Check for updates

CASE STUDY

OpenLogos in the context of custom language models for the business sector: A historical perspective on the roots of natural language generation

[version 1; peer review: 1 approved, 1 approved with reservations]

Anabela Barreiro[1], Lou Cremers[2]

[1]Human Language Technology Laboratory, INESC-ID Lisboa, Lisbon, Portugal
[2]Syn-tactic Language Technology, Venlo, Netherlands Antilles

v1 **First published:** 28 May 2024, **4**:104
https://doi.org/10.12688/openreseurope.16866.1
Latest published: 28 May 2024, **4**:104
https://doi.org/10.12688/openreseurope.16866.1

Abstract

In the framework of Multi3Generation COST Action (CA18231), we have provided the OpenLogos tools and resources to the community with the belief that (i) they can be beneficial to Early Research Career researchers and (ii) they can also assist startup companies with limited resources in initiating endeavors to create their Natural Language Generation, Machine Translation, Paraphrasing or other Natural Language Processing systems. In this article, we explore OpenLogos capabilities in an industry case scenario. We provide historical background on the proven importance of the Logos commercial system in shaping the documentation process of a multinational. Understanding the historical significance of Machine Translation is crucial for grasping our present position in Generative Language Models and our role in the domain of Generative Artificial Intelligence. We explore how novel Custom Generative Language Models, drawing inspiration from the Logos Model, can prove beneficial across various industries worldwide.

Plain language summary

Natural Language Generation has a rich history rooted in the pioneering field of Machine Translation, which was among the earliest and most intricate language applications, marking a significant milestone in language technology's evolution. The first Natural Language Processing systems that made an entry in the market were translation systems, being Logos the oldest and most thoroughly documented commercial model. Recognizing the historical significance of Machine Translation (and the Logos Model) in human language technology is essential for understanding the current status of Generative Large Language Models and our standing in the domain of Generative Artificial Intelligence. The crucial significance of this

Open Peer Review

Approval Status ✓ ?

	1	2
version 1 28 May 2024	✓ view	? view

1. **Jeffrey Allen**, SAP, Levallois-Perret, France
2. **Amparo Alcina**, Universitat Jaume I, Castellón de la Plana, Spain

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

article stems from its historical context, particularly addressing a gap in Machine Translation development regarding the linguistic processes of recent Natural Language Generation models. Initially, we highlight the challenges associated with language generation in the industry, encompassing ethical considerations and the preservation of corporate language integrity. Subsequently, we examine an early commercial Machine Translation system (Logos) that incorporated a robust and transparent NLG component, showcasing human intervention at each translation stage. Finally, we suggest the development of alternative or complementary models inspired by the Logos Model, aiming to enhance the transparency and traceability of Natural Language Generation systems, particularly in applications like Machine Translation and Paraphrasing.

Keywords
Natural Language Generation; Custom Generative Language Models; Machine Translation; Corporate Language; Multi3Generation COST Action; Logos Model; OpenLogos

This article is included in the COST Actions gateway.

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

Corresponding author: Anabela Barreiro (anabela.barreiro.papers@gmail.com)

Author roles: Barreiro A: Conceptualization, Formal Analysis, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Cremers L**: Formal Analysis, Methodology, Software, Validation, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: This article is based upon work from COST Action Multi3Generation (CA18231) supported by COST (European Cooperation in Science and Technology).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Barreiro A and Cremers L. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Barreiro A and Cremers L. **OpenLogos in the context of custom language models for the business sector: A historical perspective on the roots of natural language generation [version 1; peer review: 1 approved, 1 approved with reservations]** Open Research Europe 2024, **4**:104 https://doi.org/10.12688/openreseurope.16866.1

First published: 28 May 2024, **4**:104 https://doi.org/10.12688/openreseurope.16866.1

1 Introduction

Language models are here to remain, with the potential to yield positive or negative consequences. They symbolize the modern revolution within societies, and the scope of their impact varies from minor to substantial, depending on geographical regions and specific fields of human activity. The emerging Natural Language Generation (NLG) era is marked by a blend of challenges and opportunities, along with risks and benefits, mirroring the core of human nature and its ability to generate and understand language, the trait that distinguishes humans from other species. This is largely attributed to the inherent flexibility and comprehensive nature of language, which spans a wide spectrum, embracing diverse subjects and genres, encompassing variations in style, objectivity or subjectivity, technical and non-technical terminology, and the full range of expression clarity[1]. For the industry sector, the success of a company's communications depends on the quality of its corporate language. The first image of a company is provided by its language, in the way that it is projected on the company's website, in the production of documentation, in the way its staff express themselves, and so on.

Several years ago, the OpenLogos tools and resources became available to the community through SourceForge[1] [2]. Presently, within the framework of the Multi3Generation COST Action (CA18231)[2] [3], we are broadening their accessibility via GitHub[3] [4]. We are proceeding with the hope that these tools and resources will be beneficial for educational purposes concerning Generative Large Language Models (GenLLMs). Our target audience is Early Research Career (ERC) researchers and startup companies initiating projects in NLG, Generative Artificial Intelligence (GenAI), Machine Translation (MT), paraphrasing, or other Natural Language Processing (NLP) areas, where they often encounter constraints due to limited or restricted resources.

The Multi3Generation COST Action (CA18231), referred to as Multi3Generation hereafter, encourages collaborative endeavors to amplify the significance of the NLG research domain for the benefit of society at large. It emphasizes inclusivity by considering all valuable resources and methods, regardless of whether they stem from Large Language Models (LLMs) or more linguistically-oriented models. This article places a strong emphasis on using the carefully curated linguistic resources developed for a commercial system inspired by the Logos Model[5–7]. Linguistic models are defined by their incorporation of strategies that involve human participation (human-in-the-loop), allowing human intervention in any phase of the NLG process. Hybrid methodologies combining LLMs with human-centered approaches contribute to transparent NLG models that are both adaptable and comprehensible to a wide audience. Within the language spectrum specified in Barreiro *et al.*[1,4], in the Logos Model, there are several different levels of user engagement, which constitute a smaller or greater challenge to NLG. In addition, the OpenLogos resources include knowledge that contributes to language understanding. This article addresses the challenges it poses in the realm of industry.

