Wikimedia Research Fund 2025 - Research Proposal

Lexeme based approach for the development of technical vocabulary for underserved languages: A case Study on Moroccan Darija

Anass Sedrati, Reda Benkhadra, Mounir Afifi
Wikimedia MA User Group

Jan Hoogland
Nimar Rabat, Leiden University

Abstract

This project addresses the challenge of developing technical vocabulary for low-resource, unstandardized languages, using Moroccan Darija as a case study. The research aims to create a methodology that empowers Wikimedian communities to generate new terms, thereby enriching their wikis and promoting knowledge equity.

The methodology will leverage Wikidata lexemes and involve editors in the creation of new, grammatically sound, and semantically accurate technical words. This approach seeks to overcome the limitations of current ad-hoc methods, which lead to inconsistencies, edit conflicts, and hinder the development of comprehensive knowledge bases in underrepresented languages.

The project will work on data collection, linguistic analysis, and the development of accessible guidelines, resulting in a methodology that will be applied to generate and evaluate new terms to be assessed through a community survey. The research findings will be disseminated to facilitate adoption by other language communities, with project outcomes including enriched Wikidata lexemes, datasets, scripts, and a scientific publication.

Introduction

Date: July 1, 2025 – June 30, 2026.

The project, "*Lexeme-based approach for the development of technical vocabulary for underserved languages*," addresses the scarcity of domain-specific academic terminology in low-resource unstandardized languages, a challenge exemplified by Moroccan Darija. While English dominates academic research (Marginson & Xu 2023, Ortega, 2020), many languages struggle to keep pace with emerging scientific concepts (Amano et al. 2016). This linguistic disparity creates significant obstacles for knowledge equity and accessibility within Wikimedia projects, particularly Wikipedia. This is even more relevant when languages are unstandardized, as there are no official bodies working on the language, leaving this work to the locutors themselves.

Impact Statement

The lack of standardized terminology directly impedes the growth, quality, and inclusivity of Wikipedia. **Unequipped** editors often face the burden of individually translating or creating equivalents for English scientific terms, or relying on often inaccurate machine translations (Nigatu et al. 2024). These ad-hoc solutions consume valuable editor time and can result in conflicting term proposals, necessitating time-consuming moderation by administrators. Such conflicts can diminish editor engagement and discourage contributions, ultimately hindering the development of comprehensive and reliable knowledge in underrepresented languages (Smirnov et al. 2023).

In the 2019 Wikimedia Wishlist Survey, a major concern raised by the community was the inconsistency of scientific terms across different languages and the lack of a unified connection between them (Meta Wiki, 2019). This problem **undermines the goal of providing a unified and conceptually linked knowledge base**. Rather than developing a tool to address this, the present research investigates the potential of Wikidata lexemes as a **structured and multilingual solution**. With this approach, our aim is to empower Wikimedian communities to create richer, more accurate, and more consistent technical content in their own languages, bridge linguistic gaps, and foster greater participation and knowledge equity.

More specifically, the project aims to develop methodologies that Wikipedia editors can use to propose new terms in their native language—terms that adhere to the language's grammatical, syntactic, and morphological rules while accurately conveying the intended meaning. By leveraging a large dataset containing terms, their etymologies, derivatives, meanings, and interrelations, we will identify and extract linguistic patterns. These patterns will then be applied to additional concepts, enabling the generation of new terms that effectively address the challenges and constraints outlined above.

Hypotheses

This research aims to test the following hypotheses:

1. Wikidata lexemes provide a sufficient foundation for generating new, accurate technical terms in low-resource unstandardized languages.
2. Linguistic patterns extracted from existing lexical data can be effectively applied to create new terms that adhere to a language's grammatical, syntactic, and morphological rules.
3. A (Wikimedia) user-friendly methodology can be developed, enabling editors without specialized linguistic expertise to contribute meaningfully to the creation of new technical vocabulary in their native languages.

Research Questions

This project seeks to answer the following research questions:

1. How can Wikidata lexemes be effectively leveraged to generate new technical terms in low-resource languages?
2. What linguistic patterns can be extracted from existing lexical data to facilitate the creation of new terms that adhere to a language's grammatical, syntactic, and morphological rules?

