Decoding Complexity: A Mathematical Framework for Enhanced Translation Comprehension

Éric André Poirier and Antsa Nasandratra Nirina Avo

Département des langues modernes et de traduction, Université du Québec à Trois-Rivières, Eric.Poirier@uqtr.ca
Département des sciences de l'éducation, Université du Québec à Trois-Rivières, Antsa.Nasandratra.Nirina.Avo@uqtr.ca

Abstract

Machine translation tools have demonstrated substantial progress in enhancing translation accuracy since the emergence of artificial intelligence. However, challenges persist in reasoning (or the lack thereof), considering contexts, addressing specific word games, and interpreting very long or very short sentences—those exceeding 50 and falling below 7 words (Bowker, 2023 : 893). Additionally, accurately translating technical or specialized terms and their variations remains a hurdle. This research introduces a categorical mathematical formalization of the comprehension stages in translation, along with a model for calculating acceptances (specific meanings of words) during the verification of meaning hypotheses. The goal is to elucidate the comprehension process and integrate contextual considerations. The formalism delineates a series of fundamental cognitive operations involved in comprehension. Furthermore, it advocates for evaluating meaning hypotheses using logical modalities, particularly hypostases, described as phrases (groups of words)—a unit of discourse rather than language—signifying the structure of arguments conveying the speaker's knowledge. The strength of our proposed mathematical model lies in its independence from both source and target languages, as well as the subjectivity of text authors or translators. Additionally, the assessment of meaning hypotheses relies on verifiable logical modalities, ensuring a reliable, explicable, and controllable outcome.

Keywords: translation understanding, mathematical formalization, logical modalities, exponentiation, hypostases.

Foreword

The crafting of this scholarly article was a collaborative effort involving two authors and two systems engaged in intricate interactions. In this section, we provide a succinct overview of the primary stages and methodologies employed in this endeavor. Initially, the content was drafted in French and subsequently translated into English utilizing an online machine translation service, specifically DeepL Translator, accessed on January 25, 2024. Following this initial translation, the resultant English text underwent meticulous revision and rephrasing, adhering to academic conventions, paragraph by paragraph, utilizing the Chat-GPT AI generative tool provided by OpenAI. Additionally, select paragraphs originally authored in English and integrated during the revision phase were refined and reviewed with the aid of the Chat-GPT AI generative tool. Throughout the entire writing process, stringent attention was paid to content, terminology, and linguistic coherence, with both authors, fluent in Quebec French and Malagasy, respectively, and proficient in English as a second or third language, meticulously editing and harmonizing the text to the best of their abilities.

Introduction

The advent of deep learning in the field of artificial intelligence (AI), employing layered neural networks, has significantly advanced research in cognitive tasks, specifically in data recognition, classification, and automatic natural language processing (NLP), encompassing machine translation (MT). Although machine translation has undeniably benefited from these strides, particularly remarkable progress has been achieved through the utilization of word embedding technology. This innovation enables the vectorization (digitization) of word occurrences within textual data. Harnessing these enhanced digital text processing capabilities, Large Language Models (LLMs) have given rise to generative AI. This development facilitates the generation of coherent and contextually relevant texts, fostering seamlessly fluid dialogical interactions with human users, thereby surpassing the Turing test.

Thanks to the extensive parallel translation corpora of Large Language Models (LLMs) compiled from publicly available online documents, often bypassing copyright protections, AI-based Machine Translation (MT) systems have significantly enhanced the accuracy of translating individual words and word groups from a source language to a target language. Despite these advancements, there remains considerable room for improvement in the realm of translation reasoning (Poirier, 2017). This limitation arises from the inherent nature of MT systems, which operate through transcoding, lacking the capability to engage in reasoning, and making decisions devoid of a comprehensive understanding of the underlying textual meanings. Furthermore, the indispensable validation role of humans persists (Poirier and Roy, 2023).

Recognizing this gap, we emphasize the critical need to develop a translation formalism that sheds light on the reasoning and information processing inherent in human translation. In pursuit of this objective, this article concentrates on the initial stage of the translation process, specifically the comprehension of the source text. How can we formally characterize the cognitive operations involved in human translation? Can we construct models for the cognitive processes related to comprehension? To better address these inquiries, we opt to deconstruct the cognitive processes executed by translators during the translation task.