The rest of the paper is organized in the following way: Section 2 discusses the impact and challenges of NLG on the industry, which may go from ethical issues to the preservation of the integrity of the corporate language. Section 3 showcases the implementation of MT in Océ Technologies using the commercial Logos system. This is the main Section of the paper, aiming to offer a historical overview of how the Logos Model was tailored to translate corporate language into several languages. Section 4 suggests the development of Custom GenLLMs based on OpenLogos. Section 5 presents a simple test designed to clarify the characteristics of a generative system like eSPERTo, which is based on OpenLogos, when compared to ChatGPT. Finally, Section 6 presents conclusions and future work.

2 Natural Language Generation in an industrial context

LLMs are becoming useful for many people in the professional world and new tools are gaining an increasing amount of followers in the industry. Microsoft released the Azure AI Studio, allowing the use of OpenAI models like ChatGPT and GPT-4 with companies' data. These systems can help less skilled business people respond to customers' emails, and analyze and present business data in an expert-like manner. The language used by the private sector and the information conveyed by the corporate world need to meet quality standards (cf. non-conformance report (NCR)). For less skilled users, NLG tools deliver impressively detailed human-like text and they represent a significant support to those users. In this regard, tools that perform better than humans can provide small companies with some competitive chances. Nevertheless, content creation by GenAI tools can be either a blessing or a curse in the business world. Not only the ethical aspects and concerns of the new models have been a topic of discussion that needs to be addressed with the seriousness and prudence that they deserve, but also aspects related to the regulation or certification, not so much of GenAI itself, but of the parties involved in creating and developing the tools of the future, especially when these tools are used professionally[4].

Another important aspect to be taken into consideration concerning NLG models is related to corporate identity and copyrights. Larger companies have the notion that their main asset is their corporate language. Allowing an

[1] https://sourceforge.net/projects/openlogos-mt/

[2] https://multi3generation.eu/

[3] https://openlogos.github.io/

[4] Questions arise regarding whether professionals have received adequate training and possess the responsible mindset and qualifications necessary to conduct their work ethically and fairly within their respective industries.

external system to use such important companies' assets may be risky in the loss of their identity. Compliance with information security (ISO) standards is essential for companies that make their data available to third parties to be relaxed about using these parties' LLMs without the risk of losing their intellectual property to others. To further extrapolate, the negative impacts of this on society, we are aware that the channels of communication can be easily concentrated into a few hands of corrupt people in more primitive societies, and technologies can dangerously turn into tools of totalitarian control in non-democratic regimes. We recognize that, if used wisely, they can also provide new opportunities in the workplace and businesses across the planet. The problems to be addressed are the development of monopolies and the marginalization of those most in need of these tools. Ideally, NLG models should promote equality and be employed democratically, ensuring that training and inclusion are prioritized for those who require them the most.

The various subsections of Section 3 meticulously elaborate on the processes developed over the years at Océ Technologies, guided by robust support from the Logos MT system. We seize this opportunity to showcase and trace the roots of language generation, namely the generative characteristics of the Logos Model. The Logos tools handle the majority of the tasks, but human intervention is required to ensure reliability, understandability, and controllability.

3 Logos relevance to industry language models

3.1 Historical Background

Back in the nineties, Logos MT initiated a sequence of advancements that fundamentally revolutionized the documentation and localization procedures at a Dutch multinational company, Océ Technologies, culminating in a highly advanced and entirely automated localization process. Over a decade of translation automation, starting in 1995, Océ discovered that a sophisticated Rule-Based Machine Translation (RBMT) system like Logos is not an off-the-shelf product that merely needs to be installed and will then perform its miracle. It takes a significant effort by dedicated and committed individuals to train the system, adjust the translation workflow, and adapt the source material to accommodate MT. Companies that acquired the system but lacked the resources or the expertise would ultimately experience failure. Nonetheless, when implemented effectively, MT, in conjunction with various associated technologies, resulted in substantial cost savings, amounting to millions of Euros, within the internal localization department of the multinational Océ Technologies, as illustrated during that period. Despite achieving significant success and documented cost reductions, the internal localization endeavor was terminated in early 2006 due to poorly motivated management decisions which culminated in the influence of a few individuals with different objectives who succeeded in outsourcing the entire localization process to a major vendor. In the remainder of this article, we will describe how the introduction of Logos MT initiated a sequence of organic developments that revolutionized the processes of documentation creation and translation. For a decade, various new components were incorporated, such as Translation Memory, a Terminology Management System, Controlled Language, a fully automated translation workflow, and an XML-based Content Management System.

At Océ Technologies a group of researchers had been working on NLP for several years. In 1994, they received a request from the International Training Centre (ITC) inquiring if there was a way to address the substantial expenses incurred in translating service and user documentation. Within a year, a prototype MT system was developed, capable of processing an English FrameMaker service manual and generating a corresponding German version. The existing parser from prior projects could be recycled, and a transformational component from

Radboud University in Nijmegen became accessible. The outcomes showed great potential and persuaded the management to consider implementing it at the Information Technology and Communication (ITC) center. At the 'Translating and the Computer' conference in 1994, Logos Eschborn office representatives showcased the Logos system. After comparing the outcomes of the Océ system with those of Logos, it was collectively agreed that it was a worthwhile endeavor to enhance Logos with the required vocabulary and rules and conduct another test. In 1995, the decision was made to procure the Logos system instead of attempting to create an in-house prototype for a comprehensive multi-language end-user system.

3.2 Introducing Logos MT

As Océ already had an established workflow, the biding approach was to integrate MT at the final stage of the process, considering that the technology was still unproven. In the initial months, most efforts were dedicated to the compilation of terminology, inputting data into the system, refining and adapting the grammatical rules, and conducting test translations. Over time, the team gained insights into sentence structures and phrasings that posed challenges for Logos (or MT in general) and began tailoring the source text of the test project (a service manual) to better align with the requirements of MT. When the initial well-prepared manual was translated using Logos and subsequently submitted for post-correction (post-editing), it became evident that the translator needed only 25% of the usual time to produce the final translation. Achieving a 75% reduction in time was a notable success. The team was motivated to create a workflow that would facilitate the step-by-step translation of all service manuals using Logos.

