolzano [4], Twardowski [30], Husserl [17]). Essentially, traditional theories of meaning can be traced back to one of the following two schemas 2, where the difference lies in whether reference to the world is immediate (as in the triangular model of reference), with our linguistic and mental acts, both propositional and non-propositional, directly referring to the world, or mediated (as in the tetrahedral model) by abstract logical entities, which are the *contents* of our mental act-or-states or linguistic utterances.

Wittgenstein presents a contrasting view where *words are seen as tools* rather than mere designators. He believes that even simpler language games, such as a builder communicating with an assistant

[4]See [14] and the Constellation project for an overview of LLM architectures as of July 2023.

[5]Generally, attention mechanisms are applied in parallel across multiple "heads", allowing the model to focus on different parts of the input text. Each layer also includes mechanisms to stabilise training and enhance information flow.



(a) Triangular model. **(b) Tetrahedral model.**

Figure 2: Models of the relation of language, mind, and world.

using only four words to request specific stones, already go beyond Augustine's account, and, more generally, the traditional theory of meaning as exemplified by the two schemas in Figure 2.

He rightly stresses that what we refer to as *reference* – or "signification", "designation", "aboutness" – is of secondary importance compared to the way we use words. To grasp the nature of linguistic meaning, the focus should be on how words are employed in practice. Their so-called "reference" should only be considered afterwards. Now, while for RBMT a relevant issue is their context insensitivity (they cannot utilize context beyond individual sentences, making it difficult to disambiguate meanings that depend on broader textual cues), and while EBMT performs better because the example database includes correct usage—failing only when a specific use case is missing—statistical approaches seem to come closest to the Wittgensteinian idea of meaning as use. They "do not perform a deep analysis of the sentence to be translated, but identify groups of words that work together" [26, p. 17]. Their effectiveness lies in identifying consistent language equivalences and recurring patterns through statistical analysis. Since meaning is derived from word usage, statistical methods effectively reveal regularities in how words are used. At the same time, the in-context learning demonstrated by LLMs approximates Wittgenstein's meaning-as-use by encoding linguistic corpora accessed during LLM training. Recent studies [6] show rapid advancements in LLM-based translation, indicating continuous improvements in this direction. This enhancement of performance suggests that overcoming the conception of meaning as reference is an important step in addressing the problem of ambiguity posed by Dreyfus' Challenge.

Husserl on the problem of syntactic structure. In Paragraph 14 of his *Fourth Logical Investigation*, Edmund Husserl discusses "Laws which discourage Nonsense and Laws which discourage Absurdity", termed logical grammar or "logical morphology". This concept anticipates the concept of formal languages of [7] and its key features. Establishing formal languages requires the specification of an initial set of expressions ("meaning categories" in Husserl's terminology) and generation procedures to create new expressions ("a priori laws... that govern the combination of meanings into new meanings" [17, IV, 10, p. 317]). This cumulative constitution allows for the generation of all constructs from "fundamental forms". The laws of logical grammar prevent grammatically ill-formed expressions (*Unsinn*) and form the "ideal scaffold" (*ideales Gerüst*)[6] of language, representing the ideal structure of natural languages. Husserl sees

[6][17, IV, 14, p. 338]; [18, p. 526].

this as an effort to realise the rationalists' programme of a "universal grammar" from the seventeenth and eighteenth centuries [17, IV, 14, p. 336].

Teaching grammatical correctness to a machine has been a significant challenge, representing one of the major difficulties that computer scientists and linguists faced in achieving effective machine translation before the advent of large language models. Look back again at the result of the EBMT. While the result "Il l'a vue baissait parce qu'elle avait peur que le ballon la frappe" could be interpreted to have the intended meaning, it was grammatically ill-formed. The example database did not encode atomic constituents of English and French, nor does the described lookup method encode grammatical laws such as to produce only well-formed terms.

