Check for updates

METHOD ARTICLE

A linguistic approach to Natural Language Generation: The language spectrum, language models, and OpenLogos

[version 1; peer review: 1 approved with reservations, 1 not approved]

Anabela Barreiro [1], Elena Lloret [2], Branislav Gerazov [3]

[1]OpenLogos and INESC-ID, Instituto de Engenharia de Sistemas e Computadores, Lisboa, Lisboa, 1000-029, Portugal
[2]Department of Software and Computing Systems, University of Alicante, Alicante, Valencian Community, Spain
[3]FEEIT, Ss Cyril and Methodius University in Skopje, Skopje, North Macedonia

v1 **First published:** 24 Jun 2024, **4**:123
https://doi.org/10.12688/openreseurope.17349.1
Latest published: 24 Jun 2024, **4**:123
https://doi.org/10.12688/openreseurope.17349.1

Abstract

Natural Language Generation, a well-established research area within the field of Natural Language Processing, has achieved significant milestones such as in machine translation and paraphrasing. Large Language Models, while improving text quality, show varying effectiveness across different linguistic registers and cultural contexts, raising concerns mostly due to (i) difficulties in understanding the internal workings of the models; (ii) difficulties in understanding the generative processes, which are not transparent; (iii) difficulties in human intervention in the generative function; (iv) propensity for hallucination and incorrect information content, (v) potential irresponsible use of extensive resources from unidentified sources, and (vi) the risk of misuse, among the most obvious hurdles. This article has three main goals: (1) suggest a linguistic approach to Natural Language Generation, (2) explore the broad language spectrum, including formal to informal styles and objective to subjective language, among others, and (3) propose language models inspired by the Logos Model for enhanced transparency, traceability, and customization. The Logos Model is at the core of the best-documented pioneer commercial machine translation system called Logos, which preludes the history of Generative Artificial Intelligence. The open-source version, OpenLogos, offers valuable resources for machine translation, paraphrasing, abstractive summarization, and various other tasks related to Natural Language Generation, and it serves as an excellent resource for training researchers across multiple fields associated with Artificial Intelligence. The OpenLogos resources were made publicly available through the Multi3Generation COST Action (CA18231).

Plain language summary

Open Peer Review

Approval Status ✗ ?

	1	2
version 1 24 Jun 2024	✗ view	? view

1. **Hugo Gonçalo Oliveira**, University of Coimbra, Coimbra, Portugal

2. **Ehud Reiter**, University of Aberdeen, Aberdeen, UK

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

Natural Language Generation is an established area of study with a rich history of verified accomplishments dating back to the groundbreaking Machine Translation, which stands as the earliest and foremost intricate language application to enter the technology market. Recognizing the historical significance of Machine Translation in the context of Natural Language Generation is essential for comprehending where we stand today with regards to Generative Large Language Models and our current position on Generative Artificial Intelligence. Large Language Models such as Chat-GPT or Bard, while offering free usage (with paid versions available), play a significant role in improving text quality. However, these models' effectiveness varies significantly across different linguistic registers and cultural contexts and they have also sparked widespread concerns, especially in fields like education and various industries such as healthcare and machinery. In these sectors, the potential for language errors to result in life-threatening consequences is a crucial concern, given their substantial impact on human life integrity. In this article, we highlight the generative attributes of the Logos Model, which is the oldest and most extensively documented commercial Machine Translation model, to illustrate and trace the origins of Natural Language Generation. We showcase some resources from the English-Portuguese and the English-Spanish language pairs. These resources encompass the fundamental components of Natural Language Generation, specifically, the translation of the most elementary lexical items (words) and the second most basic lexical items (multiwords). The article proposes a linguistic-based approach to Natural Language Generation, introducing a broad spectrum of language features, including variations from formal to informal styles, the choice between objective and subjective language, distinctions between technical and non-technical terminology, and the range from clear to ambiguous ways of expression, encompassing various subjects and genres. Finally, it suggests alternative or supplementary models influenced by the Logos Model, aiming to enhance the transparency and traceability of Natural Language Generation systems in applications such as Machine Translation, Paraphrasing, or Summarization. The OpenLogos resources were made publicly available through the Multi3Generation COST Action (CA18231).

Keywords
Computational Linguistics, Natural Language Generation, Generative Artificial Intelligence, Language Models, Language Spectrum, OpenLogos, Multi3Generation COST Action

This article is included in the COST Actions gateway.

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

Author roles: Barreiro A: Conceptualization, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Lloret E**: Funding Acquisition, Investigation, Project Administration, Supervision, Writing – Review & Editing; **Gerazov B**: Conceptualization, Data Curation, Resources, Software, 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 grant agreement No. CA18231 (Multi3Generation: Multi-task, Multilingual, Multi-modal Language Generation), as supported by the COST Association (European Cooperation in Science and Technology). Also, Elena Lloret acknowledges project "PID2021-123956OB-I00", funded by MCIN/ AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; project CLEARTEXT "TED2021-130707B-I00", funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/ PRTR; and project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/2021/021), funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Barreiro A *et al*. 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, Lloret E and Gerazov B. **A linguistic approach to Natural Language Generation: The language spectrum, language models, and OpenLogos [version 1; peer review: 1 approved with reservations, 1 not approved]** Open Research Europe 2024, **4**:123 https://doi.org/10.12688/openreseurope.17349.1

First published: 24 Jun 2024, **4**:123 https://doi.org/10.12688/openreseurope.17349.1

1 Introduction

The Multi3Generation COST Action (CA18231)[1] [1], referred to as Multi3Generation hereafter, encourages collaborative endeavors to amplify the significance of the Natural Language Generation (NLG) research domain for the benefit of society. 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. LLMs have significantly advanced text generation, but there are certain drawbacks associated with these models, which pose substantial challenges when used improperly and without sufficient knowledge and training. They also present several types of multifaceted challenges that need to be addressed scientifically. The first challenge involves difficulties in comprehending the models themselves, exacerbated by their opacity that makes understanding the generative processes a complex task. Human intervention within the generative function is also limited, leading to outcomes in bots that exhibit a proclivity for hallucination and the inclusion of misleading information. Additionally, adapting to diverse styles and domains poses challenges. Furthermore, there is apprehension about the potential for irresponsible use of significant resources from undisclosed sources, compounding other concerns related to the risk of misuse, which could have a more significant impact within the domains where confidence in data and data processes is crucial. Conversely, traditional machine translation (MT) models prioritize the use of meticulously curated resources. They feature a clearly defined design that is accessible to and understood by non-experts, and their processes are characterized by transparency, adaptability, traceability, and human manageability.

This article emphasizes the generative characteristics of the oldest and most extensively documented commercial MT model, the Logos Model[2–4].[2] The aim is to illustrate and trace the origins of NLG. We can state that Logos forms the cornerstone for narrating the history of commercial Generative Artificial Intelligence (GenAI). The article discusses the robust generation component of the Logos system in the 1990s[5], showcasing pioneering language technology. The illustration of this is done using resources from the English-Portuguese and English-Spanish language pairs. Our proposal suggests developing alternative and/or complementary models inspired by the Logos Model, its philosophy, design, tools and resources. 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, there are several different levels of user engagement, which constitute a smaller or greater challenge to NLG. This is primarily due to the inherent flexibility and extensive nature of language. The article advocates for a linguistic-based approach to NLG, covering various language features. It concludes by recommending alternative or supplementary models to enhance transparency and traceability in NLG systems, particularly in applications like MT, Paraphrasing or Abstractive Summarization. The proposed models are firmly rooted in the field of linguistics, the science of language, which is a frequently neglected aspect in many existing models. Linguistic models are defined by their incorporation of strategies that involve human participation (human-in-the-loop), allowing human intervention on any phase of the NLG process.