3.3 Terminology management

The maintenance of domain-specific terminology stands out as the most crucial factor for the successful operation

of MT. In any type of MT system, whether rule-based, statistical, or neural, The system must be familiar with the terminology present in the text. Accurate and consistent terminology is essential for all systems to generate precise and consistent translations. Translating documents solely relying on general language dictionaries typically yields readable translations, suitable for grasping the main content ('gisting'). However, without domain-specific terminology, editing these documents for publication becomes a time-consuming task. Any unknown or ambiguous word can result in a wrong translation (mistranslation in Logos terminology). The use of domain-specific dictionaries enables the system to generate 100% accurate translations, even for complex sentences.

Particularly in the initial stages, it required several months to develop bilingual glossaries, either by extracting terminology from translation memories or by processing pre-existing digital dictionaries or glossaries. Most (commercial) systems typically need to include basic information like source word, target word, part-of-speech, and gender. The Logos semantic engine goes beyond, enabling the inclusion of detailed semantico-syntactic and contextual information, such as whether a word represents something that is animate or inanimate or if it is a cylinder-like object or has any specific object shape, and so on. Another powerful feature is the ability to preprocess source and/or target sentences by defining specific patterns, effectively preventing and correcting mistranslations. For example, the English verb like start has seven different translations in German, depending on the preposition that follows it, and another fourteen translations depending on context. The expression to start an enterprise is translated as gründen whereas to start machinery or equipment is translated as anschalten. Logos stands out as a highly powerful system due to its capability to facilitate the formulation of rules tailored for diverse translations in varying contextual settings.

The higher the investment in refining the rules and lexicon entries of a top-notch MT system, the superior the output quality and the reduced cost of post-editing. Certainly, a precisely established terminology foundation is crucial for all authoring and localization procedures and brings additional benefits: (i) improves the quality and uniformity of documentation; (ii) diminishes lengthy review cycles; (iii) prevents engineers from duplicating efforts (reinventing the wheel terminology-wise) or rediscovering established terminology; (iv) enhances customer satisfaction. When formulating domain-specific terminology for Logos, the team encountered typical terminological challenges: (i) overly complex terms, such as size deflector unit release lever; (ii) obvious synonyms (developing roller, developer roller, development roller); (iii) less obvious synonyms (bypass versus manual feed; (iv) product name inconsistencies (IntraLogic, Intralogic, Intra Logic); (v) one source, different translation synonyms (EN: image, PT: imagem, gráfico; non-translatables, i.e., phrases/terms, most named entities such as names (proper names, names of organizations, etc.) that require to be protected from translation (Engineering Exec). In the beginning, the term base was exclusively designed for Logos. Subsequently, it was transformed into a web-based intranet application. This extension enabled R&D developers and other Océ staff to look up and submit terms. Typically, term creation occurs in R&D, and an intranet terminology base offered a means to influence terminology creation at its origin (at the source level).

3.4 Machine Translation and post-editing

Upon initiating a translation process incorporating MT, we had to consider the translator's role. In the mid-90s, MT was not as widely accepted as it is today, and translators were somewhat reluctant to engage in post-editing instead of using traditional/conventional translation. Collaborating with a localization provider that viewed post-editing as a prospect/opportunity rather than a challenge/threat for the industry was crucial. The company Delta International in Germany demonstrated itself as such a provider. The success of any technology relied on the cooperation of translators. If translators were reluctant to cooperate, all technology would fail. Hence, Océ conducted sessions with translators to explain/clarify the meaning and purpose of MT. They were also shown the actual machines described in the manuals. The aim was to use the technology to maintain translation costs within acceptable limits, especially as the volume and number of target languages expanded.

Regarding the post-editing task, it enabled translators to focus on their expertise in domain and language, leaving the system to handle more routine tasks. Efforts were made to establish feedback loops where the experiences of translators could contribute to enhancing the rule system and dictionaries. However, this initiative faced challenges and was not ultimately successful, given the additional workload it imposed on the translators. The need arose to set up post-editing guidelines to prevent/refrain translators from enhancing/polishing MT outputs into prose. This precaution/measure was crucial to maintain the fundamental purpose of using MT in the first place. In particular, technical service manuals should undergo post-editing focused solely on ensuring grammatical correctness and clarity of the text. Post-editing is acceptable as long as the MT meets a certain basic quality standard. To prevent the potential disenchantment of translators, precautions were put in place, such as sending only suitable material through MT and the subsequent post-editing phase to avoid 'losing' translators in the process. The initial meticulously planned project that underwent MT demonstrated a remarkable 75% reduction in translation costs. Over the years, the average savings ranged from 50% to 60%, as there was typically limited time available for project preparation. Individuals reviewing machine-translated content anticipate flawless translations at a 100% accuracy level. However, for effective post-editing, the key factor determining efficiency is not the absolute number of errors in

the MT but rather the time saved by the translator in the post-editing process compared to creating a translation from scratch. Another perspective on MT is the common belief that it diminishes the workload for translators. At Océ, a positive side-effect for the translators was that the use of MT resulted in an increased workload as it facilitated the translation of extensive volumes of documentation that might not have been translated without the assistance of MT.

3.5 Translation Memory

Shortly after implementing MT, it became evident that there was a need for a module enabling the reuse of previously translated and approved text segments. Translation Memory tools were emerging, and XL8 was initially employed. However, XL8 proved challenging to integrate into an automated workflow and was subsequently replaced by Trados. In the beginning, Translation Memories (TMs) were not available. Therefore, the team developed tools that enabled the creation of TMs from an extensive collection of previously translated user and service manuals spanning approximately ten languages.

For a period, Trados was employed to analyze new documentation against translation memory. However, at a certain point, this process became excessively time-consuming, particularly with the expanding volume of translation memories. An in-house Translation Memory (TM) tool was developed to resolve the challenges faced with Trados. The newly developed tool could analyze new documentation against the complete database of TM memories and select the best-matching memory. It generated two sets of memories for the translator: one set comprised entirely of clean sentences (without any formatting) that matched the search set, and another set consisting of the original best-matching memories, including all the original formatting. This design eliminated the reliance on Trados formatting penalties, making the translation process more independent.