Research by [22] shows that syntactic information can be identified in attention vectors, suggesting a correlation between attention mechanisms and contemporary generative linguistics [11]. These findings imply that LLMs perform extensive parallel computations to navigate a vast linguistic search space, reconstructing the input's syntactic structure *probabilistically* within a semantic space defined by text corpora. The challenge for LLMs is to reverse-engineer a precise *syntactic structure* (output) from a *semantic space* that reflects the *use* of language, akin to the structures discussed by [15, 16]. Such encoding of syntactic information partially solves Husserl's problem of syntactic structure shedding light on another aspect of Dreyfus' challenge.

Pólya on the problem of holism and problem-solving. In his well-known text *How to Solve It?*, George Pólya identifies different phases in the heuristic process, namely, (i) *understanding the problem as a whole* and (ii) *establishing a programme that leads to the solution* [27, p. 5ff]. Dreyfus argued that early computer scientists focused too much on the second aspect. They sought to eliminate the initial phase by enhancing search algorithms, positing that sufficiently powerful algorithms could render the typically human holistic approach to information processing unnecessary. However, it is precisely this initial phase that has consistently led artificial intelligence to dead ends. Thus, one might contend the importance of incorporating a general understanding of the problem into the program. This holistic understanding seems to be crucial for capturing nuances, idiomatic expressions, and contextual dependencies, often lost in purely algorithmic approaches. Nonetheless, this presents a formidable challenge, as it appears that the human mind/brain does *not reckon* when comprehending a problem as a whole.

By employing LLMs, machine translation effectively reduces to a standard LLM completion task, which, from an abstract perspective, we assert consists of the reverse engineering/search problem of Figure 1. Thus, when conducting LLM-based automated translation, the network searches for a mathematically encoded solution to a (computationally hard) problem.[7] From the perspective of results achieved, contemporary LLM-based systems for automated translation are performing relatively well and advancing rapidly.

How can refined versions of *brute force* yield results nearly as effective as those derived from the specifically human form of information processing that Pólya refers to as *understanding* the problem as a whole? The possible answer is twofold: on the one hand, LLMs heavily parallelise the computations involved in the search via GPUs to shorten task completion time; on the other hand, one may argue that training on vast text corpora, i.e., the encoding of syntax into a semantic space, provides a proxy (or rough approximation) for the point (i) identified by Pólya and long neglected by research in machine translation. The latter reading suggests that we have found another step forward in overcoming Dreyfus's Challenge.

[7]Note in passing that LLMs demonstrate relatively good performance in solving problems of a verbal, physical, or practical nature [25]. Nevertheless, they perform rather poorly in tasks requiring genuine reasoning [23]. On the contrary, the software AlphaProof, combining generic reinforcement learning algorithms with symbolic reasoning, has showed very good capabilities in solving problems in advanced mathematics [1]

4 Conclusion and future perspectives

The strong performance of LLMs in translation questions Dreyfus's critiques of artificial intelligence, at least from a *purely behaviourist* perspective. This progress is attributed to three factors: (1) reconfiguring translation as a search problem in a defined semantic space; (2) access to extensive sets of linguistic tokens; (3) the introduction of attention mechanisms for assessing syntactic relationships.

Point (1) supports a Wittgensteinian view of semantics based on language use, keeping the syntax-semantics interface discussion open. Point (2) connects to Husserl's observations regarding the difficulties of mechanically acquiring syntax, emphasising the importance of substantial linguistic data, despite its biological implausibility. Point (3) aligns with Pólya's notion that effective problem-solving requires planning and contextual modelling, despite current methodologies *do not provide* a plausible model for problem *understanding* [23].

While the increasing performance of neural networks raises questions about their status as models of (biological) intelligence, we aim to further explore these philosophical critiques by examining parallels between neural architectures and models of human intelligence in pattern recognition and problem-solving. Given the importance of these three philosophical issues in the progress of machine translation in the past, we suggest that they remain important conceptual benchmarks for the development of new architectures and models based on the insights underlying them. Building on this suggestion and going beyond our case studies, further inquiry could operationalize the concepts of syntactic structure (similar to [22]), holistic language interpretation, and meaning-as-use. With these advances, more extensive tests with a solid empirical basis for our discussion could be pursued.

