REVIEW

Paraphrase and translation: the importance of being close

[version 1; peer review: 2 approved]

Diana Santos [1,2], Anabela Barreiro [2,3]

[1]Department of Literature, Area Studies and European Languages, University of Oslo, Oslo, Norway
[2]Linguateca, Oslo, Norway
[3]Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Lisbon, Portugal

v1 **First published:** 18 Feb 2025, **5**:52
https://doi.org/10.12688/openreseurope.19108.1
Latest published: 18 Feb 2025, **5**:52
https://doi.org/10.12688/openreseurope.19108.1

Abstract

This article explores the concept of paraphrasing within computational linguistics, seeking to enrich its understanding by drawing parallels with translation studies and especially machine translation. It highlights the existence of two distinct yet related tasks: paraphrase generation and paraphrase detection, as well as points out the many (sometimes implicit) contact points in evaluation in both translation and paraphrasing. We claim that the concept of near-synonymy or near- equivalence is a shared concern of both disciplines, and its formalization should be pursued.

Plain language summary

This article examines paraphrasing—a way of expressing the same meaning using different words or phrasing—and its relationship with computational linguistics and translation studies. It focuses on two key tasks in paraphrasing: generation (creating paraphrases) and detection (identifying them). Both tasks are essential for applications like machine translation, summarization, question answering, and plagiarism detection. The study emphasizes how paraphrasing and translation share the concept of near-synonymy or near-equivalence, where slight differences in word choice preserve the same core meaning. While translation involves converting text between languages while keeping the meaning intact, paraphrasing focuses on rephrasing within the same language for different purposes—like adapting for a specific audience or simplifying complex ideas. Despite their differences, both processes require balancing the original intent with contextual adjustments. The article also highlights how concepts from translation studies can inform paraphrasing research. For example, the Portuguese phrase "Por Outras Palavras" ("In Other Words") has been used as the title for both translation and paraphrasing workshops, underscoring their conceptual overlap. By exploring these connections, the study aims to deepen our understanding of paraphrasing and improve methods for creating and detecting paraphrases in natural language processing.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 18 Feb 2025	✓ view	✓ view

1. **Stella Markantonatou**, Athena Research Center, Athena, Greece
2. **Eman Abdel Reheem Amin**, Majmaah University, Al Majma'ah, Saudi Arabia

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Paraphrase, Translation, Computational Linguistics, Paraphrase Generation, Paraphrase Detection, Closeness, Formalization, Equivalence

This article is included in the COST Actions gateway.

This article is included in the Advances in Natural Language Generation collection.

Corresponding author: Diana Santos (d.s.m.santos@ilos.uio.no)

Author roles: Santos D: Conceptualization, Methodology, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Barreiro A**: Funding Acquisition, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Framework Programme for Research & Innovation as part of the COST Action Multi3Generation (CA18231), as supported by the COST Association (European Cooperation in Science and Technology). The initial work was supported by FCT -- Fundação para a Ciência e a Tecnologia with reference UIDB/50021/2020, and post-doctoral grant SFRH/BPD/91446/2012.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Santos D and Barreiro A. **Paraphrase and translation: the importance of being close [version 1; peer review: 2 approved]** Open Research Europe 2025, **5**:52 https://doi.org/10.12688/openreseurope.19108.1

First published: 18 Feb 2025, **5**:52 https://doi.org/10.12688/openreseurope.19108.1

1 Motivation

The identification of paraphrases has emerged as a significant area of study in computational linguistics, being applied across various domains such as machine translation, language generation, summarization, text reuse identification, text simplification, question answering and plagiarism detection. Recognizing different ways of expressing the same or a similar meaning is crucial for enhancing the effectiveness of all these applications. Paraphrase identification is also often considered a component of text entailment detection, the task of determining whether one text logically follows from another. This broad applicability highlights the importance of developing robust methods for detecting and generating paraphrases in natural language processing.

In this article, we examine multiple definitions and operationalizations of paraphrases found in the literature, showing two distinct definitions of paraphrases that have emerged in the field – those based on systematic relationships between linguistic expressions out of context and those based on empirically determined semantic equivalence or closeness at the sentence level.

We deem it essential to emphasize the deep connection between paraphrasing and translation, as both hinge on concepts like near-synonymy or near-equivalence, allowing slight differences in word choice or phrasing that remain true to the original intent. Translation aims to maintain semantic equivalence while converting text from one language (the source language) to another (the target language). This process goes well beyond merely swapping words, as it often demands an awareness of different cultural contexts, histories, and linguistic norms to accurately convey meaning, ensure clarity and preserve the original intent of the author. Paraphrasing, in contrast to translation, remains within the same language but may involve rephrasing text to suit different audiences, styles, contexts, or varieties. While it does not require the cross-linguistic shift that translation does, paraphrasing still demands a careful balance of preserving the original meaning while adjusting the expression.