3.6 Controlled Language

Initially, guidelines were established for authors, with a particular focus on MT. Even simple rules made a significant impact, as demonstrated by the example in Table 1. The insertion of an article disambiguates the sentence and leads to a correct translation. The English source sentence paint surface between lines was translated in Italian as vernice di superficie tra le righe and in German as Farbeoberfläche zwischen Zeilen. The corrected source sentence paint the surface between the lines was translated in Italian as verniciare la superficie tra le righe, and in German as Malen Sie die Oberfläche zwischen den Zeilen.

To enhance the efficiency of MT, efforts were directed towards the front end of the process. Controlled Language (CL) emerged as a logical step to regulate input to an MT system by focusing on terminology and sentence structure. At that time, Océ opted for a CL system that was relatively easy to implement but imposed significant restrictions on sentence structures by adhering to the rules for Simplified English.

During the preparatory phase, a dictionary for Simplified English was developed, containing a meticulously selected set of technical domain terms. The implementation was accompanied by a training course explaining the principles behind CL. This course played a crucial role in overcoming the initial resistance of the technical writers. Interestingly, some writers became overly enthusiastic and engaged in "black" authoring, attempting to eliminate every colored error code from their text. However, this behavior was discouraged for several reasons. Firstly, the tool was not infallible, and "black" text did not necessarily mean error-free text. Additionally, it consumed too much time during the writing process. Although the adoption of CL initially posed time challenges, it brought about significant benefits: (i) consistent use of terminology; (ii) a standardized writing style; (iii) grammatical English from non-native writers; (iv) improved efficiency of the content management system because of text standardization and text re-use; (v) improved quality performance of the MT system; (vi) reduction of superfluous information; (vii) improved readability and intelligibility.

In broad terms, MT benefits from the application of Controlled Language since the rules generally align with those applicable to MT. Nevertheless, there are also notable drawbacks: (i) CL and MT pursue distinct goals and operate on different analysis engines; (ii) The English verb set encompasses numerous general verbs that pose challenges for translation within an MT system; (iii) Some writers deemed this particular implementation of CL to be excessively restrictive; (iv) Implementing CL demands

Table 1. Comparison between the translation of incorrect versus the translation of correct sentences.

English	Italian	German
Source Sentence	MT of Source Sentence	Corrected Sentence
paint surface between lines	vernice di superficie tra le righe	Farbeoberfläche zwischen Zeilen
Corrected Source Sentence	MT of Corrected Source Sentence	MT of Corrected Source Sentence
paint the surface between the lines	verniciare la superficie tra le righe	Malen Sie die Oberfläche zwischen den Zeilen

a significant initial effort; (v) The tool's error messages were occasionally unclear, incorrect, or inconsistent. Regarding legacy material, the company opted against rewriting it in CL. However, in the example in Table 2 where this was undertaken, there was a 35% reduction in word count.

3.7 Moving to a Content Management System

At some point in time, it became evident that transitioning the documentation creation process to an XML-based Content Management System (CMS) would significantly enhance the efficiency of both the localization and document creation processes. This shift allowed authors to focus on content rather than layout considerations. It was crucial to have the CMS in place before implementing CL because the CMS would contain generic, clearly defined, and well-written text segments from the outset, enabling their reuse across various products and publications. This approach simplified the localization process, dealing with XML instead of grappling with multiple complex document formats such as HTML, FrameMaker, and Word, among others. The separation of layout from content also had a 'cleaning' effect on TMs, eliminating formatting information that previously impacted negatively the matching process from the TM.

3.8 Machine Translation: some numbers

For nearly a decade, Océ Technologies used Logos to translate service manuals and e-learning material from English into French, Italian, German, and Spanish. Initially, e-learning material was chosen for translation using MT to control costs. Later, German was also employed as a source language following the acquisition of a German subsidiary of Siemens. In 2003, the translation volume reached 5.3 million source words, resulting in the translation of 26 million target words across up to 17 languages. The average savings on external translation costs for the entire workflow were approximately 65%, including MT.

3.9 Automating the workflow

Right from the start, a workflow and administrative system were established. The implementation of the workflow system was essential to minimize manual tasks. For example, the traditional process of pre-translating FrameMaker manuals using TM and MT, sending them for post-editing, and converting them back involved more than 300 individual steps, each susceptible to errors. Complicating matters further was the fact that errors often became evident only after the entire process was completed, including manual post-editing. These errors surfaced when attempting to revert to the original FrameMaker format. Automating the workflow was imperative due to the limited staff dedicated to the localization process, which needed to cope with increasing volume and tasks. The entire process was managed by a team of just four individuals, responsible for designing and developing workflow modules, localization tools, project management, research, and administration. For each format that required handling, a predefined sequence of tasks was established. The system, based on the type of source material, determined the necessary sequence of conversion tasks, automatically built the relevant TM, computed matching metrics, and initiated MT if deemed suitable. A translator package was generated, encompassing sources, TM, and instructions. Authorized staff could log into the workflow system, fill out the required information (deadlines, languages, encodings, source location, etc.), and the system would initiate the process accordingly. Since the inception of the project, it was essential to track the savings achieved with MT. Consequently, the team progressively developed an administrative system that not only automated parts of the process but also maintained records of costs and time schedules. The administrative segment of the workflow systematically recorded all tasks, computed key metrics for each project (such as word count, percentage of memory reuse, number of words processed by MT, project duration, costs, and assigned vendor), and generated an overview of ongoing translation projects. This overview included estimates of costs, savings, and delivery times, providing a comprehensive snapshot at any given time. Ultimately, one of the localization vendors opted not to conduct their calculations anymore but instead adopted the in-house ones. The administration played a crucial role initially to justify the acquisition of Logos. The return on investment was realized within the first year. To enhance the translator's efficiency within the workflow, Ophelia, a comprehensive web-based localization suite, was created. This suite included workflow modules and a robust translation editor designed to simplify and streamline the translator's tasks. Upon opening the translation package in the Ophelia editor, the translator gained immediate

Table 2. German translation of regular English sentence versus translation of Controlled Language English sentence.