The rest of the paper is organized in the following way: Section 2 presents a comprehensive review of the present state of the art in NLG. Section 3 outlines the natural language spectrum and the need to scrutinize the purposes, objectives, and goals of each text type to adapt the computational system to generate customized language. In this Section, we provide a context for the broad array of language uses, variations in styles, objectives, and so on (referred to as 'the language spectrum') and how this diversity gives rise to significant hurdles in the field of NLP, particularly in NLG and Natural Language Understanding (NLU), challenges that were recognized when developing the commercial Logos system decades ago[2]. Our assessment is only superficial and has different purposes and goals but the same objective of guaranteeing control over information and knowledge. Section 4 highlights the benefits of employing a linguistically-oriented strategy for MT in particular and NLG in general, with a focus on using resources that incorporate knowledge. It presents different methodologies being developed throughout the years driven by a strong linguistic contribution and theoretical background (or combination of backgrounds) that can be further explored to increase efficiency, transparency, and public confidence in NLG technology. Section 5 is subdivided into two subsections, each illustrating target language components. Subsection 5.1 explains SAL, using examples in English-Portuguese. Subsection 5.2 describes SemTab, providing further clarification through examples in English-Spanish. In Section 6, we will discuss what is missing in most NLG models and how the integration of linguistic knowledge can help GenAI systems. Section 7 presents our vision for what the future path could entail. Finally, Section 8 offers several conclusions.

2 Background in Natural Language Generation

From a historical perspective, research into NLG has experienced several trends. The work of Reiter and Dale, 1997[13] was pioneering in proposing an NLG pipeline to develop applied NLG systems. This NLG pipeline is distinguished between the following stages[14]: macroplanning, microplanning, and surface realization. To address each of these general stages is not trivial and they may be subdivided into other subtasks, such as document planning and structuring (for macroplanning); sentence aggregation, lexicalization, or referring expression generation (for microplanning), or linguistic realization and structure realization (for surface realization).

From the 90s until approximately 2016, this was the predominant trend, and research groups addressing NLG focused on one of these tasks (e.g. macroplanning or surface realization). The types of approach was either rule-based, where SimpleNLG was one of the reference available systems[15]. SimpleNLG was originally developed for English; however, due to its

[1] https://multi3generation.eu/

[2] Complementary references are: 2,6–12

simplicity and adaptability, versions for other languages, such as German[16], Mandarin[17], Spanish[18], or Galician[19] were also proposed. In parallel, NLG applications, such as machine translation[20], abstractive summarization[21], or paraphrasing[22] were being researched independently, proposing specific methods and approaches (either rule-based, machine-based, or hybrid) without taking profit from the developments achieved by the "pure" NLG community. During this period, different initiatives, such as SIGGEN[3], the Special Interest Group on Natural Language Generation (created in 1993), as well as NLG workshops and conferences emerged. After a series of NLG workshops organized in the 90s, the First International Conference on Natural Language Generation (INLG) appeared in 2000, and it has been going on until nowadays. Around 2015, with the promising results of Deep Learning (DL) techniques in other NLP tasks, something started to change in NLG. However, whereas for Natural Language Understanding (NLU), DL advancements were done at a fast pace, these techniques were not so easy to apply for NLG, most of the novel NLG methods still relied on the macroplanning, microplanning, and surface realization pipeline[23]. In particular, the work of Barros *et al.*[24] show that a hybrid surface realization approach combining Factorized Language Models with specific semantic information encoded using rules led to competitive results in the state of the art. Moreover, its flexibility to be plugged into a macroplanning approach that uses Positional Language Models[25] turned into a cost-efficient NLG approach that could be applied to a wide range of tasks, such as summarization[26] or headline stance detection[27]. It was not until the development of Transformers architecture in 2017[28] when a new paradigm for NLG was adopted, that one of neural NLG[29], resulting in a big shift that became more evident after Generative methods and LLMs, such as GPT[4], Llama[30], Bard[5], or Vicuna[31], all of them grouped under the concept of "Generative AI". What makes it a different paradigm is the way NLG tasks are tackled. Whereas until then, each task was independently addressed, nowadays with generative architectures, different tasks can be approached using a similar methodology (taking as a basis a pre-trained model and then fine-tuning it with a specific dataset). This has revolutionized the way NLG is perceived and addressed – now talking about integrated architectures[32], where no distinction between macroplanning, microplanning, and surface realization is done anymore – and gaining popularity not only within the research community but also in society and industry.

Focusing specifically on the task of Machine Translation, traditional MT methods, which rely on rule-based and statistical approaches, continue to hold relevance in the era of DL. While neural MT models have made significant advancements in recent years, traditional MT systems provide valuable resources, especially when dealing with less-resourced languages, low-resource domains, and precise domain-specific translations. They offer control and transparency, making them indispensable in certain contexts. One of these systems is OpenLogos, a free software version of the Logos MT system that was used in production for over thirty years. It represents a vital resource for the MT community, particularly for Early Career Researchers (ECRs) seeking to explore the foundations of MT and contribute to its evolution. OpenLogos provides both a learning platform and a framework for the development of rule-based translation systems. ECRs can explore, modify, and expand OpenLogos to address language pairs and domains that may not be well-covered by neural MT models. The OpenLogos source code, as well as the OpenLogos tools and resources have been made available on the internet[6–8]. The platform facilitates both open access, as well as their collaborative development by the research community. Currently, the platform hosts the English to Portuguese dictionary, which can serve as a basis for the development of future low-resourced language dictionaries. Semantico-Syntactic (SemTab) rules for translating English to Spanish are also made available. Finally, the platform hosts a webpage that describes the OpenLogos project and hosts the Semantico-Syntactic Abstraction Language (SAL) tutorial. The webpage offers a user-friendly resource for newcomers to the OpenLogos community. Besides the useful resources it offers the possibility of creating translation and paraphrasing systems from scratch, OpenLogos can function as an educational tool where individuals can learn the pipeline to process incrementally from the NLU of words and phrases in the source language to the NLG of their corresponding translations in the target language. By learning the process, which is transparent and where human intervention can take place at any stage of the translation process, individuals can also explore the evolution of machine translation, spanning from its early systems to the contemporary landscape.

3 The Language Spectrum

Natural language is the primary and richest tool for expression and communication. It is what distinguishes humans from other animals. It is integral to everyday life, either expressed in speaking, writing, and reading. Machines can process relatively well the natural language of the type used in contracts and legal documents, scientific, technical, and academic texts, newspapers, letters, and essays, but machines perform poorly when language is stylistically rich, contains figures of speech, metaphors, idioms, etc. resulting from a creative activity or process, which is incomparable between humans and machines.[9] Creative texts produced by humans, such as fiction,

[3] https://aclweb.org/aclwiki/SIGGEN

[4] https://chat.openai.com/auth/login

[5] https://bard.google.com/

[6] https://sourceforge.net/projects/openlogos/

[7] https://www.hlt.inesc-id.pt/openlogos/demo.htm

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

[9] In our understanding of the philosophical question: 'What is creativity?', and when comparing human-machine competencies, words such as 'cognition', 'learning', and alike do not represent the same. In a conventional sense, to describe the capabilities of a machine, using words that involve purely human processes, such as *thought*, *heart*, *soul*, *emotion*, *sentiment*, *intention*, *willingness*, *desire*, *conscience*, *inspiration*, *sensibility*, *sensitivity*, *enlightenment*, *judgment*, and other processes of a psychological, moral, or ethical nature that cannot be forcibly injected into a non-sentient or non-moral/ethical machine, may not be the best way to move forward. The essence of the human and machine 'beings' is different and this nature cannot be altered now or ever. Metaphorically speaking what is blood and flesh cannot ever be transformed into hardware or the other way around. Similarly, the immaterial properties of a human being can never be transformed into software or the other way around.

plays, humor, and poetry, among others, are characterized by singularity, uniqueness, and originality, protected by Intellectual Property Rights (IPR) in international laws and treaties regarding author's rights. Figure 1[2] illustrates the wide spectrum of language, from very objective texts to subjective texts that are the fruit of human creativity.