The connection between translation and paraphrasing can be seen in the title of one of translation studies key texts, "In Other Words"[1], where paraphrasing is conceptually linked to translation. And the other way around, the Portuguese phrase "Por Outras Palavras" (literally "In Other Words") has served as the title for two paraphrasing workshops[2,3].

1.1 Translation versus paraphrasing

Paraphrasing and translation both imply "closeness of meaning", whether in the same or different languages. When considering the significance of translation versus paraphrasing, one might argue that paraphrasing is less essential, given that translations are often needed to bridge language barriers for audiences who do not understand the original language. Paraphrasing may appear simpler since it operates within the same language. Walter Benjamin[4] observed that translation involves transferring a text from one cultural and linguistic framework to another, adding layers of complexity that paraphrasing typically does not face. However, when adapting a paraphrased text for different audiences or cultures, it may require adjustments beyond simple rewording. Much like translation, paraphrasing can demand an awareness of cultural contexts.

In any case, what is undeniable and intriguing is that, as discussed in Section 3 ("Interconnections"), there have been many instances where computational linguistics researchers have drawn from work in either paraphrasing or translation without explicitly emphasizing the theoretical closeness between the two. These overlaps have fostered a natural, albeit often unspoken, interconnection between the two disciplines. Finally, the increasing role of large language models (LLMs) in refining and paraphrasing texts has sparked important questions about authorship, originality, and accountability, and this also calls for a deeper understanding of paraphrasing and the development of robust paraphrase detection methods. The goal would be to ensure that the nuances between paraphrased and original content can be accurately identified, addressing concerns related to ownership and academic integrity.

2 Paraphrase in computational linguistics

In computational linguistics, tasks related to paraphrasing generally fall into two categories: paraphrase generation and paraphrase detection. These two tasks often imply distinct underlying definitions of paraphrase. Paraphrase generation has typically emphasized competence, i.e., stressing the linguistic system's ability to produce paraphrases. In contrast, paraphrase detection is more concerned with performance, looking at how paraphrases are identified or recognized in actual language use. While paraphrase generation begins with a singular meaning and explores various ways to express it, paraphrase detection focuses on determining whether two different sentences convey the same meaning or a close enough/approximate meaning.

Starting by paraphrase-related research in the area of language generation, Smadja and McKeown[5] explored a system called Cook, which uses extensive collocation knowledge extracted from a domain-specific corpus to generate diverse sentences from what they refer to as "semantic messages". The system's ability to merge collocations while managing multiple constraints demonstrates a significant potential for generating paraphrases (pp. 238). A notable aspect of their system is its adherence to what they describe as the "delocality of semantic constraints", a concept attributed to Kukich *et al.*[6], referring to the phenomenon where the same semantic information is distributed across different parts of the sentence, sometimes separated by considerable distance. Similarly to Smadja and McKeown, Kukich and colleagues had developed a system designed to help engineers generate documentation for telephone network planning operations. Their system demonstrated the flexibility of paraphrase generation by producing an average of 150 different sentence variations for a single message type.

A key development in paraphrase research came from the article by Barzilay and McKeown[7], "Extracting Paraphrases from a Parallel Corpus". They introduced a performance-based view of paraphrasing, which has since become the dominant approach.

This shift in perspective emphasized the practical application of paraphrase detection by using large datasets, such as parallel corpora – already used in translation – to extract paraphrased pairs and better understand their variations in real-world contexts. They emphasized that existing resources predominantly focused on lexical paraphrases, i.e., alternative expressions at the word level, noting the absence of phrasal or syntactically-based paraphrases. In their view, this limited the scope of how paraphrases are understood and used in computational tasks. According to these authors, previous approaches used manual collection of paraphrases for generation, something which is time consuming and application-specific; or a) used lexical resources such as WordNet (synonymy and/or other relations) to compute paraphrases; b) employed morphological and syntactical processors applied to term lists, what they call "term paraphrases"[1]; or c) used (linguistic) explanations of the meaning of regular polysemous adjectives[8]. In all these three latter cases, they claim that the process may result in unnatural/unattested examples, and this is why they propose to identify actually used paraphrases. However, Barzilay and McKeown[7] faced a significant challenge in evaluating the paraphrases generated by their method, because they had to rely on human judgment. They defined a paraphrase as demonstrating "approximate conceptual equivalence" but encountered difficulty in determining whether the evaluation should involve context or not. In some cases, only by considering the context could one judge whether two phrases were indeed paraphrases. But, including the context meant that the paraphrases would become dependent on specific situations, reducing their generalizability. Context-dependent paraphrases could not be used across different settings, making them less valuable as general linguistic resources. On the other hand, excluding context in evaluations risked missing nuances, as paraphrases might only work as substitutes in certain contexts. The challenge then was whether to include larger text snippets to capture context more effectively, an approach which would result in longer paraphrase pairs – not explored in their paper. One can clearly see that a trade-off exists between generalizability and accuracy in determining paraphrastic relationships. We believe their paper was a cornerstone for the paraphrase detection task, which turned into a larger endeavour in the years to follow.