Original Sentence (English)	Controlled Language Sentence (English)
In several windows, an icon shows the current status/activity of a printer. See the list below for a description of each status.	These icons show the status or activity of the copier.
Translation of the Original (German)	MT of the Controlled English Sentence (German)
In mehreren Fenstern zeigt ein Ikon den laufenden Status/Aktivität eines Druckers. Siehe die Liste zwecks einer Definition jedes Status unten.	Diese Ikone zeigen den Status oder die Aktivität des Kopiergeräts.

access to TM, terminology, and MT suggestions, with all analyses completed during package creation. Edits were automatically updated in the TM, and the editor provided contextual information and a real-time preview of the manual, facilitating the translation of GUI text.

3.10 Impact of Logos on Corporate Language

After nearly a decade of implementing Logos MT in a corporate production setting, it is evident that it has significantly influenced the entire documentation process and more. Among the positive impacts, there is: (i) acceleration of the translation process: Logos MT expedited the translation workflow; (ii) substantial cost reduction: The use of MT resulted in a considerable decrease in translation costs; (iii) enhanced terminological consistency: MT contributed to maintaining consistency in terminology; (iv) faster time to market: The implementation of MT contributed to quicker turnaround times for projects; (v) efficiency in large projects: Larger projects could be handled by fewer translators, leading to improved consistency. However, the efficiency of MT is contingent on the specific context (content-dependent) and is not universally applicable. The constraints are: (i) volume and technical nature of text: There needs to be a substantial volume of technical text within a specific domain for MT to be effective; (ii) dedicated effort: Achieving and sustaining high-quality MT requires dedicated efforts, including the creation of detailed bilingual glossaries, modification of syntactic or semantic rules, and integration into a comprehensive workflow; (iii) specialized staff requirement: Efficient operation of MT requires personnel with specialized skills; (iv) source text quality: The source text must be grammatically correct and of reasonable quality for optimal MT performance.

4 Generative language models based on OpenLogos

In the realm of NLG, the quality of data and the tools employed play pivotal roles. It is crucial to understand the construction, strengths, and limitations of these tools, ensuring human control, curation, and enhancement of the generative system's quality. The initial systems prioritize these procedures, they were based on design models that humans could understand as long as they were properly trained. In the early days, things were straightforward, data was scarce, computational power was limited, and the inception and principles were collectively adhered to diligently for the sake of good results. As engineering advanced, particularly in the realm of language engineering systems, the complexity and capabilities of these systems grew rapidly, often exceeding the ability to predict or control their behavior, which is a trait that is far from desirable. As a metaphor, the created should not surpass the creator, or when control is lost, it can lead to consequences affecting everyone and everything. In contemporary times, AI heavily depends on systems that often function autonomously, drawing upon extensive datasets of uncertain origin, and generating results that may not be thoroughly scrutinized, verified, or validated. This does not inherently suggest their inadequacy; rather, when equipped with appropriate resources and data, these systems have the potential to surpass human capabilities, aided by the exponential availability of data. The current trend of generative models leans towards rectifying errors after the generation process occurs, with many business models centering on this approach, namely the influx of developers building on the top of AI. For example, ChatGPT presents users with a final text whose construction may be grammatically correct and semantically sound, but may not suit the purpose of the user who needs to reformulate the text. The user needs to refine and edit unclear and/or unsuitable text to ensure accuracy and coherence as well as needs to address any errors, and inconsistencies, and eliminate what are commonly referred to as "hallucinations". This contrasts with older transparent systems that demanded more human resources for foundational development and ongoing monitoring. The choice of the "fast track" also led to ethical dilemmas regarding the traceability and accuracy of resources in new systems. Ethical values, such as the source of the resources and veracity of the data must be protected by international and democratic laws and regulations.

INESC-ID has conducted some research to develop tools and resources for paraphrase and translation generation. Our research took a divergent path from mainstream approaches, focusing on experiments centered on transparency and a "human-in-the-loop" methodology. Drawing from our business experience and understanding, we recognized the potential value of language models aligned with the principles of the Logos Model, given their linguistic depth, transparency, and facilitation of human intervention at every stage of the process. Leveraging OpenLogos, we conducted human evaluation tasks on MT systems of various types. The Logos Model and its resources provide a foundation for developing novel systems, regardless of their specific purpose, as long as they involve natural language processing. With this in mind, we used OpenLogos resources to create a paraphrasing system for Portuguese, named eSPERTo, which stands for System for Paraphrasing in Editing and Revision of Text[8]. eSPERTo is extensively documented and positive results have been observed through experiments evaluating its effectiveness in tasks such as question-answering and summarization, among others.

4.1 A Paraphrase generation system: eSPERTo

Figure 1 illustrates eSPERTo results for the paraphrasing of the Portuguese sentence Ontem, o escritor de Marrocos fez uma aposta com a mulher portuguesa (Yesterday, the Moroccan writer made a bet with the Portuguese woman).

4.2 Paraphrastic resources: parafrasário

Akin to, but more extensive than a dictionary, the notion of a "paraphrasary," corresponding to the term "parafrasário" in Portuguese, has been conceptualized as a long-term endeavor as a means to address a gap in the creation of linguistically sophisticated resources for large multilingual generation projects centered on paraphrasing and translation. Experimental results stemming from this project have already been made public[9]. Within the Multi3Generation COST Action framework, the idea materialized in the launching of multilingual