The image reflects language characteristics and properties, used for different purposes and objectives, covering distinct activities, and applied in multiple applications. Language can be used to express different types of content and registry (style), which can be more or less formal and more or less clear. Normally, the level of clarity is bigger in formal language. Associated with language production, there is a key aspect related to the communicative goal of the message to be produced. This is very important for the generation process[33,34], but it has not been paid the deserved attention to be integrated within generation. In this respect, and speaking very broadly, language can be used either to inform or to cause feelings and emotions. Informative texts, such as contracts and other legal documents, academic, scientific and technical texts tend to be formal and use specific terminologies that are understood by the specialists in each field of knowledge. Moving on the spectrum, newspapers are informative but can be more or less formal and clear, and express different types of content in different styles, the style of the journalist communicating the news or conveying the message. The language should be transparent and the message reliable, even if there is a common feeling that transparency, quality, and reliability are downsizing lately. Essays can be considered to be in the frontier between objective and subjective language and both content and style may start merging between formal and informal and the style between conventional and idiosyncratic. As the natural language is used in fiction, plays, humor, and poetry, the clarity may start reducing as there is more subjectivity and more or less informality and idiosyncrasy using figures of speech, word playing, intentional ambiguity, and fluidity. There is an increasingly broader range of stylistic awareness and variety in the use of words and forms of expression, which affects the performance of NLP systems.

4 Linguistics in Natural Language Generation

From a theoretical linguistics perspective, words can be considered the basic element of language. A word has a surface representation formed by the characters it is composed of, but it also has a meaning assigned by its use in context. The dissociation between the representation of a word and its meaning is taken for granted by humans, but it is not so trivial for machines, which have to learn words based on raw data. LLMs are mostly taking into account the representation of the word at a surface level, rather than its meaning, even if it appears that they understand the language that they are generating. The apparent power of inference and deduction is a result of human feedback, i.e., a result of direct or indirect input

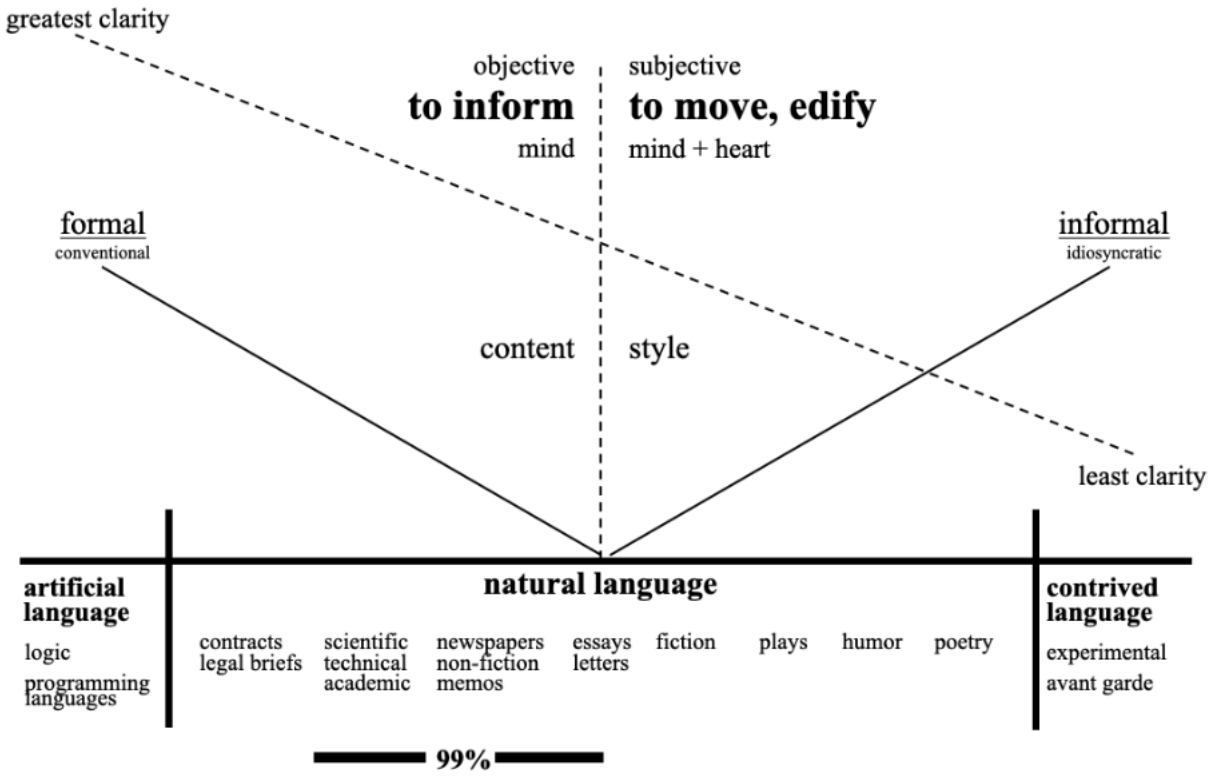


Figure 1. The Language Spectrum. (This figure has been reproduced with permission from Scott, 2003[2]).

by the model's users and trainers. Currently, one of the strategies to correct the problems of these systems is to have human trainers teach, guide, or enforce machine learning (ML). We challenge this method of human intervention with the development of more resources and tools based on linguistic knowledge because they are essential for the development of ethical AI and contribute to human control over computational systems and not the other way around. Even if LLMs have proven to allow users to perform various tasks or progress faster in other tasks, the field of AI needs linguistic expertise because language professionals are the ones who can legitimately and rightly assess, provide feedback, and express their positions concerning the linguistic quality assurance of NLP/NLG systems achieved via AI. Similarly, one can say that AI needs support and expertise from fields such as neuroscience, which is also implied on critical problems entailed in learning and knowledge acquisition by the human brain, the examination of cognitive functions, and conscience, among others.

Our perspective on NLG is that future technology should not neglect the deep study of languages and their efficient computational modeling. This approach values the development of knowledge-based linguistic resources by researchers with extensive experience in computational linguistics, as they are the only ones who can question the intrinsic limitations of language models. We believe that it is worthwhile investing in quality resources and good design of controllable systems rather than investing in post-editing of uncontrolled output[35] (notwithstanding the interest in discovering if/what an uncontrollable machine can do that a human cannot do or understand). Following this line of reasoning, efforts have been initiated to use resources obtained from OpenLogos, specifically the lexical and Semantico-Syntactic Abstraction Language, for the development and enhancement of new linguistic resources and tools. One of these endeavors aimed to design, develop, and facilitate the improvement of the eSPERTo paraphrase generation system[11].[10] Similar efforts can be undertaken for Summarization, Question-Answering, and other NLP systems. Section 5 presents two important features of the Logos Model: the Semantico-Syntactic Abstraction Language (SAL) and the Semantico-Syntactic Tables (SemTab).