Dolan and Brockett[9] developed the Microsoft Research Paraphrase Corpus (MRPC) using a significantly large data set, employing two key techniques on clustered news articles. The first involved finding pairs of sentences with a relatively small Levenshtein distance[10], a measure of the difference between two sequences by counting the minimal number of single-character edits required to transform one sentence into another. The second technique focused on identifying the initial sentences of articles with considerable lexical overlap and similar sentence lengths, as these similarities increased the likelihood that the sentences were paraphrases of one another. The MRPC corpus[2], containing 5,800 sentence pairs, was released in 2005 with human judgments indicating whether each pair was a paraphrase or not. Since its creation, the MRPC has become an invaluable resource for the development and evaluation of paraphrase detection models. As detailed by Finch *et al.*[12], these human judgments assessed whether the sentences were close enough in meaning to be considered paraphrases. Two annotators independently labeled the data, and a third annotator resolved disagreements. The inter-annotator agreement was 83%, underscoring both the complexity of the task and the ambiguity in labeling, with about 67% of sentence pairs classified as paraphrases. Once again, we have genuine sentence pairs conveying the same core information, but human verification is required to ensure this alignment. Given the wealth of this resource, Finch *et al.*[12] explored the effectiveness of evaluation metrics typically used in machine translation to assess "semantic equivalence classification" at the sentence level, considering paraphrases as semantic equivalence data. They also used data from the PASCAL challenge[13], which focused on textual entailment. In this context, sentences were classified as entailing (or not) the other sentence in the pair. Mutual/double entailment between sentences indicated a paraphrase. Like the MRPC, PASCAL judgments were human-validated.

Plagiarism detection corpora have also been created, as noted by Potthast *et al.*[14], who differentiated between real, simulated, and artificial forms of plagiarism. Although plagiarism and text reuse have distinct characteristics, they can still be seen as a form of paraphrase, as suggested by Madnani *et al.*[15], who use machine translation evaluation metrics to assess the precision of paraphrase detection. In another study, Gupta *et al.*[16] examine techniques for detecting crosslingual plagiarism, particularly when machine translation is used to plagiarize text written in another language. This form of plagiarism introduces unique challenges, as content is disguised through translation.

Over time, several paraphrase datasets have been compiled, many of which were based on translation, following the pioneering ideas of Barzilay and McKeown[7], but on a much larger scale. For example, Wieting and Gimpel[17] released a dataset containing 50 million paraphrased sentences.

Recently, this understanding has led to the annotation of paraphrase pairs at the phrase level. For example, Wang *et al.*[18] introduced ParaTag, a dataset specifically designed to capture fine-grained paraphrasing at the constituent phrase level. Interestingly, their study revealed that in almost 90% of instances labeled as paraphrases in the MRPC dataset, at least one phrase was missing from the paraphrased version. This indicates that while the general meaning remained, parts of the original sentence were frequently left out. Researchers have thus returned to emphasizing phrases after previously shifting from individual words and phrases to entire sentences.

[1] Jacquemin *et al.* [11, pp. 25] present the following example: "the word *genes* is tagged as a plural noun and morphologically connected to *genic*, *genetic*, *genome*, *genotoxic*, *genetically*."

[2] Available from https://www.microsoft.com/en-us/download/details.aspx?id=52398

Another important methodological breakthrough was brought in 2005 by Callison-Burch and associates. Bannard and Callison-Burch[19], Callison-Burch[20] used parallel corpora to find paraphrases, asserting that this resource is far more prevalent than multiple translations as proposed by Barzilay and McKeown, on the one hand, and, on the other hand, simpler to process than monolingual corpora, which require parsing and clustering, as used by Lin and Pantel[21]. Their approach offers a statistical measure to evaluate the quality of paraphrase candidates. Later, Callison-Burch and colleagues demonstrated the benefits of using paraphrasing to enhance machine translation performance[22]. In 2013, they released a large paraphrase database in English and Spanish[23], annotated with several contextual features and evaluation information.

After Callison-Burch doctoral studies, Barreiro's research[24] introduced the notion of bilingual/multilingual paraphrases, arising from aligning texts across one language pair (bilingual) or multiple language pairs (multilingual). She proposed the concept of a resource similar to a dictionary but designed for paraphrases at the multiword and phrasal levels, naming it "paraphrasary" ("parafrasário", in Portuguese). Using a revised version of Callison-Burch's Linear-B tool, called CLUE-Aligner[25], her approach identifies manually a set of paraphrastic units within parallel sentences, enabling analysis of equivalences not only across languages but also within Portuguese language varieties[26,27]. The methodology, further refined, detects non-contiguous paraphrastic units and aligns them for both paraphrastic and translation unit pairs[28].

Paraphrase evaluation faces two primary challenges: one concerning the assessment of semantic importance and the other regarding the influence or dependence of context. As Dras[29] aptly put it, paraphrases involve "a pair of text units considered interchangeable". But how to define what is interchangeable?