The objective of this article is to formally delineate certain cognitive processes involved in comprehension, executed by professional translators during the translation of documents from a source language (A) into another language (B), referred to as the target language. The act of translation entails numerous intricate cognitive operations, and our model seeks to elucidate the fundamental operations of comprehension that exhibit recurrence in the practices of most translators, irrespective of the type of text or language pair involved.

The imperative for a model that transcends translator subjectivity, language arbitrariness, and the determinism of stylistic conventions underscores the significance of the mathematical model proposed herein. Our approach adopts an intuitionistic formalism of reasoning, drawing inspiration from Daniel Gile sequential model of translation (Gile, 2005). This model is designed to accommodate the translation of diverse text types with any language pair, ensuring a level of universality and objectivity in its application.

Methodology

The model presented in this paper seeks to elucidate the processing of two categories of translational entities: individual words or word clusters within the language, and phrases or syntagms within the discourse. This elucidation is undertaken with the primary objective of comprehending the source text as an initial phase in the translation process. Notably, some elements within these two sets of entities overlap. Phrases consistently amalgamate words or groups of words, and occasionally a phrase in discourse is condensed into a single word or a group of words, as observed in titles, intertitles, enumerations, tables, and similar contexts.

We present a mathematical model of translation comprehension grounded in analogy, which broadly captures various implication relations (both unidirectional and bidirectional) between two elements—be they words, groups of words, or phrases—within a set, whether in language or discourse. Analogy finds dual applications in translation. Firstly, it serves to characterize linguistic relations, elucidating semantic nuances such as synonymy (mutual implications or similar meanings), antonymy (opposite meanings without reciprocal implication), heteronymy (implied relationships between words or groups), hyponymy (implications from one word or group to another), meronymy (part-whole relationships, such as handlebars and bicycle), and partonymy (belonging to the same class, exemplified by the relationship between dog and cat within the class of pets). This description of implication-based meaning relations draws inspiration in part from Greg Lessard's work (2014).

To obviate ambiguity, the term "terms" is defined herein as the syntactic and semantic building blocks constituting a sentence. These terms assume diverse functions and can be interconnected through various relationships, be they semantic or syntactic in nature.

In contrast, the concepts of comprehension and reformulation pertain to cognitive processes associated with a tangible object (the terms) and the abstract notion of knowledge (Michel, 2021) building and its complex interactions with unknown information. Specifically, reformulation involves expressing the terms or phrases in alternative terms to validate comprehension. Demonstrating understanding of a term is an iterative process that retraces the cognitive steps taken until the accurate meaning of the term is apprehended, thereby enabling its reformulation.

The formalism

Phrases[1] are formally regarded as a concise mathematical category, aligning with Mac Lane's perspective (Mac Lane, 2013). In this categorical framework, the objects are sets of syntagms within a document slated for translation, and the morphisms between these objects encapsulate the syntactic and/or semantic associations among terms. These associations exhibit associativity and adhere to the properties characteristic of category morphisms. If $(X, Y, Z) \in Ob(C)$, there is $(Hom\_C\ (X,Y) \times Hom\_C\ (Y,Z) \rightarrow\ Hom\_C\ (X,Z)$.

[1] Construed here as terms or words (although the literature also allows for the definition of relationships between words, wherein a cluster of interconnected terms forms a group of words).

Additionally, when considering three terms X, Y, Z having semantic and/or syntactic relationships (wherein X is related to Y, and Y is related to Z), it follows that X and Z also share a semantic and/or syntactic relationship. For instance, in the case of the terms "cars," "fuel," and "environment," a semantic relationship exists between "car" and "fuel," and between "fuel" and "environment," implying a relationship between "car" and "environment.". If (X,Y)≠(X^',Y^') then Hom_C (X,Y) ∩ Hom_C (X^',Y^')=∅.

Moreover, if two pairs of terms lack any relation, there will be no intersection between the two relations linking these term pairs. For example, consider two pairs composed of distinct terms, such as "(cars, fuel)" and "(school, computer)." In the absence of apparent semantic or syntactic relations within these term pairs, there will be no connections between the semantic and syntactic relations of "(cars, fuel)" and "(school, computer)."