3. How can a methodology be developed to enable Wikipedia editors, without specialized linguistic expertise, to contribute to the creation of new, standardized technical vocabulary in their native languages?

By investigating the research questions and testing the hypotheses, the current project aims to provide a robust and accessible methodology for generating new terms, directly addressing the challenges faced by editors and promoting greater linguistic diversity and knowledge equity within the Wikimedia movement.

Related work

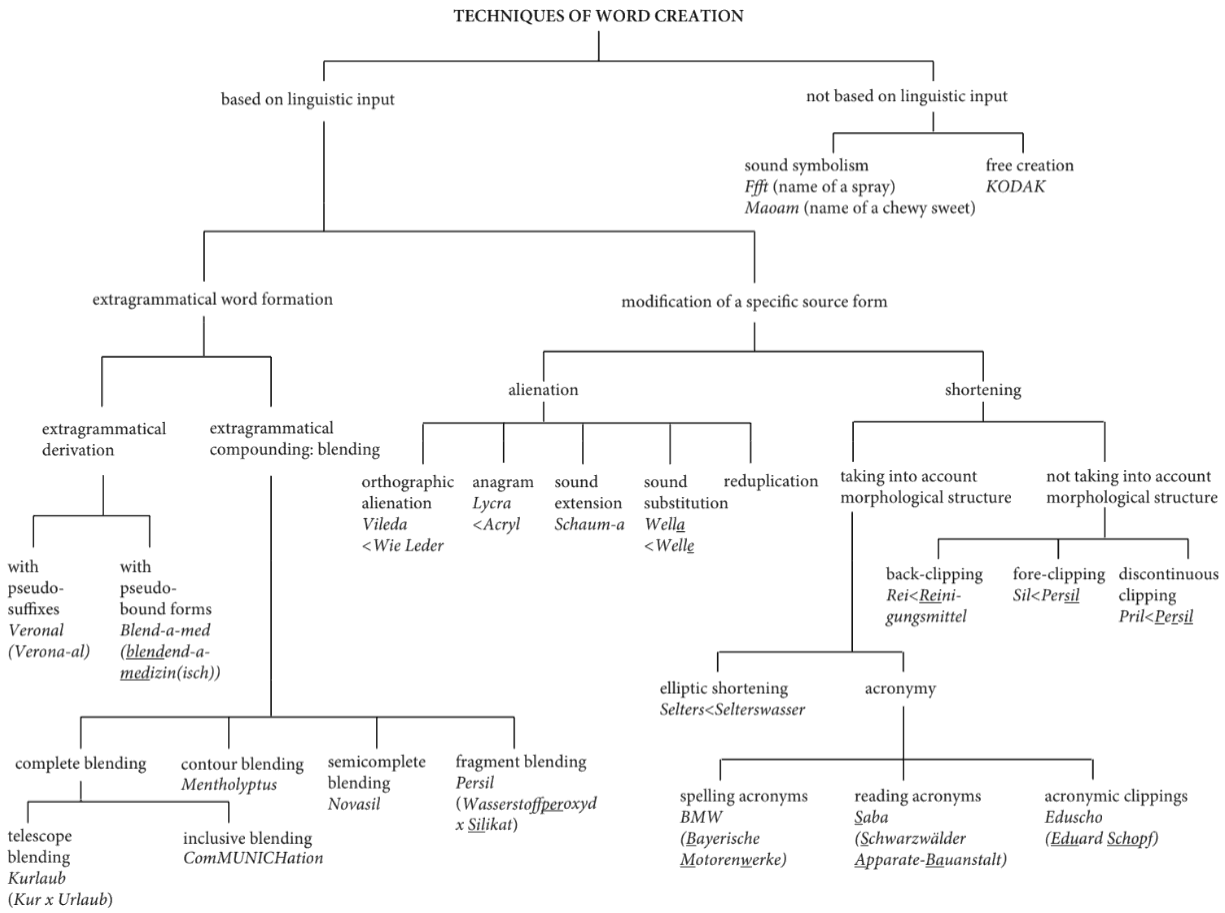
The challenge of terminology creation, particularly in technical domains, is a significant issue for languages lacking standardization, and with limited resources. Toury (1985) highlighted the inherent difficulties these languages face in producing the diverse range of discourse types found in more dominant languages, stating that "*Minority languages are weak by definition, not only vis-à-vis the majority language of the Community, but also — due to the inherent difficulties of producing in them the entire range of discourse types — with regard to most of the languages that are possible candidates to serve as source languages for translating into them.*" This weakness is acutely felt in the context of scientific and technical discourse, where specialized vocabulary is essential.