First, not all parts of a sentence carry the same weight when it comes to recognizing semantic equivalence or non-equivalence, as Madnani *et al.*[15] pointed out. Rather than viewing paraphrastic relationships as strictly binary, they suggested using a continuous scale between 0 and 1 to reflect varying degrees of similarity.

Another challenge is the issue of context. Assessing phrases or words outside their original context requires careful consideration of suitable independent contexts. The idea of simply relying on the original context, as done by Barzilay and McKeown, or Mota *et al.*, has faced criticism, notably from Callison-Burch in his PhD thesis. He argued that this method is too lenient, as it may permit phrases to be judged as paraphrases without adequately testing their versatility across different contexts.

The evaluation of paraphrases ranges from wholly subjective methods, where individuals rate or rank paraphrases based on their perceived closeness, to more "objective" approaches that involve quantifying the frequency of words identified as paraphrases across various texts and languages. Other techniques include analyzing the similarity of n-grams or lexical units between sentences, or using word embeddings. However, a central question persists: is there a universally accepted definition of what constitutes a paraphrase? Despite progress in these evaluation methods, reaching a consensus on this issue continues to be a challenge.

3 Interconnections between paraphrasing and translation

In this section, we pinpoint and elucidate the diverse connections between paraphrasing and translation within computational linguistics, classifying them in five cases: (1) using translation to obtain paraphrases; (2) using paraphrases to improve translation; (3) using paraphrases to evaluate machine translation; (4) using machine translation evaluation metrics to evaluate paraphrases; and (5) combining/mixing translation and paraphrasing in the same task.

Case 1 represents one of the most straightforward connections between paraphrasing and translation: variations that emerge from different (human) translations of the same source text. These translations naturally serve as paraphrastic candidates, when they occur at the sentence level, as shown by Barzilay and McKeown's work, or through retranslation of the same phrase or word across bilingual corpora, as Callison-Burch has proposed. Wieting and Gimpel[17] leveraged Czech-English machine translation using an English-Czech bilingual corpus to increase the volume of English paraphrases available.

Case 2 focuses on using paraphrases to improve (machine) translation quality by offering multiple ways to express the same meaning. As early as 1990, Santos [30, pp. 330] suggested using paraphrasing as a final step in the machine translation process to enhance output quality. The translation process was divided into two phases: "structural transfer", which performed the primary translation, and "style transfer", which involved selecting paraphrases in the target language[3].

Callison-Burch was the first to actually demonstrate the use of paraphrases to improve coverage in statistical machine translation (SMT). When direct translations are unavailable, paraphrases can fill the gap using existing translations from bilingual corpora, potentially across multiple languages. His approach enhanced statistical systems by expanding the number of translation options.

Barreiro[24] studied the efficiency of using short paraphrases of support verb constructions to enhance machine translation. She developed eSPERTo, a tool for paraphrasing and translation tasks, originally known as SPIDER[31], incorporating lexicon-grammar resources within a human-in-the-loop framework[32]. Then, Barreiro and colleagues introduced a

[3] Although the "style transfer" module was never fully implemented — since the initial translation option was typically accepted without further refinement — the concept laid the groundwork for future applications. A practical example of how this approach could work in generating translations is demonstrated by Stede[33].

method for enhancing machine translation by using monolingual parallel corpora from two varieties of Portuguese, as presented in Barreiro and Mota[26].

Case 3 provides another crucial link between paraphrasing and translation research, namely the use of paraphrases to evaluate machine translation. Notably, BLEU[34], a well-known machine translation evaluation metric, uses reference translations treated as paraphrases to assess machine-generated translations, measuring their closeness to human translations. As Madnani *et al.* [15, pp. 182] note:

> the task of an MT metric is essentially one of identifying whether the translation produced by a system is a paraphrase of the reference translation.

Unlike typical paraphrase evaluations, which often consider individual paraphrases, BLEU evaluates translations comparatively. While guidelines for diversity in reference paraphrases are lacking, researchers have explored various facets of evaluating paraphrases in their respective studies. In 2023, Kanayama[35] introduced a novel approach to evaluating Japanese machine translation output by proposing the use of paraphrasing rules instead of traditional n-gram based evaluation metrics.

Case 3 transitions into Case 4, where machine translation evaluation metrics are repurposed to assess the quality of paraphrases. This shift leverages the metrics typically used for translation evaluation to gauge the similarity and effectiveness of paraphrased expressions, helping measure the degree of semantic alignment between original and paraphrased texts. Quirk *et al.*[36] introduced the concept of "monolingual machine translation", while Finch *et al.*[12] adapted BLEU for paraphrase evaluation. Similarly, Wan *et al.*[37] combined BLEU's features with dependency analysis to improve paraphrase assessment. Building on these efforts, Madnani *et al.*[15] broadened this scope by applying eight different machine translation evaluation metrics to evaluate paraphrase identification. These works highlight the growing inter-section between paraphrase assessment and machine translation evaluation, producing another implicit argument for the closeness of both tasks.