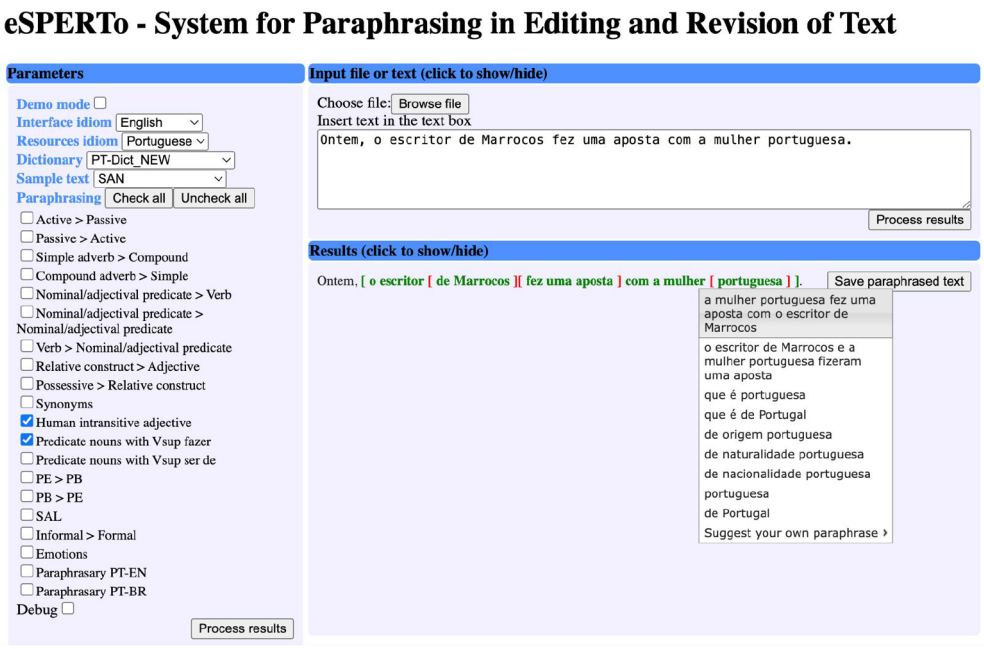

Figure 1. eSPERTo paraphrasing coordination of human-type subjects with a transitive pronominal verb expressing reciprocity: apostar/fazer uma aposta (bet) in Portuguese.

'paraphrasaries' for tailored use in monolingual or multilingual applications. In MT, achieving high-quality translation relies on robust paraphrasing knowledge and skills. Paraphrases play a pivotal role in incorporating semantic knowledge to ensure precise translation rather than mere approximations. There is a need to revisit alignment tools and methodologies collaboratively, establishing guidelines for aligning paraphrases and translations. This entails assessing various degrees of equivalence and broadening the process to encompass more creatively composed texts beyond commonly used corpora. With large paraphrasaries, NLG will improve considerably.

5 eSPERTo *versus* ChatGPT

In this Section, we introduce a straightforward test to elucidate the features of a generative system such as eSPERTo with ChatGPT, aiming to delineate the disparity between the generative process and the resultant output of each system. Example 1 presents a list of generated paraphrases for the Portuguese sentence illustrated in Figure 1. Additional paraphrasing capabilities are available for this sentence. However, for simplicity, we'll present only the most evident ones here.

(1) *PT* - Ontem, o escritor de Marrocos fez uma aposta com a mulher portuguesa.

EN - Yesterday, the Moroccan writer made a bet with the Portuguese woman.

PARA - Ontem, a mulher portuguesa fez uma aposta com o escritor de Marrocos.

EN - Yesterday, the Portuguese woman made a bet with the writer from Morocco.

PARA - Ontem, o escritor de Marrocos e a mulher portuguesa fizeram uma aposta.

EN - Yesterday, the writer from Morocco and the Portuguese woman made a bet.

PARA - Ontem, o escritor de Marrocos fez uma aposta com a mulher que é portuguesa.

EN - Yesterday, the Moroccan writer made a bet with the woman who is Portuguese.

PARA - Ontem, o escritor de Marrocos fez uma aposta com a mulher que é de Portugal.

EN - Yesterday, the Moroccan writer made a bet with the woman who is from Portugal.

PARA - Ontem, o escritor de Marrocos fez uma aposta com a mulher de origem portuguesa.

EN - Yesterday, the writer from Morocco made a bet with the woman of Portuguese origin.

PARA - Ontem, o escritor de Marrocos fez uma aposta com a mulher de naturalidade portuguesa.

EN - Yesterday, the writer from Morocco made a bet with the woman of Portuguese naturality.

PARA - Ontem, o escritor de Marrocos fez uma aposta com a mulher de nacionalidade portuguesa.

EN - Yesterday, the writer from Morocco made a bet with the woman of Portuguese nationality.

PARA - Ontem, o escritor que é de Marrocos fez uma aposta com a mulher portuguesa.

EN - Yesterday, the writer who is from Morocco made a bet with the Portuguese woman.

PARA - Ontem, o escritor marroquino fez uma aposta com a mulher portuguesa.

EN - Yesterday, the Moroccan writer made a bet with the Portuguese woman.

PARA - Ontem, o escritor que é de origem marroquina fez uma aposta com a mulher portuguesa.

EN - Yesterday, the writer who is from Moroccan origin made a bet with the Portuguese woman.

PARA - Ontem, o escritor que é de naturalidade marroquina fez uma aposta com a mulher portuguesa.

EN - Yesterday, the writer who is of Moroccan naturality made a bet with the Portuguese woman.

PARA - Ontem, o escritor que é de nacionalidade marroquina fez uma aposta com a mulher portuguesa.

EN - Yesterday, the writer who is of Moroccan nationality made a bet with the Portuguese woman.

PARA - Ontem, o escritor de Marrocos fez uma aposta com a mulher de Portugal.

EN - Yesterday, the writer from Morocco made a bet with the woman from Portugal.

PARA - Ontem, o escritor de Marrocos apostou com a mulher portuguesa.

EN - Yesterday, the writer from Moroccan naturality made a bet with the Portuguese woman.

The paraphrases just illustrated show that paraphrastic units changed to all parts of the sentence except for the adverb ontem (yesterday). The paraphrastic equivalents for the subject o escritor de Marrocos (the writer from Morocco) were o escritor que é de Marrocos (the writer who is from Morocco), o escritor marroquino (the Moroccan writer), o escritor que é de origem marroquina (the writer who is from Moroccan origin), o escritor que é de naturalidade marroquina (the writer who is from Moroccan naturality), o escritor que é de nacionalidade marroquina (the writer who is from Moroccan nationality). Similar paraphrastic units were generated for the complement a mulher portuguesa (the Portuguese woman). The support verb construction fez uma aposta (made a bet) presents its equivalent single verb apostou (bet) as its semantic equivalent. Additionally, eSPERTo also paraphrased the sentence by swapping the positions of the subject o escritor de Marrocos (the writer from Morocco) and the complement a mulher portuguesa (the Portuguese woman) because the verb, in this case the support verb construction fez uma aposta (made a bet), is a transitive pronominal verb expressing reciprocity of the type [A verb B = B verb A], i.e., if the Moroccan N made a bet with the Portuguese N, then the Portuguese N also made a bet with the Moroccan N. Due to the nature of the verb, which has permutable properties, then it is possible to combine the subject with the complement producing a coordinated subject o escritor de Marrocos e a mulher portuguesa (the writer from Morocco and the Portuguese woman) that requires the verb to be plural fizeram uma aposta (made a bet). The linguistic processes involved in transforming the original sentence are transparent, with clearly delineated boundaries indicating where transformations can occur. This shift represents visible paradigmatic changes. The methodology is consistent and human-validated.