This problem is particularly evident in Moroccan Darija, as demonstrated by Sedrati and Ait Ali (2019), and is further corroborated by observations within several online communities such as Wikipedia. Even in bigger languages such as Modern Standard Arabic, the lack of standardized technical terminology has led to conflicts and inconsistencies: "*The scientific content in Arabic Wikipedia is considered weak compared to other knowledge fields. This includes Engineering, Medicine and IT. One of the problems that faces the scientific content is terminology. In many countries, science is taught in English in universities, and students find it difficult to accept alternatives for the terminologies they know, so they depend on the English articles in this case. Trying to provide the content in Arabic is also difficult and faces a lot of edit-wars because some users prefer to use Arabized terms even if they are non-popular or difficult to use.*" (Salman 2015).

This lack of readily available terminology sources has hindered knowledge sharing, often forcing contributors to either engage in extensive searches for appropriate terms, thus wasting valuable volunteering time, or resort to creating neologisms, which is a practice generally prohibited by Wikipedia's original research policy (Millosh 2015).

These conflicts and challenges highlight the urgent need for a systematic and user-friendly approach to terminology creation in languages lacking them.

While numerous methodologies for word creation exist, as detailed by authors such as Algeo (1978), Cannon (1986 & 1989), Grésillon (1984), Kobler-Trill (1994), Soudek (1978), and Steinhauer (2000), these techniques are often insufficient to address the **specific needs of low-resource languages**. The existing methods do not fully account for the **lack of standardized orthography, limited lexicographical resources, and the need for community involvement in the terminology creation process**, which is a core value for Wikimedians. In the case of Moroccan Arabic, works by Jan Hoogland (2007, 2008, 2014) provide valuable insights into the challenges of lexical gaps, bilingual dictionaries, and orthography, further emphasizing the specific linguistic context.



Techniques of word creation (Ronneberger-Sibold 2010)

Recent work has also explored the use of Wikidata lexemes as a resource for lexical data, as seen in the work of Macfarlane (2016), Morshed (2024), Magwenzi (2023), and Cartoni et al. (2020). These studies demonstrate the potential of Wikidata for supporting lexical representation and generation.

Nielsen (2019, 2023) has investigated the use of Wikidata lexemes in the context of Danish, a language with more resources than Darija, highlighting the potential of this approach across different language contexts.

Finally, research by Takrour (2015) and Tachicart & Bouzoubaa (2022, 2025) has examined vocabulary generation in Moroccan Arabic, including morphological naturalization and rule-based approaches. However, these approaches **do not fully address** the challenge of creating new terminology in a **collaborative and accessible way for non-experts**, specifically in the context of Wikimedian communities who need to rely on themselves to create new words.

Gap and Advancement

This project aims to advance research and scientific understanding by addressing a critical gap: The lack of a practical, accessible methodology for generating new terminology in low-resource, non-standardized languages such as Moroccan Darija. While previous work has explored general word creation techniques, lexical resources like Wikidata, and specific challenges in Darija, our research will

uniquely combine these elements to develop a methodology specifically tailored to the needs of Wikimedian communities. By leveraging Wikidata lexemes and actively involving editors in the terminology creation process, our methodology will provide a novel and much-needed solution, empowering native speakers to contribute to the growth and standardization of their language in technical domains, while also enriching their wikis.

Methods

Our research employs a mixed-methods approach to develop and evaluate a methodology for generating new terminology in low-resource unstandardized languages, taking Moroccan Darija as a use case. Here below is a description of the different activities and processes that will take place during our project implementation.

