الذكوات البيض

اسم مشتق من الذكوة وهي الجمرة الملتهبة والمراد بالذكوات الربوات البيض الصغيرة المحيطة بمقام أمير المؤمنين علي بن أبي طالب{عليه السلام}

شبهها لضيائها وتوهجها عند شروق الشمس عليها لما فيها

موضع قبر علي بن أبي طالب {عليه السلام}

من الدراري المضيئة

{در النجف}فكأنها جمرات ملتهبة وهي المرتفع من الأرض، وهي ثلاثة مرتفعات صغيرة نتوءات بارزة في أرض الغري وقد سميت الغري باسمها، وكلمة بيض لبروزهاعن الأرض.وفي رواية إنّها موضع خلوته أو إنّها موضع عبادته وفي رواية أخرى في رواية المفضل عن الإمام الصادق {عليه السلام}قال: قلت:يا سيدي فأين يكون دار المهدي ومجمع المؤمنين؟ قال:يكون ملكه بالكوفة، ومجلس حكمه جامعها وبيت ماله ومقسم غنائم المسلمين مسجد السهلة وموضع خلوته الذكوات البيض

Republic of Iraq
Ministry of Higher Education &
Scientific Research
Research & Development
Department

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
دائرة البحث والتطوير

No.:
Date:

الرقم: ب ت ع/ ٥٠٤٩
التاريخ: ١٤/٨/٢٠٢٢

ديوان الوقف الشيعي / دائرة البحوث والدراسات

م/ مجلة الذكوات البيض

السلام عليكم ورحمة الله وبركاته ...

إشارة الى كتابكم المرقم ١٠٤٦ والمؤرخ ١٢/٢٨ /٢٠٢١ والحاقاً بكتابنا المرقم ب ت ٥٧٤٤/٤ في ٢٠٢١/٩/٦ ، والمتضمن أستحداث مجلتكم التي تصدر عن الوقف المذكورة أعلاه ، وبعد الحصول على الرقم المعياري الدولي المطبوع وانشاء موقع الكتروني للمجلة تعتبر الموافقة الواردة في كتابنا أعلاه موافقة نهائية على أستحداث المجلة.

... مع وافر التقدير

أ.م.د حسين صالح حسن
المدير العام لدائرة البحث والتطوير / وكالة
٢٠٢٢/١/١٢

نسخة منه الى:
- قسم الشؤون العلمية / شعبة التأليف والنشر والترجمة / مع الأوراق.
- الصادرة.

مهند ابراهيم
١٠ / كانون الثاني

وزارة التعليم العالي والبحث العلمي – دائرة البحث والتطوير – القصر الأبيض – المجمع الثربوي – الطابق السادس

إشارة إلى كتاب وزارة التعليم العالي والبحث العلمي / دائرة البحث والتطوير
المرقم ٥٠٤٩ في ٢٠٢٢/٨/١٤ المعطوف على إعمامهم
المرقم ١٨٨٧ في ٢٠١٧/٣/٦
تُعدّ مجلة الذكوات البيض مجلة علمية رصينة ومعتمدة للترقيات العلمية.

الذكوات البيض

مجلة علمية فكرية فصلية محكمة تصدر عن
دائرة البحوث والدراسات في ديوان الوقف الشيعي

العدد (١٧) السنة الثالثة جمادى الآخرة ١٤٤٦ هـ تشرين الأول ٢٠٢٥ م
رقم الإيداع في دار الكتب والو ثائق (١١٢٥)
الرقم المعياري الدولي ISSN 2786-1763

الذكوات البيض

العدد (١٧) السنة الثالثة جمادى الآخرة ١٤٤٦ هـ – تشرين الأول ٢٠٢٥ م

الذكوات البيض

مجلة علمية فكرية فصلية محكمة تصدر عن
دائرة البحوث والدراسات في ديوان الوقف الشيعي

العدد (١٧) السنة الثالثة جمادى الآخرة ١٤٤٦هـ - تشرين الأول ٢٠٢٥م

العنوان الموقعي
مجلة الذكوات البيض
جمهورية العراق
بغداد /باب المعظم
مقابل وزارة الصحة
دائرة البحوث والدراسات
الاتصالات
مدير التحرير
٠٧٧٣٩١٨٣٧٦١
صندوق البريد / ٣٣٠٠١
الرقم المعياري الدولي
ISSN ١٧٦٣-٢٧٨٦
رقم الإيداع
في دار الكتب والوثائق(١١٢٥)
لسنة ٢٠٢١
البريد الالكتروني
إيميل
off reserch@sed.gov.iq
hus65in@gmail.com

دليل المؤلف .....