5 Use case: the Logos Model

As previously stated, the Logos Model stands out as the most extensively scientifically published MT model. The OpenLogos (Logos Model) method stands as a well-established and widely recognized commercial system with extensive implementations across both research and commercial spheres. OpenLogos has left a mark on the field, playing a pivotal role in shaping the development of systems such as the Anusaaraka MT system[36] and the eSPERTo paraphrasing system[11,37]. These implementations underscore the method's significance and contributions it offers, demonstrating its versatility and applicability across diverse domains and applications. Various articles have delved into its philosophy, model, resources, and customization. A crucial element of ensuring sufficient resources revolves around establishing a well-defined taxonomy and data modeling, coupled with a structured data pipeline[2]. We defend that the absence of such a detailed modeling setup undermines the strategic value of any AI initiative. From here on we will focus on a taxonomy where natural language is represented as a refined Semantico-Syntactic Abstraction Language (SAL), also referred to as a hierarchical ontology. Upon entering the system, natural language undergoes an immediate transformation into SAL strings. The system incorporates a supplementary database of combinatorial elements known as SemTab rules that interact with SAL elements. These rules are numeric representations of semantico-syntactic rules, complementing the dictionary with a database of multiwords and phrasal units.

5.1 Semantico-Syntactic Abstraction Language (SAL)

SAL contains categories for all parts of speech and has four levels of abstraction: one syntactic level (word class), and three levels called superset (higher level), set (middle level), and subset (lower level). In SAL, the parts of speech are categorized into 12 classes, which are further divided into open and closed word classes. Open word classes are nouns, adjectives, adverbs, and verbs. Closed word classes are articles, prepositions, auxiliary verbs, pronouns, interrogatives, negative particles, conjunctions, and punctuation. Table 1 lists the parts of speech in SAL numeric representations (word classes (WC)).

When processing the sentence, word strings (literally) are converted into SAL entities or SAL patterns. We believe that it makes sense to train ML systems to learn new SAL patterns based on paraphrasary entries, instead of on the conventionally-used n-gram alignments. For example, English verbs are organized in a taxonomy of supersets and sets. There are no subsets as the subset codes are used for verb-specific verb IDs. Verbs have eleven supersets. Nine of these supersets represent transitive verbs, three represent intransitive verbs. Figure 2 illustrates the three types of intransitive verbs: existential, operational, and motional. Verb superset and set codes denote the valence or argument structure common to its verb members. In the figure, maroon denotes the verb superset, and red denotes the verb set.

Figure 3 illustrates the SAL superset for motional intransitive verbs (INMO) and its two sets, the motional intransitive that selects the preposition type *into* (INMOinto-type) (or other prepositions denoting directed motion), and the motional intransitive that selects the preposition type *in* (INMOin-type). The sentence *they walked into the room* illustrates the INMOinto-type verb and the sentence *they danced in the streets* illustrates the INMOin-type.[11]

[10] eSPERTo project REF: EXPL/MHC-LIN/2260/2013

[11] Figure 2 and Figure 3 have been reproduced with permission from Bernard Scott, chief technology architect of the Logos Model. This information was obtained from an internal tutorial of the Logos Corporation enterprise, where Anabela Barreiro worked for 7 years.

Table 1. Parts of Speech in SAL.

Part of Speech	Abbreviation	Word Class
Nouns	[N]	[WC 01]
Adjectives	[ADJ]	[WC 04]
Determiners	[DET]	[WC 14 15]
Arithmates	[AR]	[WC 16] (See also WC 04, 14)
Pronouns	[PRO]	[WC 05]
Relatives/Interrogatives	[INTREL]	[WC 18]
Verbs	[VT / VI]	[WC 02]
Auxiliary Verbs	[AUXMO]	[WC 12]
Adverbs	[ADV]	[WC 03 06]
Prepositions	[PREP]	[WC 11 13]
Negatives	[NEG]	[WC 17]
Conjunctions	[CONJ]	[WC 19]
Punctuation	[PUNC]	[WC 20]

Table 2 depicts the OpenLogos dictionary designed for translating between English and Portuguese. The implementation of this dictionary was made in a relational database, specifically Oracle. Although the database contained numerous tables, not all are depicted in the figure for simplicity. The dictionary includes punctuation, signs, single words, and multiwords of several types, including named entities, lexical bundles, phrasal and prepositional verbs, among others.

5.2 Semantico-Syntactic Tables (SemTab)

The Logos Model reckons the incorporation of databases in what is referred to as the Semantic-Syntactic Tables (SemTab). These tables conventionally represent grammatical rules that encompass various types of multiwords, such as complex compounds, lexical bundles, prepositional nouns, adjectives, and verbs, and support verb constructions, among other multiwords, some of which are commonly known as phrasemes or simple phrases[35]. All of these phrases are integrated into the SemTab function. By providing the English to Spanish SemTab rules, we illustrate the different translations of the same verb depending on its context, such as in ACCEPT(VT) N(PROPOSAL) = ACEPTAR N(PROPUESTA) and ACCEPT(VT)