Data Collection

The project will gather a comprehensive dataset in Moroccan Darija from a variety of sources. This process involves multiple stages:

1. Source Identification - We will identify and prioritize high-quality sources, both online and in paper format. Online sources will encompass digital dictionaries (including Wiktionaries) and other reputable online content in Moroccan Darija. Offline sources will include scanned versions of relevant dictionaries and linguistic texts.
2. Data Gathering - Data will be collected using two main techniques. For physical documents, Optical Character Recognition (OCR) software will be employed to extract text. For online sources, web-scraping techniques will be utilized to systematically gather lexical data.
3. Data Preparation - The raw data obtained will undergo a rigorous preparation and cleanup process. This will involve correcting errors introduced by OCR, standardizing the data format, and organizing the words and their associated information (definitions, etymologies, etc.) into a format compatible with Wikidata. We will also establish the relationships between these words to capture their semantic and morphological connections.
4. Wikidata Integration - The prepared data will be uploaded to Wikidata using *QuickStatements* and *OpenRefine*, which are well established tools for efficient data import and management for Wikidata. This step ensures that the generated lexical data is structured, accessible, and interoperable.

Data Analysis and Methodology Development

The core of this research involves analyzing the collected lexical data to identify patterns in terminology generation and developing a robust methodology. This will be conducted in collaboration with a linguistics researcher, specialized in our use case language, Moroccan Darija. The steps that will be followed in this part are:

1. Pattern Extraction - We will analyze the data to extract existing patterns of terminology generation in Moroccan Darija, considering various linguistic factors such as morphology, syntax, and semantics. This analysis will examine how new terms are formed, adapted from other languages, and used in different contexts.

2. Methodology Creation - Based on the extracted patterns, we will develop general guidelines and methods for generating new words. This methodology will be designed to be accessible and applicable by individuals without specialized linguistic expertise.
3. Methodology Application - The developed approach will be applied to a chosen list of terms. This list will be drawn from a set of known concepts (e.g., a subset of the list of articles every Wikipedia should have) to ensure that the generated terminology is relevant and widely applicable.

Community Survey

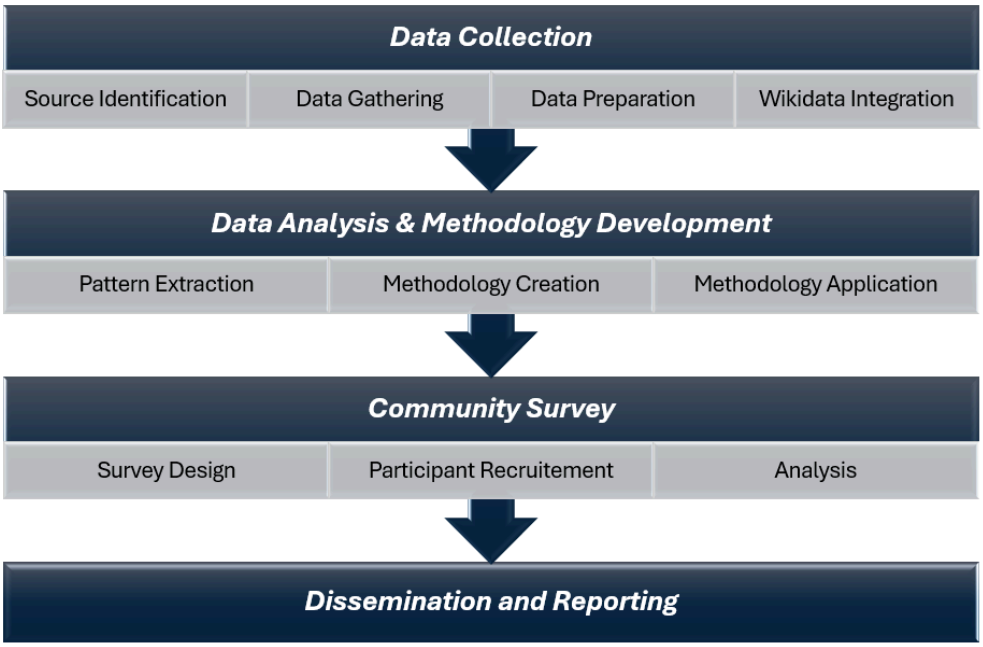
To assess the acceptance and usability of the newly generated terminology, a survey will be conducted among native speakers of Moroccan Darija, both within and outside the Wikipedia movement. The aim of this survey will be to gather feedback on these new words. The activities involved in this part are:

1. Survey Design - The survey will present a set of approximately 100 words generated using our developed methodology. Participants will be asked to assess the terms for clarity, accuracy, naturalness, and appropriateness for use in technical articles.
2. Participant Recruitment - The survey will be shared with Moroccan Darija speakers, prioritizing Wikimedians, through various channels, including Wikis, social media, and relevant community networks, to ensure a diverse range of perspectives. We will actively engage with Wikipedia editors and other interested stakeholders to maximize participation.
3. Analysis - The survey results will be analyzed to evaluate the effectiveness of the terminology generation methodology. We will measure the level of agreement among respondents regarding the acceptability of the generated terms and identify any patterns or areas for improvement.