١–أن يتسم البحث بالأصالة والجدّة والقيمة العلمية والمعرفية الكبيرة وسلامة اللغة ودقة التوثيق.

٢– أن تحتوي الصفحة الأولى من البحث على:

أ. عنوان البحث باللغة العربية .

ب . اسم الباحث باللغة العربي، ودرجته العلمية وشهادته.

ت . بريد الباحث الإلكتروني.

ث . ملخصان: أحدهما باللغةِ العربية والآخر باللغةِ الإنكليزية.

ج . تدرج مفاتيح الكلمات باللغة العربية بعد الملخص العربي.

٣–أن يكونَ مطبوعًا على الحاسوب بنظام(office Word) ٢٠٠٧ أو ٢٠١٠) وعلى قرص ليزري مدمج (CD) على شكل ملف واحد فقط (أي لا يُجزّأ البحث بأكثر من ملف على القرص) وتُزوَّد هيأة التحرير بثلاث نسخ ورقية وتوضع الرسوم أو الأشكال، إن وُجدت، في مكانِها منَ البحثِ، على أن تكونَ صالحةً مِنَ الناحيةِ الفنيَّة للطباعة.

٤–أن لا يزيدَ عدد صفحات البحث على (٢٥) خمس وعشرين صفحة من الحجم (A4).

٥. يلتزم الباحث في ترتيب وتنسيق المصادر على الصفية APA

٦–أن يلتزم الباحث بدفعِ أُجُور النشر المحدَّدة البالغة (٧٥،٠٠٠) خمسة وسبعين ألف دينار عراقيّ، أو ما يعادلها بالعملات الأجنبية.

٧–أن يكونَ البحثُ خاليًا مِنَ الأخطاءِ اللغوية والنحوية والإملائيَّة.

٨–أن يلتزم الباحث بالخطوط وأحجامِها على النحو الآتي:

أ. اللغة العربية: نوع الخط (Arabic Simplified) وحجم الخط (١٤) للمتن.

ب . اللغة الإنكليزية: نوع الخط (Times New Roman) عناوين البحث (١٦). والملخصات (١٢)

أما فقرات البحث الأخرى؛ فبحجم (١٤) .

٩–أن تكونَ هوامش البحثِ بالنظام الأكتروني(تعليقات ختامية) في نهاية البحث. بحجم ١٢.

١٠–تكون مسافة الحواشي الجانبية (٢,٥٤) سم،والمسافة بين الأسطر (١) .

١١–في حال استعمال برنامج مصحف المدينة للآيات القرآنية يتحمل الباحث ظهور هذه الآيات المباركة بالشكل الصحيح من عدمه، لذا يفضّل النسخ من المصحف الالكتروني المتوافر على شبكة الانترنيت.

١٢–يبلَّغ الباحث بقرارِ صلاحيَّة النشر أو عدمها في مدَّةٍ لا تتجاوز شهرين من تاريخ وصولهِ إلى هيأةِ التحرير.

١٣–يلتزمُ الباحث بإجراءِ تعديلات المحكّمين على بحثهِ وفق التقارير المرسلة إليهِ وموافاةِ المجلة بنسخةٍ معدَّلةٍ في مدَّةٍ لا تتجاوزُ (١٥) خمسة عشر يومًا.

١٤–لا يحق للباحث المطالبة بمتطلبات البحث كافة بعد مرور سنة من تاريخ النشر.

١٥–لاتعاد البحوث الى أصحابها سواء قُبلت أم لم تُقبل.

١٦–تكون مصادر البحث وهوامشه في نهاية البحث، مع كتابة معلومات المصدر عندما يرد لأول مرة.