References

[1] AlphaProof and AlphaGeometry teams. 2024. AI Achieves Silver-Medal Standard Solving International Mathematical Olympiad Problems. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/. [Accessed 12-10-2024].
[2] Yehoshua Bar-Hillel. 1960. The Present Status of Automatic Translation of Languages. *Adv. Comput.* 1 (1960), 91–163.
[3] Boethius. 2014. *On Aristotle On Interpretation.* Bloomsbury Publishing. Translated by A. Smith.
[4] Bernard Bolzano. 1837. *Wissenschaftslehre. Versuch einer ausfürlichen und größtentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter.* Vol. 1–4. Seidel. In: Berg J et al (ed), Bernard Bolzano-Gesamtausgabe [BGA], Frommann (Holzboog), Stuttgart-Bad Cannstatt 1969 ff, series 1, vols 11–14. Translation of selected parts: George R (ed, trans.) (1972) Theory of science. Oxford and Berkelay/Los Angeles.
[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter,

Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[6] Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. *CoRR* abs/2403.16512 (2024). https://doi.org/10.48550/ARXIV.2403.16512 arXiv:2403.16512

[7] Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

[8] Noam Chomsky. 1995. Language and nature. *Mind* 104, 413 (1995), 1–61.

[9] Noam Chomsky. 2000. *New horizons in the study of language and mind*. Cambridge University Press.

[10] Noam Chomsky, Ángel J Gallego, and Dennis Ott. 2019. Generative grammar and the faculty of language: Insights, questions, and challenges. *Catalan Journal of Linguistics* (2019), 229–261.

[11] Noam Chomsky, T Daniel Seely, Robert C Berwick, Sandiway Fong, MAC Huybregts, Hisatsugu Kitahara, Andrew McInnerney, and Yushi Sugimoto. 2023. Merge and the strong minimalist thesis. *Elements in Generative Syntax* (2023).

[12] Hubertus Dreyfus. 1972. *What Computers Can't do. A Critique of Artificial Reason*. Harper & Row.

[13] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research* 22, 107 (2021), 1–48.

[14] Sarah Gao and Andrew Kean Gao. 2023. On the Origin of LLMs: An Evolutionary Tree and Graph for 15,821 Large Language Models. arXiv:2307.09793 [cs.DL]

[15] Peter Gärdenfors. 2004. *Conceptual spaces: The geometry of thought*. MIT Press.

[16] Peter Gärdenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press.

[17] Edmund Husserl. 1901, ^{2}1913, ^{3}1921. *Logische Untersuchungen, Vol. II: Untersuchungen zur Phänomenologie und Theorie der Erkenntnis*. Max Niemeyer.

[18] Edmund Husserl. 1970. *Logical investigations*. Routledge and K. Paul; Humanities Press. English translation of [17] by J.N. Findlay.

[19] John Hutchins. 2005. Example-based machine translation: a review and commentary. *Machine Translation* 19, 3 (2005), 197–211.

[20] Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

[21] Gottfried Wilhelm Leibniz. 1966. Logical papers, translated and edited by G.H.R. Parkinson.

[22] Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30046–30054.

[23] Gary Marcus. 2024. This one important fact about current AI explains almost everything. (2024). August 2024 online version.

[24] Augustine of Hippo. 2008. *The Confessions*. Oxford University Press. Translated by by H. Chadwick.

[25] Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in artificial intelligence* 6 (2023), 1199350.

[26] Thierry Poibeau. 2017. *Machine Translation*. The MIT Press.

[27] George Pólya. 1945. *How to Solve It*. Princeton University Press. 1957 Edition.

[28] Andrew Rothwell, Joss Moorkens, María Fernández-Parra, Joanna Drugan, and Frank Austermuehl. 2023. *Translation tools and technologies*. Routledge.

[29] Steven L Small, Garrison W Cottrell, and Michael K Tanenhaus. 2013. *Lexical ambiguity resolution: Perspective from psycholinguistics, neuropsychology and artificial intelligence*. Elsevier.

[30] Kazimier Twardowski. 1894. *Zur Lehre vom Inhalts und Gegenstand der Vorstellungen. Eine psychologische Untersuchung*. Hölder, Vienna.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.