Product in the Term Category

The term category, as thus defined, incorporates a product that results from the conjunction of two terms. This product is structured by the presence of a unique semantic and/or syntactic relation between the product (TP) of two distinct terms (T1 and T2) and another term (T3) that is linked to both terms (Nirina Avo et al., 2022) (see figure 1). To illustrate, consider the terms "cat"(T1) and "dog" (T2); their product yields the term "animal" (TP) In practice, the term "animal" establishes semantic and/or syntactic relations (f and g) with both "dog" and "cat," hence being present in the set of associated terms for both. If we then introduce the term "mammal," (T3) which also shares semantic and/or syntactic connections (p_1 and p_2) with "dog" and "cat," a semantic and/or syntactic relationship (h) emerges between the product of the terms "animal" and the third term "mammal."

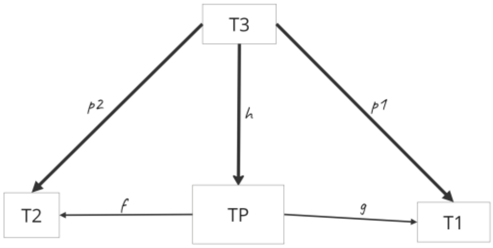

Figure 1: Product of a category

It should be noted that the list of related terms is not exhaustive, and it's easy to make a mistake if you don't rely on the morphological definition of the category. For example, let's consider the terms ostrich (T1) and pink flamingo (T2), which have the product term birds (TP), because it has semantic and/or syntactic connections with the term birds, and birds is present in the list of related terms for both ostrich and pink flamingo. To check the product, we need to look at a third term (T3) that has semantic and/or syntactic connections with the first two terms and check if it is still fibered with the product term. In other words, the terms ostrich and pink flamingo have semantic and/or syntactic connections with omnivore (T3). As a result, there is a unique semantic and/or syntactic connection (h) between the term omnivore and birds. The term flying birds cannot be used because flying birds has no semantic and/or syntactic connections with ostrich and therefore does not meet the conditions of a fiber product in a category.

Final and Exponential Objects in the Category of Terms

The fundamental objects of comprehension fall into two categories, distinguishing between their roles in naming familiar objects of knowledge (and, incidentally, for naming unfamiliar objects) or designating both known and unknown objects of knowledge, as proposed by Frath (2015). Firstly, we encounter knowledge, closely intertwined with extralinguistic reality – where rationality extends beyond language, especially in terms of reasoning. This category comprises words or groups of words, termed as sets of associated terms, falling under the umbrella of language, serving the purpose of naming. These terms refer to various elements such as objects, concepts (common nouns), events (holidays or historical facts), routine facts, or places (sites, cities, locations, regions), as well as entities or beings distinct from places (proper nouns). They articulate relationships between words or groups of words, facilitating the expression of connected knowledge or knowledge assembled for the communicative needs of speakers.

Secondly, analogy relations delineate syntactic dependencies among elements within syntagms on the discursive plane, a concept embedded in our formalism as the contextualization of terms within a sentence. The dependency relation can be described as a form of mutual implication between subject and predicate in discourse: the subject implies the predicate, and vice versa, evident in the definitional periphrases used to describe each other – the subject being what is discussed, and the predicate being what is asserted about it. Conversely, the verb unidirectionally implies the complement, as the complement relies on the verb. Analogical relations also encompass the associations of adverbs with verbs, adjectives, or nouns, as well as the relations of adjectives to nouns, all characterized by unidirectional implication.

The category of terms, as delineated, is both Cartesian and closed, featuring the document theme as the final object and the associated terms as the exponential object^context. This exponential object^context encapsulates the contextual application of relations to the associated terms. Notably, there exists a consistent semantic and/or syntactic relationship between the document terms and the document theme, contextualizing them within the document, thereby ensuring the commutativity of the diagram depicted in Figure 1.

In practical terms, each chosen term in a document corresponds to a set of associated terms (λ_f: selection → associated terms), where the context influences the selection to filter out pertinent associated terms (f: selection × context → associated terms). This categorization reflects the hypothesis of meaning attributed by the translator to the selection. The relevance of this categorization is consequently subject to evaluation (eval: associated terms^context × context → associated terms) through modal logic reasoning to ascertain its accuracy.