Dissemination and Reporting

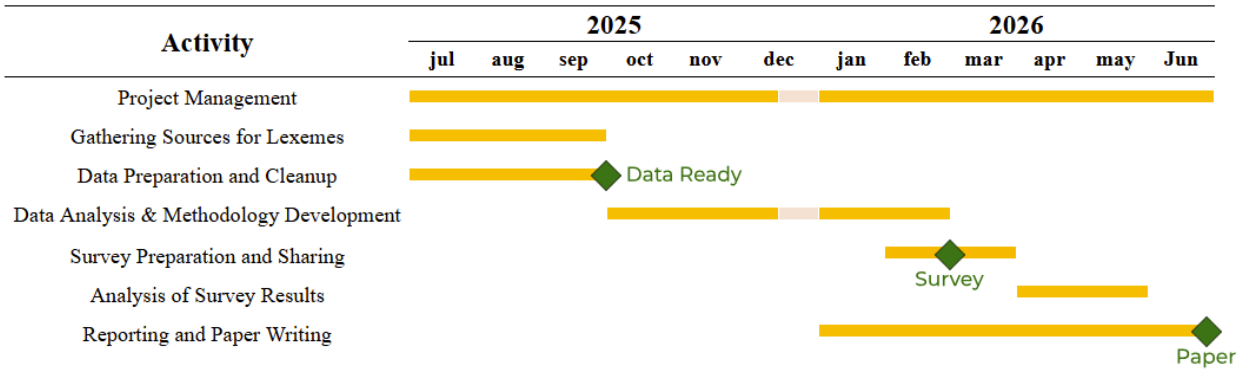
The findings of this research, including the developed methodology and its evaluation, will be disseminated through a project report, conference presentations, and an academic publication. We will also explore ways to integrate the methodology and its results into Wikimedia resources and tools to facilitate its adoption by the Wikimedian communities.

A visual representation of our approach is provided below, outlining the process flow and the input-output relationships between individual activities.



Project Workflow

The timeline below summarizes the planned activities to be performed during the project implementation phase (July 2025 - June 2026).



Project Timeline

Expected output

By its completion, this project will deliver several research outputs, all aiming to support the development of low-resource unstandardized languages, and improve Wikimedia content. The list of key outputs is presented in the table below.

Output	Description	Audience	Benefit
Datasets	Curated collection of compiled datasets containing Moroccan Darija lexical data.	• Linguists • Computational linguists	Standardized, reliable data source for linguistic analysis.

		• Language researchers	
Scripts	Software scripts developed for data collection and processing	• Developers • Researchers • Wikimedians	Available resources to replicate for other Wikimedia communities and projects.
Enriched Darija Wikidata Lexemes	Expanded set of structured and interconnected Moroccan Darija lexemes in Wikidata.	• Wikidata Community • Darija Community	More effective language research, computational applications, and online presence for Darija.
Terminology Generation Methodology	Well-documented and user-friendly methodology for generating new technical vocabulary in low-resource, unstandardized languages	• Wikimedia Communities • Language Activists	Empowering communities to create and standardize their own technical vocabulary in their native languages.
Scientific Publication	Peer-reviewed academic publication detailing the research process, findings, and our developed methodology.	• Scientific community interested in Wikidata and linguistics.	Advancing knowledge in computational linguistics, lexicography, and language revitalization areas.
Project Report	Report summarizing the research activities and results	• WMF • Wikimedia Communities • General Public	Knowledge sharing and dissemination of the project work, its outcomes and impact

Risks

We have summarized the different risks that we anticipate for this project in the table below, gathering a definition for each risk, its impact, probability, and how we intend to mitigate it.