To give readers an idea of how the eSPERTo system compares with ChatGPT, we use ChatGPT version 3.5 to paraphrase the sentence in Example (1) into both Portuguese and English. The Portuguese paraphrases are presented in Example (2), and the English paraphrases in Example (3).

(2) *PT* - Ontem, o escritor de Marrocos fez uma aposta com a mulher portuguesa.

PARA - No dia anterior, o escritor marroquino fez uma aposta com a mulher de Portugal.

PARA - Na véspera, o escritor proveniente de Marrocos fez uma aposta com a mulher originária de Portugal.

PARA - No dia anterior, o autor marroquino fez uma aposta com a portuguesa.

PARA - Ontem, o autor do Marrocos fez uma aposta com a portuguesa.

PARA - Na data passada, o escritor marroquino fez uma aposta com a de nacionalidade portuguesa.

For Portuguese, ChatGPT presented different forms of the adverb ontem (yesterday), such as no dia anterior (in the previous day), na véspera (on the eve), and na data passada (in the past date). This shows lexical variety. For the subject, there was alternation between o escritor (the writer) and o autor (the author), which are synonyms in this context. There is also another paraphrastic unit not shown in eSPERTo, which is proveniente de Marrocos (originating from Morocco). The verb takes always the same form, that of a support verb construction, fez uma aposta com (made a bet with). The complement a mulher portuguesa (the Portuguese woman) is paraphrased as a mulher originária de Portugal (the woman originating from Portugal). Critical is the replacement of mulher (woman) by esposa (wife), which in the context of the sentence does not seem to be correct. Only further context would allow us to ensure that this option would be correct.

(3) *EN* - Yesterday, the Moroccan writer made a bet with the Portuguese woman.

PARA - Yesterday, the author from Morocco placed a wager with the Portuguese woman.

PARA - Yesterday, the Moroccan writer engaged in a bet with the Portuguese lady.

PARA - Yesterday, the Moroccan writer made a bet with the Portuguese.

PARA - Yesterday, the writer from Morocco made a bet with the Portuguese woman.

PARA - Yesterday, the Moroccan writer made a bet with the Portuguese woman.

The English paraphrastic options are identical to those in Portuguese, except for the adverb which remains always

the same word, 'yesterday'. The subject 'the Moroccan writer' alternates with 'the writer from Morocco' and 'the author from Morocco'. The verb alternates between the support verb constructions 'made a bet with', 'placed a wager with', and 'engaged in a bet with'. Lastly, the complement 'the Portuguese woman' alternates with 'the Portuguese lady', and dangerously with 'the Portuguese wife'. Both the Portuguese and English paraphrases generated by ChatGPT were less accurate than those produced by eSPERTo for Portuguese. ChatGPT tended to offer riskier paraphrases when substituting words. Ultimately, eSPERTo demonstrated a higher level of linguistic sophistication compared to ChatGPT. For example, it generated coordinated subjects from a subject and verb complement, allowing for permutations, and correctly, and conjugated verbs in the plural form.

6 Conclusions and future work

Recent advances in LLMs have significantly improved text generation capabilities, sometimes surpassing human abilities. However, these advancements also bring forth scientific challenges, particularly in areas like language understanding, among others. This paper primarily focuses on tracing the evolution of NLG within the context of MT. We exemplify this journey through a commercial partner's translation process, highlighting the transition from traditional methods to the adoption of the Logos MT system to streamline translation needs. Understanding the historical significance of MT is crucial for comprehending our current state in NLG and LLMs and our position within the realm of GenAI.

Expanding on this historical context, we introduce a paraphrase generation system, eSPERTo, inspired by the Logos Model and using OpenLogos resources. eSPERTo operates as a transparent, glass-box system rooted in linguistic knowledge, catering specifically to Portuguese paraphrasing. Comparing eSPERTo's output with that of ChatGPT, we find eSPERTo to be linguistically more robust, transparent, and traceable, whereas the model used by ChatGPT remains unclear due to its black-box nature.

Distinct characteristics emerge between GenAI models: linguistically-based, transparent (glass-box) systems with human involvement at every stage, ensuring greater control and refined outputs, versus opaque (black-box) models where human intervention and design clarity are uncertain. The latter represents the current trend, favoring post-generation error correction over proactive control measures akin to glass box systems. While platforms such as ChatGPT offer impressive results, it is crucial to exercise caution in their usage and recognize that these tools inherently carry biases reflective of their creators. Conversely, 'old-school' tools like OpenLogos (in MT) demand more effort and pose usability challenges but offer transparency, complete customization, and reproducible outcomes when users invest the time to comprehend them. Ultimately, human collective knowledge and wisdom should not be surpassed by inferior entities like machines as the Creator should never be replaced by the creation.

Data availability

No data are associated with this open letter.