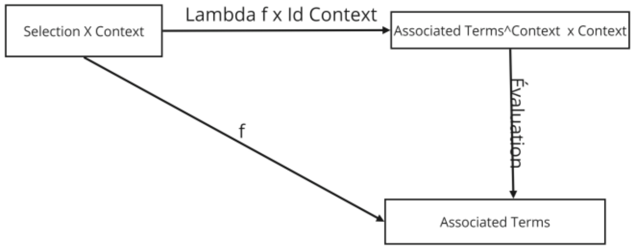

Figure 2: Exponentiation of a term

Consider the word abstract in the sentence: “I have difficulty dealing with the abstract - let's discuss particular cases.” The generally associated terms for the word “abstract” may include {general ideas, concepts, theoretical, summary, statement, content, short form, important facts, article, paper, book, painting, shape, line, colour, texture, etc.}. However, within the context of the phrase the associated terms such as idea, concept, summary, statement, painting, facts, line, colour or texture (for examples) would not make sense. To address this, a sorting operation is executed to derive the set of associated terms that are contextually relevant to the sentence. Subsequently, this sorting undergoes evaluation to verify the accuracy of the categorization – essentially assessing the validity of the meaning hypothesis, as per Daniel Gile's framework (Gile, 2005). Figure 2 illustrates this process of validating the meaning hypothesis in translation by the evaluation, presented as an exponentiation operation.

Evaluating the hypothesis of meaning

This evaluation for a term selected in translation involves a logical assessment of the significance of the sorting applied to the set of associated terms in accordance with the context.

To facilitate this evaluation, we turn to Augustin Sesmat and Robert Blanché's (1966) logical hexagon, an expansion of Aristotle's logical square that introduces a shift in the interpretation of the "contingent" modality. Advances in logical studies have demonstrated the possibility of broadening the expressivity (the ability to formulate) of these modalities.

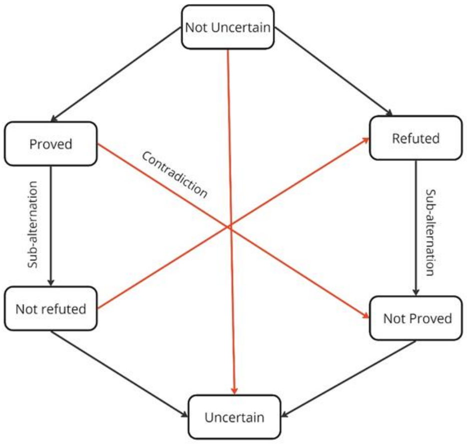

Figure 3: Logical hexagon (Sallantin et al., 2023)

Logic addresses not only the binary states of truth and falsehood but also what is prouven (and therefore true), what is refuted (and therefore false), and what remains uncertain (neither proven nor not refuted). In the realm of intuitionistic logic, proof requires a constructive nature, and refutation necessitates a counter-example (Sallantin et al., 2023). Let's consider two proof devices, namely empirical and formal, along with three logical inferences: induction2, abduction3, and deduction4. Additionally, we entertain two possible interpretations for a statement: “de dicto5” and “de re6.”

The intricacy lies in evaluating the significance of the sorting of associated terms based on the logical hexagon illustrated in Figure 3. The objective is to establish that if the relevance of the sorting is not proven “in a certain way”

[2] From propositions A and B, we deduce “A implies B.” This mirrors the physicist's approach, extracting laws and causal relationships from observations.

[3] From propositions B and “A implies B,” we deduce A. This corresponds to the methodology of a detective or archaeologist, generating hypotheses from observations and constructing possible scenarios.

[4] From propositions A and “A implies B,” we deduce B. This aligns with the approach of a mathematician, proving theorems based on axioms and demonstrations.

[5] Refers to what is said

[6] Pertains to the nature of the thing or fact spoken of

within the context, and not refuted “in another way” that avoids contradiction, then it must be adjudged as uncertain or a “hypostasis” (Nirina Avo et al., 2019). In this context, a hypostasis means the expression of what is being discussed and corresponds to what John Duns Scotus called a “haecceity” or the effect of individuation.

Table 1 delineates the six forms of reasoning obtained: empirical induction, empirical deduction, empirical abduction, formal induction, formal deduction, and formal abduction. Hypostases within the same column indicate that they are unproven for a specific type of reasoning and proof device. Hypostases within the same row suggest that they are not refuted for another type of reasoning and a different proof device. According to this table, a hypothesis is, therefore, a hypostasis not proven by empirical induction and not disproven by formal deduction.