١٧–يخضع البحث للتقويم السري من ثلاثة خبراء لبيان صلاحيته للنشر.

١٨–يشترط على طلبة الدراسات العليا فضلاً عن الشروط السابقة جلب ما يثبت موافقة الأستاذ المشرف على البحث وفق النموذج المعتمد في المجلة.

١٩–يحصل الباحث على مستل واحد لبحثه، ونسخة من المجلة، واذا رغب في الحصول على نسخة أخرى فعليه شراؤها بسعر (١٥) ألف دينار.

٢٠–تعبر الأبحاث المنشورة في المجلة عن آراء أصحابها لا عن رأي المجلة.

٢١–ترسل البحوث إلى مقر المجلة – دائرة البحوث والدراسات في ديوان الوقف الشيعي بغداد – باب المعظم)

أو البريد الإلكتروني:(hus65in@Gmail.com) (off reserch@sed.gov.iq) بعد دفع الأجور في مقر المجلة

٢٢–لا تلتزمُ المجلة بنشر البحوث التي تُخلُّ بشرطٍ من هذهِ الشروط .

مجلة علمية فكرية فصلية محكمة تصدر عن دائرة البحوث والدراسات في ديوان الوقف الشيعي

محتوى العدد (١٧) المجلد الثالث

فصلية مُحكمة تُعنى بالبحوث والدراسات العلمية والإنسانية والفكرية
العدد (١٧) السنة الثالثة جمادى الآخرة ١٤٤٦ هـ كانون الأول ٢٠٢٥م

The Intersection of Language and Artificial Intelligence: Implications for Linguistic Diversity

Lecturer Ahlam Abdulrazzaq Thiab Al-Dulaimi
Baghdad General Directorate of Al-Karkh

Abstract:
The field of Artificial Intelligence (AI) or close to it, Natural Language Processing (NLP), is literally transforming how a human being interacts with technology and information. This dissertation will spell out the profound implications of this technological revolution to the linguistic diversity in the world. On the one hand, the potential of AI to offer unequaled prospects to communication and information access is indisputable, on the other it gives rise to the dystopia of unequal proportions, and to a great part that of uneven language. This paper explores the dual nature of AI as both a very real danger to low-resource languages and as a possible solution to saving and promoting the latter. This work, developed on an analysis of the prevailing situation in the language technology sector, explores structural bias favouring a few rich and strong languages. It further identifies the inherent obstacles, such as the absence of data and the biased nature of algorithms, to the development of fair AI systems in the thousands of languages in the world. The study reveals that commercial and geopolitical forces of AI development are inclined to marginalize linguistically diverse societies. However, it also mentions new projects and technological plans that involve the use of AI to record, teach, and preserve dying languages. Given the study, it is concluded with a set of recommendations that may be applied by technologists, policymakers, and linguistic communities to create a more open and linguistically diverse digital future. The main idea here is that it requires a proactive, intentional process of guiding the development of AI, so that it is beneficial and not detrimental to the divergent world of human languages.

Keywords: Artificial Intelligence, Natural Language Processing, Linguistic Diversity, Endangered Languages, Digital Language Divide, Algorithmic Bias, Language Technology.

1. Introduction

1.1. Background of the Study

The 21st century can be defined as the period when the idea of Artificial Intelligence (AI) is evolving into everyday life swiftly. Language technologies based on AI are everywhere: there are virtual assistants on mobile phones or advanced translation options. It is a technological revolution facilitated by the progress in Natural Language Processing (NLP), which is one of the areas of AI that aim to make computers able to read, comprehend, and write human language (Jurafsky and Martin, 2023).

These tools can rip the communication barriers and will democratize the access to information globally. But there are an estimated 7,000 languages of the world, and each of these is viewed as a distinct symbol of a cultural and cognitive system (Crystal, 2000). Most of these languages are extremely not covered by the current wave of AI technologies.

Automation of NLP models requires vast amounts of electronic text and speech data, which is easily obtained in economically and politically significant languages such as Spanish, Mandarin Chinese, and English. These populations that live in smaller numbers are linguistic communities whose languages are in low resources and cannot be represented digitally to be shown to train efficient AI models. Such contrast is one of the critical nexuses of technological advancement and the ancient linguistic diversity issue.