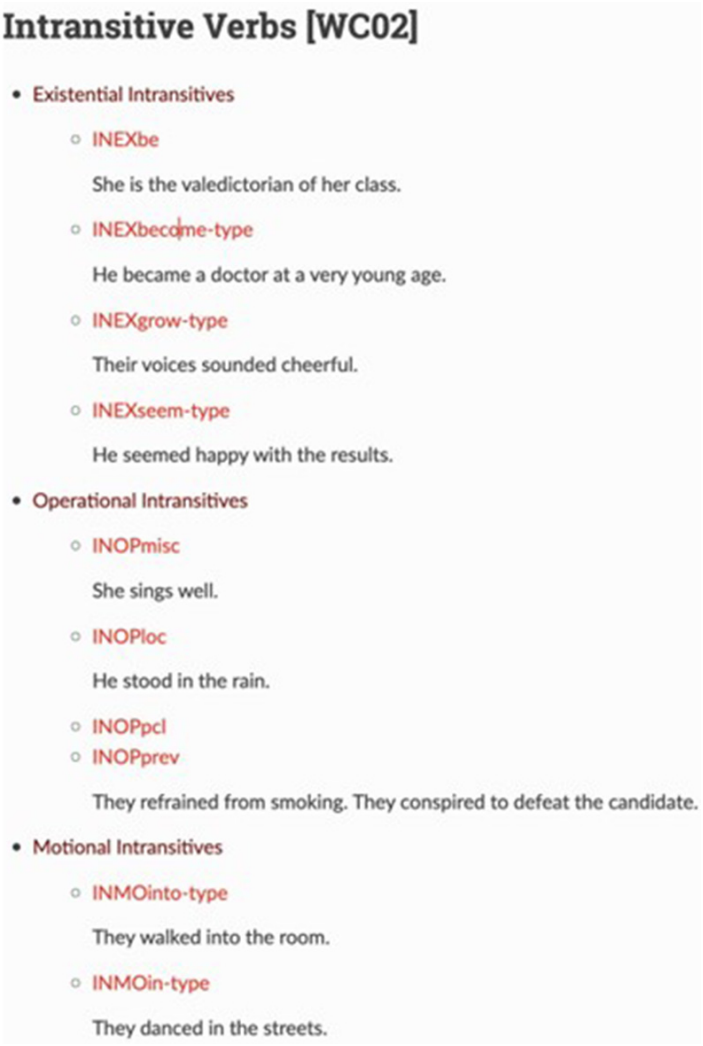
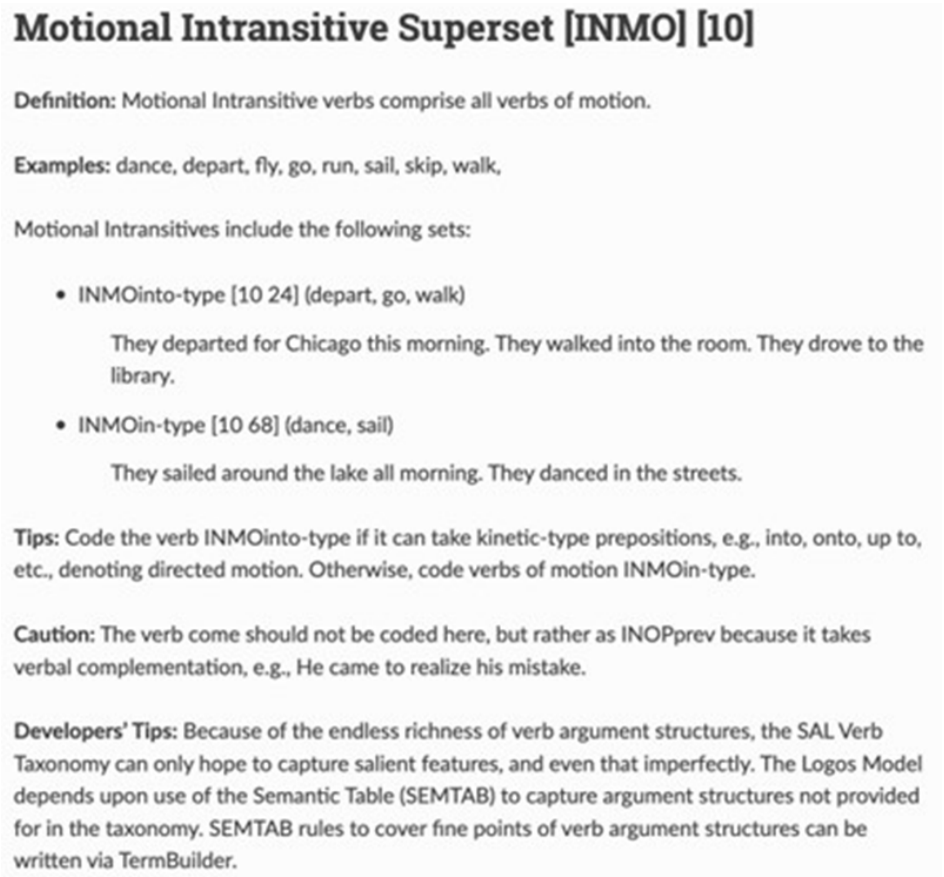


Figure 2. SAL for 3 types of intransitive verbs in English.



Motional Intransitive Superset [INMO] [10]

Definition: Motional Intransitive verbs comprise all verbs of motion.

Examples: dance, depart, fly, go, run, sail, skip, walk,

Motional Intransitives include the following sets:

- INMOinto-type [10 24] (depart, go, walk)

  They departed for Chicago this morning. They walked into the room. They drove to the library.

- INMOin-type [10 68] (dance, sail)

  They sailed around the lake all morning. They danced in the streets.

Tips: Code the verb INMOinto-type if it can take kinetic-type prepositions, e.g., into, onto, up to, etc., denoting directed motion. Otherwise, code verbs of motion INMOin-type.

Caution: The verb come should not be coded here, but rather as INOPprev because it takes verbal complementation, e.g., He came to realize his mistake.

Developers' Tips: Because of the endless richness of verb argument structures, the SAL Verb Taxonomy can only hope to capture salient features, and even that imperfectly. The Logos Model depends upon use of the Semantic Table (SEMTAB) to capture argument structures not provided for in the taxonomy. SEMTAB rules to cover fine points of verb argument structures can be written via TermBuilder.

Figure 3. SAL Motional Intransitive verbs superset and sets INMOinto-type and INMOin-type.

Table 2. English–Portuguese dictionary entries' sample with embedded SAL knowledge. The featured elements are source word (s_word), source word in text (s_wt), superset, set, subset, and target word (t_word).

s_word	s_wt	superset	set	subset	t_word	Mnemonic
Ms.	01	5	91	807	Sra.	ANtitle
Chief Executive Officer / CEO	01	5	91	91	director geral	ANtitle
Board of Trustees	01	5	94	94	Conselho de Administração	ANcoll
Albuquerque	01	9	94	226	Albuquerque	PLcity
Antartic Ocean	01	9	56	229	Oceano Antárctico	PLwater
Christmas tree	01	3	32	384	árvore de natal	COtree
everyday life	01	10	30	30	vida quotidiana	TIundif
address	02	3	32	32	endereçar	PCLexplain-type
evaluate	02	4	98	980	avaliar	OBHUMundif
far and wide	03	11	11	683	por toda a parte	spatial_x_phr
on a one to one basis	03	14	14	683	individualmente	rank_x_x
happy	04	13	88	442	feliz	PChappy-type
on and off	04	13	80	185	de quando em quando	
we	05	05	21	796	nós	personal_First_we/us

s_word	s_wt	superset	set	subset	t_word	Mnemonic
far more	06	20	20	150	muito mais	intensifier_x_x
efficiently	06	18	18	18	eficazmente	manner
on a trial basis	06	18	18	683	à experiência	
on behalf of	11	9	28	367	em nome de	cause_becauseof_onbehalfof
on account of	11	9	28	366	por motivo de	
on	13	2	72	522	em	on
on the right side of	13	5	75	649	à direita de	totheright
billion	14	9	54	605	mil milhões	number_billion_billion
some	15	03	32	151	algum	singularplural_all_some
sixty eight	16	4	23	567	sessenta e oito	number_num1-32_5-99
not	17	1	1	460	não	negative_not_not
not yet	17	1	1	461	ainda não	negative_not_not
who	18	1	1	401	quem	relative_that_who
whoever	18	1	48	403	quem quer que seja	relative_whatever_whoever
on the other hand	19	4	32	848	por outro lado	disjunctive_yet_instead

N(RESPONSIBILITY) = ASUMIR N. Figure 4 illustrates several examples of what is called in Logos jargon as "comment lines" for the SemTab tables. In the depicted examples, the translation of the English verb *appear* into Spanish varies depending on the context, specifically influenced by the verb particles (resembling prepositions) accompanying the verb, and in certain scenarios, the chosen complement noun associated with the prepositional verb. For example, APPEAR(VI) IN N(COURT) and APPEAR(VI) BEFORE N(COURT) are both translated the same way, COMPARECER ANTE N(TRIBUNAL). On the other hand, APPEAR(VI) IN N(PUBLIC) or N(PERSON) or N(ACCOUNT), etc. is translated as APARECER EN N.

Logos linguists painstakingly manually crafted thousands of SemTab rules for various language pairs within the commercial Logos enterprise. Some of these rules were intended for internal usage, while others were tailored to meet the specific requirements of external clients, typically large corporations seeking to expand their technical databases. In the late 1990s, Logos recognized that comprehensively cataloging all the phrases and multiwords in a language was an exceptionally challenging task for linguists. Given the sheer abundance of multiwords in languages, and after gaining a firm grasp of the methodology, it became evident that manually capturing all the nuanced expertise for MT systems was impractical/unfeasible and that statistics would need to be incorporated into the MT process. Consequently, a new era began in the field of MT, where statistics were employed to facilitate the recognition and coverage of multiwords.