Table 1. Hypostases (Nirina Avo et al., 2019)

	Not proved					
Not refuted	**Empirically Induce**	Classification	Paradox	Formalism	Aporia	Approximation
	Variance	**Empirically Abduct**	Index	Data	Variable	Phenomenon
	Mode	Variation	**Empirically Deduce**	Belief	Dimension	Event
	Axiom	Value	Speculation	**Formal Induce**	Structure	Invariant
	Hypothesis	Definition	Problem	Theory	**Formal Abduct**	Method
	Principle	Paradigm	Domain	Law	Object	**Formal Deduce**

In accordance with the hypostasis table and the formulation of the term category, a hypothesis of meaning is delineated as a designated morphism linking the set of correlated terms to the set of pertinent associated terms. Consequently, the translator possesses the capacity to engage in reasoned analysis of the hypothesis of meaning by assessing the pertinence of this morphism, as illustrated in Figure 2. A hypothesis of meaning, given its definition, remains uncertain since it lacks proof through empirical induction and is not refuted by formal abduction. If the result of the evaluation of the meaning hypothesis is not uncertain, it could be certain (with the caveat that the reverse is not necessarily true), when the translator can substantiate it through empirical induction or disprove it through formal abduction. The status of a hypothesis that is considered controversial or not uncertain can be due to knowledge or information that are ambiguous or vague in two ways: it cannot be empirically or formally excluded from the inferences of the text (and is therefore not uncertain) and it also cannot be empirically or formally inferred from the text (and is therefore uncertain)

This iterative process continues with a renewed search for a relevant associated term, effectively generating a new hypothesis of meaning in the opposing scenario.

To illustrate, consider the term “bank”; the attributed meaning to the term in the sentence “We should deposit this money in a bank” might be the “financial institution where people and businesses can invest or borrow money”. This hypothesis holds true if the translator can substantiate it through empirical induction or disprove it through formal abduction, affirming that, in the context of the sentence, the term “bank” is associated with a financial institution. Evaluating this hypothesis of meaning therefore comes down to assessing the relevance or robustness of the morphism between the set of terms (deposit, money, bank) and the set of relevant associated terms (depository financial institution, banking, organization, financial services, commercial services, money lending, vault, etc.), as opposed to another set of terms associated with the word “bank” (sloping land, sides of a river, lake, body of water, etc.).

The primary objective of our formal framework proposal is to establish a foundation for describing the intricacies inherent in the process of comprehension and the sequential stages constituting a hypothesis of meaning. In this paper, our endeavor has been to explain the comprehension process and the assessment of the hypothesis of meaning from a categorical standpoint. Employing a categorical formalism affords us the opportunity to ascend to a higher level of abstraction, facilitating a nuanced understanding of the diverse cognitive operations invoked by a translator in the course of translation. Concurrently, it renders the challenges posed by prevailing probabilistic and statistical models (used in machine translation) more readily comprehensible.

Discussion And Conclusion

Our model, founded upon the elucidation of analogical relations between language and discourse, primarily seeks to delineate the cognitive mechanisms employed by translators in selecting the pertinent meanings of words or groups of words from the language. These selected meanings, owing to their incorporation into the source text, metamorphose into discourse facts, or syntagms. While this cognitive operation is recurrent, it is not always systematic, as translators may engage in it as required. In instances where established knowledge objects and syntagms are concerned, translators might opt for a cognitive shortcut, directly accessing their memory of extralinguistic knowledge and recurrent discourse formulations that they have previously analyzed and understood.

The process of selecting the meanings of words or groups of words within syntagms is indispensable for the initial stage of translation—comprehending the text to be translated. This operation draws on the translator's familiar knowledge base as well as new information presented in the text. Formalization of this selection of meanings employs the operation of exponentiation within a closed Cartesian category of terms, affording it a formal status distinct from the probabilistic occurrences upon which contemporary machine translation tools rely. Semantic calculation is further formalized through evaluation using the logical hexagon and the hypothesis of hypostasis. Reasoning within category theory enhances the precision of meaning computation.