1.2. Statement of the Problem

The fundamental question that this study answers is that the supporting concept of AI across all the languages in the world and its subsequent impacts on linguistic diversity are unequal. The trend which AI research is moving toward can be seen to be creating a two-tiered world; the one, full of the privileges of AI, is those that speak high-resource languages, and the other, which are left behind, are those who speak low-resource languages. Failing to include them in online sites may cause their languages to be undermined, become less popular in the online world, accelerate language shift and endangerment (Bird, 2020). The devastating impact of AI on the linguistic diversity is immense; nevertheless, its possibilities as the means of preserving and reviving have not been utilized to the full extent. In this way, there should be a critical overview to enable the evaluation of the risks and opportunities offered by AI.

1.3. Research Questions

The key research questions to be answered in this dissertation are the following:

1. What roles do the present stages of AI and NLP technology development play in creating a digital language divide?
2. What are the main problems (e.g. lack of data, bias in the algorithms) with developing AI in low-resource languages?
3. So what is the way to use AI as a means to aid the documentation,

preservation and resilience of endangered languages?

1.4. Significance of the Study

The proposed study is timely and critical to conduct a study on a fast-growing issue. It is important because it is an interdisciplinary field, and it straddles computer science, linguistics, and sociology. It is useful to the literary discussion in that it analyzes the structuring elements behind linguistic inequality in the digital era in a systematically organized way. Moreover, this research can provide some useful suggestions to such stakeholders as policy makers, technology experts and language communities. It will support the development of more balanced and inclusive linguistic technologies that do not harm but afford the linguistic heritage of the world by reminding both the dangers and the potential opportunities.

1.5. Aims of the Study

The main aim of the study is to develop the in-depth study of the intricate connection between the development of AI and the condition of the linguistic diversity on the planet. The targeted objectives are the following ones:

1. To question the extent to which the latest direction of AI development, especially of Natural Language Processing (NLP), enhances the digital language divide and creates a threat to the linguistic diversity.
2. To determine and discuss the most significant technological and socioeconomic obstacles, including data scarcity, algorithmic bias, and market-driven incentives, that result in the creation of the biased AI in the case of low-resource and endangered languages.
3. To understand the opportunities and challenges of AI as a helpful solution to the documentation, preservation, and renewal of all the languages in the world, to emphasize the cases of successful practice and new approaches.
4. To develop evidence-based policy advice to be offered to policy-makers, the developers of technologies, and linguistic communities to manage the field of AI creation in a more inclusive and linguistically equitable direction.

1.6. Hypotheses

The current study would be based on the subsequent hypotheses and examined in terms of critical analysis of the available literature and case studies:

H1: The functional depreciation and marginalization of low-resource languages in the digital space as a result of AI development being concentrated on a subset of high-resource languages will hasten the process of language shift.

H2: Data scarcity: Data scarcity is the dominant technical challenge in creating effective NLP technologies in the vast majority of languages of the world and introduces a systemic bias that can only be adopted by the current mainstream AI models with specific intervention.

H3: Community-based methods of introducing AI into language revitalization heavily rely on the ability of community participants to focus on the particularities of the needs and the cultural background of the speakers, and not on implementing generalized and commercially-oriented AI solutions.

2. Literature Review

2.1. The Foundation of Natural Language Processing (NLP)

The presence of Natural Language Processing is a subject that borders on both computer science and AI as well as linguistics. It is focused on developing computational systems that are able to process and comprehend human language (Chowdhary, 2020).

In early NLP systems were rule based, with grammar rules being written by linguists and programmers manually, to be followed by a computer. This was a fragile method that was not able to cope with the enormous complexity and ambiguity of the human language (Manning and Schutz 1999). Machine learning and, more recently, deep learning dominate the modern day of NLP.

These systems do not have to be explicitly programmed but rather learn quite directly and without needing significant amounts of text and speech data. Models such as Google›s BERT or Open AI›s GPT (Generative Pre-Trained Transformer) are built on the transformer architecture, which so far has shown great success in such tasks as translation, summarization, and question answering (Vaswani et al., 2017, as cited in Jurafsky and Martin, 2023). The quality and amount of data on which the models are trained is however directly proportional to their effectiveness. This data-hungry character is the key cause of the variation in the performance of the various languages.