Finally, case 5 includes broader applications that integrate both translation and paraphrasing without any hierarchical preference. An example is the multilingual plagiarism detection system discussed by Gupta *et al.*[16], which employs both translation and paraphrasing techniques.

4 Closeness of meaning

Both translation and paraphrasing are intrinsically linked to the idea of "closeness" or "sameness" of meaning. Over time, several researchers in paraphrase have invoked sameness. Kukich *et al.*[6] discussed the concept of paraphrases sharing "the same semantic information". Barzilay and McKeown[7] noted that "paraphrases are different ways of expressing the same information". Finch *et al.*[12] referred to paraphrasing as "Sentence-level Semantic Equivalence". Wieting *et al.*[38] described a paraphrase table as containing pairs of text snippets "that have the same meaning". And Globo *et al.*[39] stated that paraphrasing is rewriting "without altering the meaning of the original content".

Others emphasized closeness of meanings: Barzilay and McKeown[7] described paraphrases as having "approximate conceptual equivalence". Finch *et al.*[12] discussed whether two sentences are "close enough in meaning".

Qiu *et al.*[40] argued that paraphrase recognition (PR) operates on the concept of "relaxed meaning", which allows sentence pairs with minor variations to still qualify as paraphrases. They proposed that PR can be understood as an extension of synonym detection, but at the sentence level, and acknowledged that absolute synonymy is rare in natural language, as nuances of meaning are frequently added or removed in the paraphrasing process. They claimed:

> [...] for many people the notion of paraphrases also covers cases in which minor or irrelevant information is added or omitted in candidate sentences, [...]. Such extraneous content should not be a barrier to PR if the main concepts are shared by the sentences.

So we conclude that the challenge lies in identifying which elements constitute the "core concepts" and distinguishing them from details that can be safely omitted or added without altering the fundamental meaning. We thus turn to research that seeks to formalize the concept of "closeness of meaning".

Edmonds and Hirst [41, pp. 111] introduced the idea of "near-synonymy" precisely when addressing translation. They argued that the closest word in the target language is often a near-synonym rather than an exact synonym.

> Near-synonymy rather than synonymy is the norm in lexical transfer in translation: the word in the target language that is closest to that in the source text might be a near-synonym rather than an exact synonym.

Already in 1967, Catford[42] had observed that translations often do not convey the exact same message, but instead use different linguistic resources to achieve a comparable outcome.

Kay[43], cited in Edmonds and Hirst[41], introduced the concept of "elementary meanings", the core semantic units of a sentence, stripped of peripheral elements such as connotations or idiomatic quirks: "discrete objective denotations uncolored by [...] peripheral aspects such as connotations, implications, or quirks of idiomatic usage".

Edmonds and Hirst[41] extended Kay's concept by proposing a model for encoding near-synonymy, presented as a key characteristic of natural language. Their model consists of two components: (1) a language-neutral ontology of coarse-grained representations, essentially clusters of near-synonyms that are context-independent; and (2) a set of differences

between these near-synonyms that capture relations to context, fuzziness, and degrees of necessity. These differences, which they termed a "subconceptual level", highlight subtle variations between words that go beyond core meaning. In their model, these differences are expressed through concepts, attitudes, and stylistic dimensions, which can vary across languages. This representation of near-synonymy accommodates a wide range of linguistic distinctions without limiting the kinds of nuances that can be conveyed. Their graphically illustrated examples (e.g., distinguishing between the English nouns *blunder* and *error*, and the English verbs *order* and *enjoin*) show that encoding such distinctions requires a substantial amount of information, which they acquired from existing thesauri.

From a different perspective, but along similar lines, Wierzbicka's Natural Semantic Metalanguage (NSM) approach, as discussed in works like Goddard and Wierzbicka[44], Wierzbicka[45], and Wierzbicka[46], sought to express complex meanings by breaking them down into configurations of semantic elements using a limited set of basic, universally translatable words known as semantic primes. Interestingly, the defense of this approach also draws on the concept of paraphrasing:

> The NSM analysis of meaning is based on "reductive" paraphrase, in the sense that complex meanings are "reduced", in a systematic way, to simple or simpler ones. It attempts to say "the same thing" in a paraphrase composed of maximally simple, intelligible, and cross-translatable words (semantic primes), thereby laying bare the semantic content compressed in the original expressions.

Finally, another approach to handling complex meanings, from very distant quarters, is the method used for answering definition questions through informational "nuggets", as introduced by the QA track of TREC[47]. In this framework, the evaluation of an automated response to a complex question is based on how effectively it captures key pieces of information, or "nuggets", that are deemed essential to a complete and accurate answer (and which were previously listed by the organizers).

We thus note that scholars like Edmonds and Hirst, Wierzbicka and Voorhees have developed (different) fine-grained meaning representations, which should enable a more precise understanding of both translation and paraphrasing. In all of them, meaning proximity is crucial, but discrete, and we believe this is the way to proceed.