Risk	Impact	Probability	Mitigation
Datasets are too large or resources are scarce (due to limited access or availability)	High	Medium	A sampling approach will be used to maintain the representativeness of the data. We will also allocate contingency time for data collection.
Loss of key member	High	Low	document all processes and scripts, to enable an efficient handover if needed. Prepare contingency plans for recruitment and replacement.
Low community engagement in the survey or the project in general, leading to insufficient feedback for evaluating the generated	Medium	High	Reaching out to diverse people since the start of the project, including

terminology.			interested scholars and universities working in the field. Include incentives to participate.
OCR not working in Moroccan Darija	Medium	High	Adjusting texts manually by working first on OCR in Standard Arabic and correcting to Darija later.
Difficulty in extracting clear and consistent patterns of terminology generation from the data, potentially due to linguistic variations or data sparsity.	Medium	Medium	Involve an expert linguist with expertise in Moroccan Darija, to explore different linguistic analysis techniques, and refine the scope of the analysis if necessary.
The developed methodology may not be easily applicable or understandable by Wikimedians	Low	Medium	Involve target users in the development process, and be ready to iterate if needed.
Delay in project deliverables	Low	Medium	Develop a detailed project timeline, track progress, and have regular meetings.

Community impact plan

This project is deeply committed to engaging with and impacting audiences beyond academia, particularly the Wikimedia communities. To promote the adoption and use of the research outcomes, the project will actively involve Wikimedia volunteer editors and organizers throughout its lifecycle. This will include collaborating with these communities in the design and execution of the community survey, ensuring that the generated terminology is relevant and user-friendly. We will also disseminate the project's findings through Wikimedia channels, such as project pages and community talk pages, and explore opportunities to integrate the developed methodology and resources into Wikimedia tools and workflows.

By prioritizing community involvement and tailoring the project's outputs to meet the practical needs of Wikimedians, we aim to empower the different Wikimedia linguistic communities to create richer, more accurate, and more consistent technical content, fostering greater participation and knowledge equity within the movement.

Evaluation

This project's success will be evaluated using several measures, focusing on both the project's outputs and its impact on the Wikimedia community. These measures are:

- Produce a well-documented, user-friendly methodology for generating new technical vocabulary, that will be made available to Wikimedia communities.
- Achieve a significant increase in the number of Moroccan Darija lexemes in Wikidata (currently there are only two), contributing to a more comprehensive and structured linguistic resource.
- Generate a set of at least 100 new technical terms in Moroccan Darija, and use it to test the developed methodology.
- At least 50 people (Wikimedians and other interested stakeholders) will participate in the community survey, providing feedback on the generated terminology.
- Publish a peer-reviewed academic publication detailing the research process, findings, and our developed methodology.

Budget

The total budget requested for this application is 242,000 MAD (corresponding to 25,722 USD, by conversion date April 16, 2025). The main items in this budget are personnel support (54 % of total), dissemination activities (12 % of total), and travel for conference participation (21 % of total). Kiwix will be the fiscal sponsor receiving and operating the budget on behalf of the team. Kiwix is charging 10% of the total amount as a fee.

The budget spreadsheet for this application can be found at this link.