This research, while illuminating the decomposition of comprehension and the evaluation of meaning hypotheses in translation, is restricted in its scope. Extending this categorical model to contemplate propositions consisting of two or three syntagms (subject, predicate, and possibly a sentence complement), as well as sentences comprising one, two, three, or more propositions, would enrich our understanding. Both propositions and phrases, whether they serve as subjects, predicates, or sentence complements, exhibit recursion, with a proposition theoretically accommodating an infinite number of subordinate or relative propositions, and phrases theoretically accommodating an infinite number of complements or expansions. In essence, transitioning from evaluating the relevance of a hypothesis of associated terms in a context to evaluating the relevance of the meaning of arguments associated with the terms in a context does not alter the methodological approach; rather, it enhances its clarity. It is noteworthy that the conventional focus on the meaning of an argument, whether proposed by an AI or not, is more prevalent than the meaning of a set of terms that contextually indexes a text.

The advantage of this model is that it is based on the operating principle of the hypothesis of meaning, which is the same regardless of the language in which we work. However, the model easily accounts for the bilingual or multilingual skills of text readers, because it describes in a sequential way the operation to be performed in a hypothesis of meaning calculation.

Tools

Antidote 11 (software, bilingual version 3.1). Druide informatique, Montreal, 2022

Centre de recherche inter-langues sur la signification en contexte - EA 4255 (Crisco) (2024). Dictionnaire Électronique des Synonymes (DES). on line : https://crisco4.unicaen.fr/des/

ChatGPT version 3.5, Open AI: openai.com/chat

"Review the next paragraph and adjust the style if needed for a scientific paper and an academic style. Also, please make changes to the wording for a more idiomatic phrasing. Please explain any mistake grammar or otherwise that you find." Answer to the authors, 25 January 2024.

"Correct and rewrite the next paragraph in a more idiomatic and academic style for a scientific paper." Answer to the authors, 25 January 2024.

DeepL Translate, Deepl SE: https://www.deepl.com/translator.

References

Blanché, Robert (1969). Structures intellectuelles. Essai Sur l'Organisation des Concepts. Frath, Pierre (2015). Dénomination référentielle, désignation, nomination. *Langue française*, no. 188, 33-46.

Bowker, L. (2023). Translating Research into Practice: Plain Language and Writing for Machine Translation Guidelines. Proceedings of the Association for Information Science and Technology, 2023, vol. 60, no 1, p. 892-894.

Gile, Daniel (2005). "Chapitre IV. Un modèle séquentiel de la traduction" in *La traduction. La comprendre, l'apprendre*, 101-135. Paris. Presses universitaires de France. Retrieved from: https://www.cairn.info/la-traduction-la-comprendre-l-apprendre--9782130525004-page-101.htm

Luzeaux, Dominique, Sallantin, Jean and Dartnell, Christopher (2008). Logical extensions of Aristotle's square. *Logica Universalis, 2*(1), 167-187.

Mac Lane, Saunders (2013). *Categories for the working mathematician* (vol. 5). Springer Science & Business Media.

Michel, Johann (2021). Qu'est-ce que la compréhension ? *Revue philosophique de la France et de l'étranger, 146* (2), 163-182. Retrieved from: https://doi.org/10.3917/rphi.212.0163

Nirina Avo, Antsa Nasandrata, Luzeaux, Dominique and Sallantin, Jean (2022). "Une modélisation catégorielle du débat numérique." in Danièle Bourcier, Paul Bourgine and Salma Mesmoudi (eds.), *Systèmes complexes; théories et pratiques*, Res-Systemica Libri AFSCET, 231-249.

Nirina Avo, Antsa Nasandrata, Sallantin, Jean, Randriamahaleo, Solo and Pinet, Véronique (2019). Les hypostases : une classification de rapports d'incertitudes (traduits en language courant) qui fondent, alimentent et dynamisent le débat, la discussion et la controverse scientifiques. In 7ème journée épistémologique de l'université de Montpellier, France.

Poirier, Éric (2017). Entre comparaison et raison: la qualité de la traduction automatique. *Circuit, Printemps 2017, no.* 133. Retrieved from: https://depot-e.uqtr.ca/id/eprint/9568/1/POIRIER_E_14_ED.pdf

Poirier, Éric A. and Roy, Jean-Hugues (2023). L'outil Ultrad de La Presse canadienne: la traduction automatique dans un contexte journalistique. *Traduction, terminologie, Rédaction, 36*(1), 71 - 105. Retrieved from: https://doi.org/https://doi.org/10.7202/1107567ar

Sallantin, Jean, Dominique Luzeaux and Sylvaine Jenny (2023). Le débat numérique. Le média d'un renouveau démocratique. Spartacus-idh.