2.2. Defining Linguistic Diversity and Its Importance

Linguistic diversity is the diversity of languages used in the world. This

diversity does not only have to do with the variety of vocabularies or grammatical rules; every language is a particular manner of thinking and structuring the world (Harrison, 2007). Culture, history and traditional knowledge have been transmitted by languages.

Language loss in terms of David Crystal (2000) is a loss of human intellectual heritage like the extinction of a species in biodiversity. There is a danger on this diversity though. According to the reports by United Nations, more than half of the world languages are facing the risk of extinction this century. It is a process that is justified by many forces, such as globalization, economic requirements, and integration, that is tempting speakers of their native languages to adopt stronger alternatives. These pressures are now being applied within the cyber space which is a new and powerful medium.

2.3. The Rise of the Digital Language Divide

The digital divide is the term that was initially used to denote the difference in access of the net and computer technology between various socioeconomic groups. It has gone further to the language aspect as well. Digital language divide: This is the gap between the languages that are broadly accessible to the net and the other languages that are not so prevalent (Pimienta, 2007). The web has been turning increasingly multilingual, but during its long-standing history, English has occupied the preponderant position in the internet with a small number of languages continuing to represent the large percentage of the digital content on the internet. This rift is currently being copied and even multiplied in the age of AI.

The AI systems are trained based on the existing online content. When a language is not popular on the Internet, there is no information to train the AI. It uses a vicious circle in that, it is more difficult that speakers can produce digital content in their language due to AI tools, such as keyboards, spell checkers, and voice recognitions and thus there is less data to produce further advanced AI tools (Nettle, 1999).

2.4. AI as a Threat: Homogenization and Bias

Research and development of AI has a direct threat on linguistic diversity because it is focused on a few languages. It strengthens the superiority of dominant languages and they become the default digital language. This may establish a silent yet compelling pressure on minority language speakers to make a switch to speaking one of the

dominant languages as a way to avail of services, information, and financial opportunities in the Internet (Kornai, 2013).

Moreover, depending on the language or culture of the data used to train AI models, it fosters and enhances biases. As an example, the automatic translation system based on the culturally specific objects of a low-resource language will perform the nonsensical translation or produce the incorrect ones. It is a form of algorithmic bias, namely the output of the system is biased due to biased training data (O›Neil, 2016). Those AI systems, which fail to provide linguistic and cultural reality with which they are supposed to deal, are not merely providing a poor service but are also discriminating against them.

2.5. AI as Opportunity: Revival and Preservation

Also a robust language documentation and revitalization tool, AI is a massive language danger in spite of the enormous monstrousities. NLP could also be used to assist the handling and analysis of historical language documents in the archives such that linguists are able to create under-documented language dictionaries and grammars more effectively. As an example, speech recognition can be integrated with AI to help in transcribing oral history and conversation, typically the only extant document of a dying language (Bird, 2020).

Moreover, AI has the potential to be employed in the creation of interactive learning tools to acquire a language. The chat-bots, learning games, and applications may be created to educate the new generations of speakers in their native languages. They would be able to provide personalized learning opportunities and help in learning languages particularly to diaspora populations that may be physically distanced to the rest of language speakers (Abney & Bird, 2010). The remedy is that the agenda of AI research should be transformed to cease operating with the aim of commercialization but rather operate on the suggestion of community efforts that can satisfy the requirements of the language revival researches in particular.

3. Methodology

3.1. Research Approach

This dissertation adopts the qualitative research approach using a critical literature and case study meta-analysis. It is a suitable methodology since research should be aimed to synthesize and interpret the information of various sources and come to a broader understanding of a

complex, multi-faceted phenomenon. A quantitative process would be restricted due to the unavailability of standardised and global data on AI support in all languages. Qualitative methodology can be used to better and more finely delve into the social, cultural, and technological forces at work. This study uses the critical theory perspective, which emphasizes power relations and inequality between development and implementation of technology (Fuchs, 2014).