References

1. Barreiro A, Lloret E, Turuta O: **An outlook on Natural Language Generation**. In: Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, and Dominik Ślęzak, editors, *Position Papers of the 18th Conference on Computer Science and Intelligence Systems.* Annals of Computer Science and Information Systems, PTI, 2023; **36**: 27–34.
Publisher Full Text
2. Barreiro A, Scott B, Kasper W, *et al.*: **OpenLogos machine translation: philosophy, model, resources and customization.** *Mech Transl.* 2011; **25**(2): 107–126.
Publisher Full Text
3. Barreiro A, de Souza JGC, Gatt A, *et al.*: **Multi3generation: multitask, multilingual, multimodal language generation**. In: Helena Moniz, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussá, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. EAMT 2022, Ghent, Belgium, European Association for Machine Translation, June 1–3, 2022, 2022; 345–346.
Reference Source
4. Barreiro A, Gerazov B, Lloret E: **A linguistic approach to Natural Language Generation**.
5. Scott B: **The logos model: an historical perspective.** *Mech Transl.* 2003; **18**(1): 1–72.
Publisher Full Text
6. Scott B: **Translation, brains and the computer: a neurolinguistic solution to ambiguity and complexity in Machine Translation.** *Machine Translation: Technologies and Applications.* Springer International Publishing, 2018.
Publisher Full Text
7. Scott B, Barreiro A: **OpenLogos MT and the SAL representation language**. In: *Proceedings of theFirst International Workshop on Free/Open-Source Rule-Based Machine Translation.* Alacant, Spain, November 2–3, 2009; 19–26.
Reference Source
8. 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.* Lecture Notes in Computer Science, Springer, 2011; **6609**: 365–376.
Publisher Full Text
9. Barreiro CMA: **A Multilingual paraphrasary of multiwords**. In: *Proceedings of Workshop on Multilingual, Multimodal and Multitask Language Generation, co-located with The 24th Annual Conference of The European Association for Machine Translation*. Tampere, Finland, June 2023; 47–56.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ?

Version 1

Reviewer Report 21 August 2024

https://doi.org/10.21956/openreseurope.18223.r42455

© **2024 Alcina A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Amparo Alcina
Universitat Jaume I, Castellón de la Plana, Valencian Community, Spain

Overall, the text highlights the historical importance that user interaction has had in improving machine translation systems, and the benefit that this methodology can have in improving language generation systems. This idea is, in itself, interesting. However, a greater development of this perspective is lacking throughout the different sections. The sections appear poorly related or cohesive with each other.
Section 2 draws attention to the benefits of using NLG in industry, and to the ethical aspects that must be taken into account in the use and application of NLG, such as biases, information control.
Section 3, 'Logos relevance to industry language models', presents the different tools (machine translation, terminology management, translation memory, content management, etc.) and strategies (controlled language) that are part of or have been used in OpenLogos. In Section 4, the eSPERTo language generation system is presented, which is distinguished from others, such as ChatGPT, because it uses user interaction to choose the best text options during the generation process, rather than the user having to adapt or modify the final result.
Some parts of the text are of great interest, such as the presentation of the eSPERTo system and its functionalities (Section 4). However, one would expect it to go into more detail about the characteristics of the presented system, and for the introductory sections to highlight those features that have in one way or another been related to this development. For example, since the problem of biases in language generation has been mentioned in Section 2, one would expect Section 4 to explain how this eSPERTo system addresses this problem or can improve the final result.
The description of the Logos system and its relevance (Section 3) is generic. The characteristics it describes do not confirm its 'relevance' to other commercial systems. In this sense, it would have been interesting, for example, to frame the system within the typology and methodologies of machine translation systems, which would allow its relevance to be focused. There is a lack of bibliography or documentation describing the Logos system or sources that have been consulted for the preparation of the article. It would also be expected to name other systems that have also used these strategies, such as, in general, those called interactive machine translation, although it focuses on one of them.
The evaluation of the eSPERTo system and its comparison with ChatGPT is based on a few

sentences. The analysis does not provide definitive conclusions.
In addition to providing bibliographical references from the authors themselves, other references used in the preparation of the research and the text should be provided.

Is the background of the case's history and progression described in sufficient detail?
Partly

Is the work clearly and accurately presented and does it cite the current literature?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Is the case presented with sufficient detail to be useful for teaching or other practitioners?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Translation technologies, Language Technologies, Terminology.

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, however I have significant reservations, as outlined above.

Reviewer Report 12 August 2024

https://doi.org/10.21956/openreseurope.18223.r42454

© 2024 Allen J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jeffrey Allen
SAP, Levallois-Perret, France

First of all, this article is very well-written, in English by non-native speakers, and only contains such minor points that would only represent stylistic preferences which are not important to even mention.

This article provides one of the most comprehensive explanations of the implementation done at

Océ that I have read over the past 20+ years. Lou provides more info on the benefits which helps support the claims to success which Océ and other organizations have had with Controlled Language, MT, TMs, a TMS and other related technologies and workflows. He also provided a well-explained summary of the factors involved to move away from a home-grown solution to licensed 3rd party solutions with customization.

This article also provides a follow-up on using Logos, which underwent a transition to become OpenLogos, yet I have never seen since that time any other academic or industrial research team take on the task to use the open-source version of the product. This shows how legacy products are still very useful today for current needs.

Another key point which I particularly liked was the emphasis on how the older, dictionary & rule-based generation of these technologies already contained various human-in-the-loop steps which was successful and beneficial. And he indicated how the move to statistical, neural and GenAI generations lost that key element, due to the focus on handling Big Data, which has thus turned into the quest for bringing the human and the mind back into the procedure. Many of the large multinational corporations who pursued the same path in parallel, (Caterpillar, GM, Xerox, SAS, SAP, various telecom companies, Symantec, etc) all made similar claims of success, yet many of them also closed down the projects for similar reasons. The refreshing aspect of this story is that both authors showed how it is possible to bridge the rule-based technology generation to the current GenAI generation by showing how linguistic analysis has been lacking (since 2010) and can be brought back again into equation via their proposal to use eSPERTo.

Another important factor is that the project has been conducted and published within the framework of a European funding agency for research and innovation networks. This ensures that the results are available and can be reproduced by others, rather than staying within the closed walls of corporate and industrial intellectual property.
I have only 2 minor comments on content.

1. The second paragraph of chapter 4 starts with the acronym INESC-ID which is not defined in a spelled out or paraphrase of the meaning of the organization/project.
2. The introductory paragraph states “the core of human nature and its ability to generate and understand language, the trait that distinguishes humans from other species”. I tried using the same statement in replying to a museum guide at the Exposition on Species in May 2022 at the Confluences Museum in Lyon, France. The correct answer, according to the museum guide, is that humans are the only species to destroy their own habitat and environment. Maybe a slight change in phrasing could be to say that the notion of humans being the only species to generate and understand language is a “commonly made claim”.

Is the background of the case’s history and progression described in sufficient detail?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Is the case presented with sufficient detail to be useful for teaching or other practitioners?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational Linguistics, Natural Language Processing, Controlled Language, Translation technology and workflows.

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.