Ethics and consent

We confirm that this study did not involve interviews, workshops, or surveys with individuals, ensuring adherence to ethical standards for research involving human subjects.

Data and software availability

No data or software were developed in this study.

References

1. Baker M: **In other words: a coursebook on translation**. Routledge, First edition: 1991. 1998.
Reference Source
2. Barreiro A, Baptista J, Vieira R, *et al.*: **Por Outras Palavras POP1 @ PROPOR 2018, in other words: 1st workshop on linguistic tools and resources for paraphrasing in Portuguese.** *Linguamática.* Workshop co-located with the 13th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2018). Canela, Brazil, 2018; **10**(2).
Reference Source
3. Barreiro A, Baptista J, Vieira R, *et al.*: **Por Outras Palavras POP2 @ PROPOR 2020, in other words: 2nd workshop on linguistic tools and resources for paraphrasing in Portuguese.** *Workshop co-located with the 14th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2020)*. Colégio do Espírito Santo, Évora, Portugal, 2020.
Reference Source
4. Benjamin W: **Die Aufgabe des Übersetzers**. In: Tillman Rexroth, editor, *Gesammelte Schriften, 4. Kleine Prosa, BaudelaireÜbertragungen*. Suhrkamp, 1921; 9–21.
Reference Source
5. Smadja F, McKeown K: **Using collocations for language generation.** *Comput Intell.* 1991; **7**(4): 229–239.
Publisher Full Text
6. Kukich K, McKeown K, Shaw J, *et al.*: **User-needs analysis and design methodology for an automated document generator.** In: *Proceedings of the Fourth Bellcore/BCC Symposium on User-Centered Design*. 1993.
Reference Source
7. Barzilay R, McKeown KR: **Extracting paraphrases from a parallel corpus.** In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2001; 50–57.
Publisher Full Text
8. Lapata M: **A corpus-based account of regular polysemy: the case of context-sensitive adjectives.** In: *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. 2001.
Publisher Full Text
9. Quirk C, Dolan B, Brockett C: **Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources.** In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. 2004; 350–356.
Publisher Full Text
10. Levenshtein V: **Binary codes capable of correcting deletions, insertions, and reversals.** *Soviet Physics Doklady.* 1966; **10**: 707–710.
Reference Source
11. Jacquemin C, Klavans JL, Tzoukermann E: **Expansion of multi-word terms for indexing and retrieval using morphology and syntax.** In: *Proceedings of the 35th Annual Meeting of the ACL*. ACL, 1997; 24–31.
Publisher Full Text
12. Finch A, Hwang YS, Sumita E: **Using machine translation evaluation techniques to determine sentence-level semantic equivalence.** In: *Proceedings of the Third International Workshop on Paraphrasing*. 2005; 17–24.
Reference Source
13. Dagan I, Glickman O, Magnini B: **The pascal recognising textual entailment challenge.** In: Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*.

Springer Berlin Heidelberg, 2006; 177–190.
Publisher Full Text

14. Potthast M, Stein B, Barrón-Cedeño A, *et al.*: **An evaluation framework for plagiarism detection.** In: *Coling 2010: Posters*. 2010; 997–1005.
Publisher Full Text

15. Madnani N, Tetreault J, Chodorow M: **Re-examining machine translation metrics for paraphrase identification.** In: Eric Fosler-Lussier, Ellen Riloff, and Srinivas Bangalore, editors, *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada, Association for Computational Linguistics, June 2012; 182–190.
Publisher Full Text

16. Gupta P, Barreiro A, Rosso P: **Paraphrastic cross-language plagiarism detection with semantico-syntactic unit mapping.** In: *In proceedings of the I Workshop en Procesamiento Automatizado de Textos y Corpora (WoPATeC-2012)*. Viña del Mar Chile, 2012.

17. Wieting J, Gimpel K: **ParaNMT-50M: pushing the limits of paraphrastic sentence embeddings with millions of machine translations.** In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* ACL, 2018; 451–462.
Reference Source

18. Wang S, Xu R, Liu Y, *et al.*: **ParaTag: a dataset of paraphrase tagging for fine-grained labels, nlg evaluation, and data augmentation.** In: Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. December 2022; 7111–7122.
Publisher Full Text

19. Bannard C, Callison-Burch C: **Paraphrasing with bilingual parallel corpora.** In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 2005; 597–604.
Publisher Full Text

20. Callison-Burch C: **Paraphrasing and translation**. PhD thesis, University of Edinburgh, 2007.
Reference Source

21. Lin D, Pantel P: **Dirt @SBT@discovery of inference rules from text.** In: *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*. 2001; 323–328.
Publisher Full Text

22. Callison-Burch C, Koehn P, Osborne M: **Improved statistical machine translation using paraphrases.** In: *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2006)*. 2006; 17–24.
Reference Source