6 How can GenAI models benefit from OpenLogos expert knowledge?

AI-based Natural Language Generation (NLG) technologies have the potential to revolutionize how information is presented and consumed across various industries. While these technologies are highly advanced, they still face limitations, particularly in comprehending complex domain knowledge. However, integrating AI-based NLG with expert knowledge accessible through traditional rule-based systems offers several benefits. Traditional rule-based systems, crafted by subject matter experts, boast extensive knowledge bases comprising domain-specific terminology, rules, and concepts. By merging these systems with AI-based NLG, generated content becomes more accurate, relevant, and insightful, i.e., the generated content will exhibit a good understanding of the nuances and complexities inherent to the domain. This combination also enhances contextual understanding, as rule-based systems offer deeper comprehension of specific events or situations. For example, in finance, they aid in identifying market trends and stock prices, and other financial data that are crucial for informed decision-making. Coupling this knowledge with AI-based NLG enhances content understanding of domain nuances and complexities. Moreover, rule-based systems provide a standardized approach to generating the content, as they ensure output consistency, crucial in industries like legal or finance where accuracy and compliance are vital. By combining this standardization with AI-based NLG, generated content maintains both accuracy and insightfulness. Additionally, integrating rule-based systems with AI-based NLG enhances

```
APPEAR(VI) AGAINST = APARECER EN
APPEAR(VI) FROM = APARECER DESDE
APPEAR(VI) BEFORE = COMPARECER ANTE
APPEAR(VI) FROM BEHIND = APARECER POR DETRÁS DE
APPEAR(VI) IN N(DRAMA,ETC) = APARECER EN N
APPEAR(VI) IN N(ACCOUNT) = APARECER EN N(INFORME)
APPEAR(VI) IN N(COURT) = COMPARECER ANTE N(TRIBUNAL)
APPEAR(VI) IN N(PUBLIC) = PRESENTARSE(REFL) EN N
APPEAR(VI) IN N(PERSON) = PRESENTARSE(REFL) EN N(PERSONA)
APPEAR(VI) ON N(SCREEN,PAPER,ETC) = APARECER EN N
APPEAR(VI) ON N(AGENDA) = APARECER EN N
APPEAR(VI) BEFORE N(COURT) = COMPARECER ANTE N(TRIBUNAL)
APPEAR(VI) AS N = PRESENTARSE(REFL) COMO N
APPEAR(VI) N(NOM-IT) TO N(HUMAN) = PARECER N(NOM-LE) A N
APPEAR(VT) TO-INF = PARECER-INF
APPEAR(VI-ACTIVE) IN N(PRINT-RAW) = PUBLICARSE(REFL)
```

Figure 4. English-Spanish comment lines for SemTab rules.

output quality. While the former provides domain expertise, the latter bolsters language generation capabilities, resulting in accurate and engaging content. This integration also minimizes errors and elevates overall content quality.

The Logos Model, with its well-established linguistic foundation, can significantly improve the accuracy of machine translation within GenAI. Its extensive scientific publications contribute to a robust understanding of language nuances, leading to more precise and contextually relevant translations. By integrating the Logos Model into GenAI, customization for specific domains becomes more effective. The model's documented resources and adaptability enable tailored solutions for various industries, ensuring that language generation aligns seamlessly with the specialized terminology and nuances of different fields. The Logos Model, being a pioneer in the field of machine translation, can enhance the natural language generation capabilities of GenAI. Its comprehensive coverage of language features, from formal to informal styles and technical to non-technical terminology, contributes to the production of more coherent and contextually appropriate text. The Logos Model's emphasis on well-defined taxonomy, data modeling, and structured data pipeline enhances the transparency and traceability of language generation processes within GenAI. This not only fosters a better understanding of how the system operates but also facilitates easier troubleshooting and improvement. Leveraging the Logos Model enables GenAI to excel in cross-language applications. Its proficiency in multiple language pairs, such as English-Portuguese and English-Spanish, provides a foundation for effective communication and translation across diverse linguistic contexts. Finally, the Logos Model, with its community of developers using linguistic environment tools, fosters collaboration and resource-sharing. Integrating this collaborative approach into GenAI can lead to a more dynamic and continuously evolving language generation system. In summary, integrating the Logos Model into GenAI brings about advantages such as improved translation accuracy, customization for specific domains, enhanced natural language generation, increased transparency, cross-language applicability, and the benefits of a collaborative community.

7 The road ahead

Despite the positive impact of LLMs like ChatGPT on text quality, they are under scrutiny for substantial issues in various fields. On the other hand, rule-based systems have plateaued in performance due to the difficulties involved in capturing the complexity of natural language. In the framework of COST Action Multi3Generation's interdisciplinary orientation, there is a consensus that both approaches could contribute to a mutually beneficial synergy. While AI tools demonstrate impressive generative performance, caution is warranted, recognizing their susceptibility to the biases of their creators. Traditional tools like OpenLogos may require more effort and pose usability challenges but offer transparency, full customization, and reproducible outcomes with diligent investment and understanding.

Integrating expert knowledge into modern AI-based Natural Language Generation (NLG) systems offers numerous advantages across various areas. Firstly, rule-based systems possess extensive knowledge bases encompassing domain-specific terminology, cultural norms, and social nuances, enabling AI to model intended meaning more accurately. Additionally, these systems excel in contextualization, handling specific contexts like medical reports or legal documents with precision. Moreover, rule-based systems effectively manage complex linguistic structures, grammar rules, and sentence formations,

contributing to faster learning and improved accuracy during AI training processes. Furthermore, the lexical resources and syntactic rules inherent in traditional systems can enhance vocabulary consistency and encompass rare language patterns, augmenting the datasets used for AI training. Importantly, incorporating rule-based methods for fact-checking ensures the accuracy of generated content, while personalization and customization capabilities cater to diverse audiences, enhancing engagement and effectiveness. Additionally, expert knowledge aids post-editing support providing guidelines for error correction and content improvement. Lastly, leveraging dictionaries and rules from traditional Machine Translation (MT) systems promotes inclusion by advancing Natural Language Processing (NLP) development for languages with limited resources, thus addressing disparities in NLP system development.

8 Conclusions

Linguistics serves as a foundational pillar in understanding that words possess both surface representations and contextual meanings, crucial for advancing NLG beyond superficial understanding. By challenging the current dependence on human trainers and emphasizing the development of linguistic resources, machines can autonomously learn words and meanings, fostering ethical AI development with greater human control. Linguistic expertise is vital for assessing language quality, providing feedback, and ensuring responsible AI outputs, mitigating the risks of unchecked machine-generated content. Furthermore, linguistic insights into language evolution underscore the importance of investing in controllable systems over post-editing, advocating for the proactive development of linguistic resources guided by computational linguists. The integration of linguistic features such as Semantico-syntactic Abstraction Language (SAL) and Semantico-Syntactic Tables (SemTab) from the Logos Model showcases practical applications of linguistic knowledge, offering potential enhancements to paraphrase generation systems and other NLP applications. In summary, the arguments presented advocate for the integration of linguistic expertise in NLG, emphasizing the importance of deep linguistic study, ethical considerations, human control, and the development of knowledge-based resources.

9 Ethics and consent

Ethics and consent were not required.

Data and software availability

Underlying data

All data underlying the results are available as part of the article and no additional source data are required.

Extended data

The following are the GitHub webpages for OpenLogos:

- Main OpenLogos GitHub webpage

  https://github.com/OpenLogos

- GitHub repository for OpenLogos SemTab Rules

  https://github.com/OpenLogos/SEMTAB-rules

For further resources related to OpenLogos, including installation and tools, repositories can be accessed via the following link: https://github.com/orgs/OpenLogos/repositories.