References

- Algeo, J. (1978). “The Taxonomy of Word Making” , Word 29, 122-31.
- Amano, T., González-Varo, J. P., & Sutherland, W. J. (2016). “Languages are still a major barrier to global science”. PLoS biology, 14(12), e2000933.
- Cannon, G. (1986). “Blends in English word formation”, Linguistics 24, 725-753.
- Cannon, G. (1989). “Abbreviations and Acronyms in English Word-Formation” , American Speech 64. 2, 99-127.
- Cartoni, B., et al. (2020). “Introducing Lexical Masks: a New Representation of Lexical Entries for Better Evaluation and Exchange of Lexicons”. In Proceedings of the Twelfth Language Resources and Evaluation Conference, 3046-3052.
- Grésillon, A. (1984). “La règle et le monstre: le mot-valise. Interrogations sur la langue, à partir d’un corpus de Heinrich Heine”. Tübingen: Niemeyer.
- Hoogland, J. (2007). “Lexical gaps in Arabic: Evidence from dictionaries”. In Approaches to Arabic Linguistics, 455-473. Brill.
- Hoogland, J. (2008). “Lexicography: bilingual dictionaries”. Encyclopedia of Arabic Language and linguistics, 3, 21-30.
- Hoogland, J. (2014). “Towards a standardized orthography of Moroccan Arabic based on best practices and common ground among a selection of authors”. In Árabe marroquí: de la oralidad a la enseñanza, 59-76.
- Kobler-Trill, D. (1994). “Das Kurzwort im Deutschen . Eine Untersuchung zu Definition, Typologie und Entwicklung”. Tübingen: Niemeyer.
- Macfarlane, D. (2016). “The lexeme hypotheses: Their use to generate highly grammatical and completely computerized medical records”. Medical hypotheses, 92, 75-79.
- Magwenzi, T. (2023). “The Development of lexicographical databases, tools, and resources for storing Multilingual Data in Support of the Abstract Wikipedia Project: A Literature Review”.
- Marginson, S., & Xu, X. (2023). “Hegemony and inequality in global science: Problems of the center-periphery model”. *Comparative Education Review*, *67*(1), 31-52.
- Meta-Wiki (2019). “Community Wishlist Survey 2019/Editing/Tool for easy science editing using linked dictionary”. Meta Wikimedia. https://w.wiki/DoKF
- Millosh (2015). “Developing new language editions of Wikipedia”. Meta Wikimedia. https://w.wiki/DoK9
- Morshed, M. (2024). “Using Wikidata lexemes and items to generate text from abstract representations”. Semantic Web, 15(6), 2319-2332.
- Nielsen, F. Å. (2019). “Danish in Wikidata lexemes”. In 10th Global WordNet Conference.
- Nielsen, F. Å. (2023). “Alignment of Wikidata lexemes and Det Centrale Ordregister”. In 24th Nordic Conference on Computational Linguistics.
- Nigatu, H. H., Canny, J., & Chasins, S. E. (2024). “Low-Resourced Languages and Online Knowledge Repositories: A Need-Finding Study”. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1-21.
- Ortega, R. P. (2020). “Science’s English dominance hinders diversity, but the community can work toward change”. *Science*.
- Ronneberger-Sibold, E. (2010). “Word creation”. In Variation and Change in Morphology: Selected papers from the 13th International Morphology Meeting, Vienna, February 2008 (Vol. 310). John Benjamins Publishing.

- Salman, M (2015). “Dilemma of Scientific Terms in WikiArabic”. Meta Wikimedia. https://w.wiki/DoK5
- Sedrati, A., & Ait Ali, A. (2019). “Moroccan Darija In Online Creation Communities: Example Of Wikipedia”. Al-Andalus Magreb, 26(1), 2660-7697.
- Smirnov, I., Oprea, C., & Strohmaier, M. (2023). “Toxic comments are associated with reduced activity of volunteer editors on Wikipedia”. PNAS nexus, 2(12), pgad385.
- Soudek, L. I. (1978). “The relation of blending to English word-formation: theory, structure, and typological attempts”, in: Wolfgang U. Dressler & Wolfgang Meid (eds.), Proceedings of the Twelfth International Congress of Linguists. Innsbruck: Institut für Sprachwissenschaft, 462-466.
- Steinhauer, A. (2000). “Sprachökonomie durch Kurzwörter. Bildung und Verwendung in der Fachkommunikation”. Tübingen: Narr.
- Tachicart, R., & Bouzoubaa, K. (2022). “Moroccan Arabic vocabulary generation using a rule-based approach”. Journal of King Saud University-Computer and Information Sciences, 34(10), 8538-8548.
- Tachicart, R., Bouzoubaa, K., & Namly, D. (2025). “Effective Techniques in Lexicon Creation: Moroccan Arabic Focus”. In AI-Driven: Social Media Analytics and Cybersecurity, 235-249. Cham: Springer Nature Switzerland
- Takrour, H. (2015). “Contact de langues au Maroc: Naturalisation morphologique des mots d’emprunt dans l’arabe marocain−Cas du vocabulaire de la mer”. Langues, cultures et sociétés, 1(1), 151-170.
- Toury, G. (1985). “Aspects of translating into minority languages from the point of view of translation studies”.