23. Ganitkevitch J, Van Durme B, Callison-Burch C: **PPDB: The Paraphrase Database.** In: Lucy Vanderwende, Hal Daumé III and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia, Association for Computational Linguistics, June 2013; 758–764.
Reference Source

24. Barreiro A: **Make it simple with paraphrases: automated paraphrasing for authoring aids and machine translation**. PhD thesis, Universidade do Porto, 2009.
Reference Source

25. Barreiro A, Raposo FA, Luís T: **CLUE-Aligner: an alignment tool to annotate pairs of paraphrastic and translation units.** In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, ELRA, 2016.
Reference Source

26. Barreiro A, Mota C: **A multilingual paraphrasary of multiwords.** In: Anabela Barreiro, Max Silberztein, Elena Lloret, and Marcin Paprzycki, editors, *Proceedings of the 1st international workshop on multilingual, multimodal and multitask language generation.* Tampere, Finland, European Association for Machine Translation, June, 2023; 47–56.
Reference Source

27. Barreiro A, Rebelo-Arnold I, Mota C: **A variety-based paraphrasary for Portuguese parafrasário: a case study with the varieties of Portugal and Brazil.** *Open Res Eur.* 2024; **4**(110).

28. Barreiro A, Batista F: **Machine translation of non-contiguous multiword units.** In: Wolfgang Maier, Sandra Kübler, and Constantin Orasan, editors, *Proceedings of the workshop on discontinuous structures in natural language processing.* San Diego, California, Association for Computational Linguistics, June, 2016; 22–30.
Publisher Full Text

29. Dras M: **Tree adjoining grammar and the reluctant paraphrasing of text.** PhD thesis, Macquarie University, 1999.
Reference Source

30. Santos D: **Lexical gaps and idioms in machine translation.** In: Hans Karlgren, editor, *Proceedings of COLING' 90.* 1990; 330–335.
Publisher Full Text

31. Barreiro A: **SPIDER: A System for Paraphrasing in Document Editing and Revision — Applicability in Machine Translation Pre-editing.** In: Alexander Gelbukh, editor, *Computational linguistics and intelligent text processing: 12th international conference, CICLing, Tokyo, Japan, February 20-26, 2011. proceedings, part II.* Berlin, Heidelberg, Springer Berlin Heidelberg, 2011; **6609**: 365–376.
Publisher Full Text

32. Barreiro A, Mota C, Baptista J, *et al.*: **Linguistic resources for paraphrase generation in Portuguese: a lexicon-grammar approach.** *Lang Resources & Evaluation.* March, 2022; **56**(1): 1–35.
Publisher Full Text

33. Stede M: **Lexical choice criteria in language generation.** In: Steven Krauwer, Michael Moortgat, and Louis des Tombe, editors, *Sixth conference of the european chapter of the association for computational linguistics.* Utrecht, The Netherlands, Association for Computational Linguistics, April, 1993.
Reference Source

34. Papineni K, Roukos S, Ward T, *et al.*: **Bleu: a method for automatic evaluation of machine translation.** In: Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, July, 2002; 311–318.
Reference Source

35. Hiroshi K: **paraphrasing rules for automatic evaluation of translation into Japanese.** In: *Proceedings of the second international workshop on paraphrasing (ACL 2003).* 2003; **16**: 88–93.
Publisher Full Text

36. Quirk C, Brockett C, Dolan W: **Monolingual machine translation for paraphrase generation.** In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004).* 2004; 142–149.
Reference Source

37. Wan S, Dras M, Dale R, *et al.*: **Using dependency-based features to take the "para-farce" out of paraphrase".** In: *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006).* 2006; 131–138.
Reference Source

38. Wieting J, Bansal M, Gimpel K, *et al.*: **From paraphrase database to compositional paraphrase model and back.** *Trans Assoc Comput Linguist.* 2015; **3**: 345–358.
Publisher Full Text

39. Globo A, Trevisi A, Zugarini A, *et al.*: **Neural paraphrasing by automatically crawled and aligned sentence pairs.** In: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).* 2019; 429–434.
Publisher Full Text

40. Qiu L, Kan MY, Chua TS: **Paraphrase recognition via dissimilarity significance classification.** In: Dan Jurafsky and Eric Gaussier, editors, *Proceedings of the 2006 Conference on empirical methods in natural language processing.* Sydney, Australia, Association for Computational Linguistics, July, 2006; 18–26.
Publisher Full Text

41. Edmonds P, Hirst G: **Near-synonymy and lexical choice.** *Comput Linguist.* 2002; **28**(2): 105–144.
Publisher Full Text

42. Catford JC: **A linguistic theory of translation: an essay in applied linguistics.** Oxford University Press, 1967.
Reference Source

43. Kay MW: **Webster's Collegiate Thesaurus.** Merriam-Webster, 1988.
Reference Source

44. Goddard C, Wierzbicka A: **Introducing lexical primitives.** In: Cliff Goddard and Anna Wierzbicka, editors, *Semantic and Lexical Universals.* John Benjamin Publishing, 1994; 31–54.
Publisher Full Text

45. Wierzbicka A: **Language, culture and meaning: Cross-cultural linguistics.** In: *Cognitive Exploration of Language and Linguistics.* John Benjamins Publishing Company, 1999; 137–159.