Author contributions

Anabela Barreiro, Multi3Generation's chair, coordinated the article, organized the structure of the manuscript, wrote the Abstract, Keywords, Plain Language Summary, Introduction, Section 3 on the Language Spectrum, Section 4 on the Linguistic Approach to NLG, Section 5 (The Logos Model, its Semantico-Syntactic Abstraction Language (SAL) and the Semantico-Syntactic Tables (SemTab), the Conclusions and Future Work, and revised the entire article.

Elena Lloret, Multi3Generation's vice-chair, wrote the first part of Section 2 on the state of the art and performed a thorough review and assessment of the manuscript.

Branislav Gerazov, Multi3Generation member, authored the latter portion of Section 2, providing insights regarding the impact of OpenLogos on machine translation. Additionally, he conducted a comprehensive review of the manuscript.

All authors participated in different phases of preparing the manuscript, encompassing writing, reviewing, and editing, thereby fulfilling the authorship requirements through their collaborative contributions.

References

1. 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-jussà, 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, June 1-3, 2022*. European Association for Machine Translation, 2022; 345–346.
Reference Source
2. Scott B: **The logos model: an historical perspective.** *Machine Translation.* 2003; **18**: 1–72.
Publisher Full Text
3. Scott B, Barreiro A: **OpenLogos MT and the SAL representation language**. In: *Proceedings of the First Interna- tional Workshop on Free/Open-Source Rule-Based Machine Translation*. Alacant, Spain, 2009; 19–26.
Reference Source
4. 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.
Reference Source
5. Amato S, Arrieta K: **Natural Language Generation in the logos model**. In: *Proceedings of the 24th Annual Conference of the European Association for Machine Translation.* EAMT, 2023; 32–39.
Reference Source

6. Scott BE: **The logos system**. In: *MT Summit II*. 1989.
Reference Source

7. Scott BE: **Biological neural net for parsing long, complex sentences**. Technical report. Logos Corporation, Mt. Arlington, NJ, 1992.

8. Scott BE: **Competence, performance, and the paradigm shift: a connectionist perspective**. Technical report. Logos Corporation, Mt. Arlington, NJ, 1992.

9. Scott BE: **Architecture and process of a hybrid symbolic/connectionist parser**. Technical report. Logos Corporation, Mt. Arlington, NJ, 1993.

10. Orliac B, Dillinger M: **Collocation extraction for Machine Translation**. In: *Proceedings of the Machine Translation Summit IX*. 2003; 292–298.
Reference Source

11. 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 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II*. Berlin, Heidelberg, Springer Berlin Heidelberg, 2011; 365–376.
Publisher Full Text

12. Barreiro A, Monti J, Orliac B, *et al.*: **Linguistic evaluation of support verb constructions by openlogos and google translate**. In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC' 14)*. ELRA, 2014; 35–40.
Reference Source

13. Reiter E, Dale R: **Building applied Natural Language Generation systems.** *Nat Lang Eng.* 1997; **3**(1): 57–87.
Publisher Full Text

14. Reiter E, Dale R: **Building Natural Language Generation systems.** *Studies in Natural Language Processing.* Cambridge University Press, 2000.
Publisher Full Text

15. Gatt A, Reiter E: **SimpleNLG: a realisation engine for practical applications**. In: Emiel Krahmer and Mariët Theune, editors, *ENLG 2009 - Proceedings of the 12th European Workshop on Natural Language Generation, March 30-31, 2009, Athens, Greece.* The Association for Computer Linguistics, 2009; 90–93.
Reference Source

16. Braun D, Klimt K, Schneider D, *et al.*: **SimpleNLG-DE: adapting SimpleNLG 4 to German**. In: Kees van Deemter, Chenghua Lin and Hiroya Takamura, editors, *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan, October-November 2019. Association for Computational Linguistics. 2019; 415–420.
Publisher Full Text

17. Chen G, van Deemter K, Lin C: **SimpleNLG-ZH: a linguistic realisation engine for Mandarin**. In: Emiel Krahmer, Albert Gatt, and Martijn Goudbeek, editors, *Proceedings of the 11th International Conference on Natural Language Generation.* Tilburg University, The Netherlands, Association for Computational Linguistics. 2018; 57–66.
Publisher Full Text

18. Ramos-Soto A, Janeiro-Gallardo J, Diz AB: **Adapting SimpleNLG to Spanish**. Association for Computational Linguistics, 2017; 144–148.
Publisher Full Text

19. Cascallar-Fuentes A, Ramos-Soto A, Diz AB: **Adapting SimpleNLG to Galician language**. In: Emiel Krahmer, Albert Gatt, and Martijn Goudbeek, editors, *Proceedings of the 11th International Conference on Natural Language Generation.* Tilburg University, The Netherlands, Association for Computational Linguistics. 2018; 67–72.
Publisher Full Text

20. Isabelle P, Foster G: **Machine Translation: overview**. In: Keith Brown, editor, *Encyclopedia of Language Linguistics (Second edition)*. Elsevier, Oxford, second edition edition, 2006; 404–422.
Publisher Full Text

21. Lloret E, Palomar M: **Text summarisation in progress: a literature review.** *Artif Intell Rev.* 2012; **37**(1): 1–41.
Publisher Full Text

22. Gadag A, Sagar BM: **A review on different methods of paraphrasing**. In: *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*. IEEE, 2016; 188–191.
Publisher Full Text

23. Saliby JG: **Survey on Natural Language Generation.** *Int J Trend Sci Res Dev.* 2019; **3**(3): 618–622.
Publisher Full Text

24. Barros C, Lloret E: **HanaNLG: a flexible hybrid approach for Natural Language Generation**. In: Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 20th International Conference, CICLing 2019, La Rochelle, France, April 7-13, 2019, Revised Selected Papers, Part II*. volume 13452 of Lecture Notes in Computer Science. Springer, 2019; **13452**: 507–519.
Publisher Full Text

25. Vicente ME, Lloret E: **A discourse-informed approach for cost-effective extractive summarization**. In: Luis Es-pinosa Anke, Carlos Martín-Vide, and Irena Spasic, editors, *Statistical Language and Speech Processing - 8th International Conference, SLSP 2020, Cardiff, UK October 14-16, 2020, Proceedings,*. volume 12379 of Lecture Notes in Computer Science. Springer, 2020; **12379**: 109–121.
Publisher Full Text

26. Barros C, Vicente ME, Lloret E: **To what extent does content selection affect surface realization in the context of headline generation?** *Comput Speech Lang.* 2021; **67**: 101179.
Publisher Full Text

27. Vicente ME, Sepúlveda-Torres R, Barros C, *et al.*: **Can text summarization enhance the headline stance detection task? Benefits and drawbacks**. In: Josep Lladós, Daniel Lopresti, and Seiichi Uchida, editors, *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*. volume 12822 of Lecture Notes in Computer Science, Springer, 2021; **12822**: 53–67.
Publisher Full Text

28. Vaswani A, Shazeer N, Parmar N, *et al.*: **Attention is all you need**. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems.* NIPS' 17, Red Hook, NY USA, Curran Associates Inc, 2017; 6000–6010.
Reference Source

29. Erdem E, Kuyu M, Yagcioglu S, *et al.*: **Neural Natural Language Generation: a survey on multilinguality, multimodality, controllability and learning.** *J Artif Intell Res.* 2022; **73**: 1131–1207.
Publisher Full Text