46. Wierzbicka A: **'Story' - an English cultural keyword and a key interpretive tool of Anglo culture.** *Narrat Inq.* 2010; **20**(1): 153–181.
Publisher Full Text

47. Voorhees EM: **Overview of the TREC 2003 question answering track.** In: *Proceedings of TREC-2003*. 2003.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 12 March 2025

https://doi.org/10.21956/openreseurope.20679.r51650

✔ **Eman Abdel Reheem Amin**
Majmaah University, Al Majma'ah, Saudi Arabia

Summary of the Article:
The article explores the intricate relationship between paraphrasing and translation, emphasizing the necessity of maintaining semantic and stylistic proximity to the original text. It presents a well-researched discussion on how paraphrasing can either support or challenge translation fidelity, particularly in contexts where linguistic and cultural nuances play a critical role.

The author systematically examines existing literature, theoretical frameworks, and real-world examples, offering a thorough and insightful analysis of the topic. this article is a well-structured, well-researched, and clearly articulated contribution to the study of paraphrasing and translation.

The discussion is thorough, the arguments are well-supported, and the conclusions are appropriately drawn. No major concerns were found, and the paper is well-positioned to contribute valuable insights to the field.

I recommend the article for indexing in bibliographic databases after standard editorial checks.

Does the review offer a comprehensive analysis of the research literature?
Yes

Are all statements factually correct and adequately supported by citations?
Yes

Is the review written in clear language?
Yes

Are the conclusions drawn appropriate in the context of the current literature, books and book chapters?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Applied lingustics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 11 March 2025

https://doi.org/10.21956/openreseurope.20679.r51654

Stella Markantonatou
Athena Research Center, Athena, Greece

The paper makes an interesting review of the literature on paraphrasing in computational linguistics with two main goals, one explicitly stated and one served implicitly. Explicitly stated is the goal of highlighting the conceptual similarities of paraphrasing and translation, as they both start from a “given” text and try to develop texts that simultaneously differ in form from the “given” one and convey the same meaning with it. Of course, translation works across languages while paraphrasing within the same language as the “given” text, which may simplify the task to some extend. The implicit goal is to identify the features of good paraphrases and develop a theory for them; the authors seem to consider theories relying on semantic primitives as a proposal to this end.

The authors go through important points in the history of paraphrasing in computational linguistics and explain in sufficient detail how the field evolved following the developments in NLP and came close to research on translation and eventually shared with it materials and evaluation techniques. In particular, NLP techniques based on parallel corpora allowed for searching for the naturally occurring different textual forms in a corpus of language A that were mapped on to the same translation in a language B; these texts were evaluated by humans as to whether they constituted paraphrases of each other. This idea was refined with various techniques and evaluation measurements but the main problem remained: could the texts evaluated as paraphrases substitute one another in different contexts? Which properties of the texts would determine substitutability? In simpler and somehow more vague terms, which properties of the texts are responsible for good paraphrases? This is what we have called the implicit goal of the article.

Paraphrasing relies on the hard-to-define notion of “sameness” or “equivalence” of meaning carried by texts different in form. An important finding is that the parts of paraphrases may not stand in 1:1 relation among them, rather some parts are more important while others may be omitted. A theory of “sameness” should be able to explain such observations. The authors believe that theories based on semantic primitives allowing for a direct description of meaning, are best candidates for this purpose. Theories of semantic primitives, though, have never managed to

achieve wide acceptance because they failed to convince that the defined primitives were enough or the right ones. This is a very old problem, that has motivated various approaches of indirect description of word meaning, e.g., WordNet uses Synsets (where meaning is described by lists of synonyms) while in state-of-the-art NLP, embeddings describe word/phrase meaning by the contexts in which a word/phrase occurs.

We would like to make two observations. First, the article makes an unnecessary, in our view, reference to the Chomskyan distinction between competence and performance. These notions apply to human linguistic abilities and not to computational systems, unless the authors imply that the systems they examine somehow simulate human cognition. The distinction between paraphrase generation and paraphrase discovery, also made by the authors, is enough and suitable for the systems that the paper examines. And second, the paper does not put its topic under the perspective of LLMs. Are paraphrases interesting in a world where LLMs dominate NLP or are they a necessity of the technology before LLMs?

In the overall, this is an informative and clearly written article, that conveys well-organised information even if its readers do not agree with the generalisations and the preferences of the authors.

Does the review offer a comprehensive analysis of the research literature?
Yes

Are all statements factually correct and adequately supported by citations?
Yes

Is the review written in clear language?
Yes

Are the conclusions drawn appropriate in the context of the current literature, books and book chapters?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: NLP, ontologies

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.