30. Touvron H, Lavril T, Izacard G, *et al.*: **LLaMA: open and efficient foundation language models**. Computing Research Repository, arXiv: 2302.13971. Version 1. 2023.
Reference Source

31. Chiang WL, Li Z, Lin Z, *et al.*: **Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality**. 2023.
Reference Source

32. Gatt A, Krahmer E: **Survey of the state of the art in Natural Language Generation: core tasks, applications and evaluation.** *J Artif Intell Res.* 2018; **61**(1): 65–170.
Reference Source

33. Vicente M, Lloret E: **Exploring flexibility in Natural Language Generation through discursive analysis of new textual genres**. In: José F Quesada, Francisco-Jesús Martín Mateos, and Teresa López Soto, editors, *Future and Emerging Trends in Language Technology. Machine Learning and Big Data*. Cham, Springer International Publishing, 2017; 98–109.
Publisher Full Text

34. Reiter E: **Natural Language Generation challenges for explainable AI.** *CoRR.* abs/1911.08794, 2019.
Publisher Full Text

35. Barreiro A: **Make it simple with paraphrases: automated paraphrasing for authoring aids and Machine Translation**. PhD thesis, Universidade do Porto, Porto, Portugal, 2009.
Reference Source

36. Chaudhury S, Rao A, Sharma DM: **Anusaaraka: an expert system based Machine Translation system**. In: *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010).* 2010; 1–6.
Publisher Full Text

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

Open Peer Review

Current Peer Review Status: ✗ ?

Version 1

Reviewer Report 23 October 2024

https://doi.org/10.21956/openreseurope.18750.r42996

© **2024 Reiter E.** 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.

?

Ehud Reiter
University of Aberdeen, Aberdeen, England, UK

This paper does two things:
(A) it argues that linguistically-based approaches to NLG make sense and add value in many contexts.
(B) Describes OpenLogis, which is a resource for researchers interested in linguistic approaches

With regard to (A), I certainly agree that in many cases linguistics/rule-based approaches work better than LLMs, and in other cases a combination is best. Indeed, I make related points in a book I have recently written about NLG (https://ehudreiter.com/book/). However, to be honest the paper could make this point more strongly. The argumentation is high-level and conceptual, it would be stronger if it as supported by data, case studies, examples, and more links to related work (as is usually done in position papers published in ACL, for example)

With regard to (B), the OpenLogos model sounds like it could be really useful resource, but I would like to see how it compares to other resources for linguistic/rule-based NLG and MT. It also would be useful to get more information about what resources are available. Eg is there good documentation and tutorials? Which languages are supported, and how much depth are they supported in? How is OpenLogos used from a software perspective (eg, is there a Python API). Who has successfuly used OpenLogos? Its hard to assess its utility without such information

From a presentation perspective, the authors might wish to more clearly different between statistical and neural approaches. For example, on page 7 they criticise "conventionally used n-gram alignments", but this doesnt apply to the neural models used by most current MT systems.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Natural Language Generation

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.

Author Response 23 Oct 2024

Anabela Barreiro

Thank you for your valuable feedback. We plan to incorporate your suggestions and address your criticism to improve the article and submit a revised version.

Competing Interests: No competing interests were disclosed.

Reviewer Report 10 September 2024

https://doi.org/10.21956/openreseurope.18750.r43668

© 2024 Gonçalo Oliveira H. 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.



Hugo Gonçalo Oliveira
University of Coimbra, Coimbra, Portugal

This paper presents OpenLogos as a linguistic-oriented solution for Natural Language Generation in opposition to simply using current Large Language Models (LLMs). It points out some limitations of LLMs and discusses the potential benefits of their combination. However, despite sharing the vision of the paper, I have to say that it is too theoretical, with several assumptions sounding more like "opinions", with few technical details and no empirical confirmation.

For instance, it lacks actual examples of where LLMs fail, before describing what OpenLogos provides for minimising such issues; or, technically, how OpenLogos could be combined with the LLMs for better solutions.
Ideally, some experimentation should be performed for confirming the discussed ideas.

This does not mean that most of the discussed arguments are not true. They just do not go further than what we can already read on the abundant literature on Transparent / Interpretable / Explainable Artificial Intelligence (e.g., "the integration of linguistic knowledge can help GenAI systems"), leaving some doubts on the actual contribution.
The description of OpenLogos does not go deeper than the presentation of some valuable human-crafted resources that it includes. Of course, more detailed descriptions can be found in dedicated papers, but, again, how LLMs could leverage on the available resources is left open.
Briefly, the three goals proposed in the abstract end up not being completely addressed and lack materialisation.

Moreover, the paper disregards the flexibility of LLMs, i.e., the same architecture, and often the same model, can easily be adapted to a broad range of language-oriented tasks.
This is one of the reasons that they are being used everywhere, even when they are not the best solution, only the easiest.

Additional suggestions:

On the combination of linguistic knowledge and LLMs, I suggest giving some words on Knowledge Injection, starting, e.g., with the following paper:
Martino, A., Iannelli, M., & Truong, C. (2023, May). Knowledge injection to counter large language model (LLM) hallucination. In European Semantic Web Conference (pp. 182-185). Cham: Springer Nature Switzerland.

I would not put such a well-defined boundary on the limitations of computers / current models when it comes to creative language, because these models can always be trained for better dealing with different kinds of language.
Specifically on computers and creativity, I would suggest some reading on Computational Creativity, such as the following book:Ref-2

Minor suggestions:

p5, col1:
Bard was shut down in March 2023 and could perhaps be replaced by more recent models (e.g., Gemini).

On the "similar methodology" for different tasks, in many cases, and often for LLMs, tasks are performed by the pre-trained model, not even requiring fine-tuning.

p5, col2:
The sentence starting with "It represents a vital resource for the MT community..." could be complemented of some references of previous work using OpenLogos.

The URL of the OpenLogos webpage could be added in a footnote.

"relatively well" is too subjective.

p6, col1:
"bigger" -> "greater"?

Please rephrase the sentence starting with "As the natural language is used in fiction...".

On "representation of the word at a surface level, rather than its meaning", I think that I understand what the authors mean, but LLMs do more than representing words at surface level. Their representations of words and sequences depends on the context, which is related to meaning. This, of course, does not mean that they "understand language".

The sentence starting with "The apparent power of inference and deduction is a result of human feedback..." is true, but I do not see it as a limitation towards linguistic-oriented approaches, which also rely on input input / handcrafted resources, that may be manually fixed, if necessary.

p11, col2:
References of existing studies should be provided after the statement "they are under scrutiny for substantial issues in various fields."

Form:
Footnotes and bibliographic references are marked equally, which may add confusion.

Several figures (2, 3, especially 4) are blurred, which can harm readability.

In Table 2, the meaning of the numbers in the columns (s_wt, superset, set, subset) is not clear. Perhaps they could be replaced by human-interpretable values or just removed?

I am not sure if this is the case, but the "plain language summary" sounds like it was partly generated by a LLM.

References

1. Martino.A, Iannelli.M,Truong.C: Knowledge injection to counter large language model (LLM) hallucination. In European Semantic Web Conference. Reference Source
2. Veale.T,Cardoso.F. A.: Computational creativity: The philosophy and engineering of autonomously creative systems. Reference Source

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence, Natural Language Processing, Computational Creativity, Semantic Web.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 23 Oct 2024

Anabela Barreiro

Thank you for your valuable feedback. We plan to incorporate your suggestions and address your criticism to improve the article and submit a revised version.

Competing Interests: No competing interests were disclosed.