3.2. Data Collection and Sources

Various academic and institutional sources were used as a source of data in this study. The primary sources include:

• Linguistic, Computer science and sociology: The theoretical background was set using books related to academic works on linguistics, computer science, and sociology.

• Scholarly Articles: The peer-reviewed articles presented specific information about the NLP models, language endangerment, and digital inequality.

• Technical Reports: The technical Reports of the technology companies, research labs and non-profit organizations provided us with the insight in the existing condition of language technology.

• Online Databases: The trends were discovered using publicly available information, such as the UNESCA Atlas of the World Languages in danger and the lists of NLP materials.

The choice of the sources was made with consideration to the principles of accessibility and reliability, and considering depth sources that could be found freely in order to present the scientific studies and guarantee the uprightness of the research.

3.3. Analytical Framework

The framework under the guiding the analysis of collected data concentrates on three major areas:

1. Resource Availability: It is the evaluation of the unequal access to digital content (text corpora, speech records, etc.), as well as to NLP services (e.g., parsers, translation systems, etc.), based on language.

2. Technological Performance: This involves the discussion of the output of the currently existing AI applications, including machine translation and voice assistant, on the high-resource and low-resource languages, in certain instances as a case study.

3. Socio-Technical Impact: This is to analyze the social and cultural

consequences of these technological differences such as the language shift potential, as well as the language revitalization opportunities.
In this context, the triangulation of information sources will help develop a complete image of the intersection between AI and linguistic diversity, which is the goal of the research.

4. Findings: The Current State of AI and Language

4.1. The Imbalance in Language Representation

The most remarkable observation is that there is a severe unequal distribution of AI resources among the linguistics around the world. Although there are more than 7,000 languages, a very small portion of them is well-supported by the modern NLP technology. A study also indicates that the world is determined by as few as 20 languages that are digitally available and on which most commercial AI applications are built (Joshi et al., 2020). It can be represented in the form of a pyramid.

Top languages include a few of the so-called high-resource languages such as English, Mandarin, and Spanish, where there is plenty of data and AI tools can be highly developed. Beneath this lie a layer of comparatively well supported «median resource» languages such as German, Japanese and French. The enormous bottom of the pyramid is comprised of thousands of low-resource languages which have minimal or no digital footprint and AI support.

Chart 1: The Pyramid of Language Resources for AI

This disequilibrium does not occur as a technical neutral process; it is the representation of the current geopolitical and economic balances of power. The economically dominating countries invest most in AI research and development of their languages.

4.2. Case Study: Machine Translation Services

Examples of the potential and the danger of AI in languageMachine translation (MT) services such as Google Translate and DeepL are both very strong indicators of both the promise and the kind of danger AI represents. In the case of high-resource pairs of languages (e.g. French-English), the translation quality can be exceedingly high and it can be of human-level quality in certain types of text (Bahdanau, Cho, and Bengio, 2014, as cited in Goodfellow, Bengio and Courville, 2016). Through these services, cross-cultural communication and access to information is made easy.

Their performance is however very low with low–resource languages. Any translation of these languages results in a nonsensical or grammatically incorrect translation. The reason is the fact that a lot of systems have English as a pivot language. An example of this is that to translate it out of Swahili into Welsh, the system can do Swahili to English and then to Welsh. Every step of this process creates errors thus leaving out a poor quality final translation.

Table 1: Quality of Machine Translation for Different Language Types

Language Pair	Resource Level	Typical Translation Quality	Example of Error
Spanish English	High	High	Very high; fluent and accurate for most common domains.	Minor stylistic awkwardness.
Hungarian Japanese	Medium	Medium	Good, but with noticeable grammatical errors.	Incorrect verb conjugation or word order.
Yoruba Icelandic	Low	Low	Very poor; often unintelligible or misleading.	«The cat sat on the mat» becomes «Automobile is sleeping in the house sky.»
Kʼicheʼ (Mayan) English	Endangered	High	Extremely poor or completely unavailable.	System fails to recognize the language or produces random words.

This language imbalance ensures that these low–resource language speakers could not depend on these tools to communicate, conduct business or even obtain education which further discriminates them in the globalized world.

4.3. Case Study: Voice Assistants and Speech Recognition

Voice assistants such as those of Siri (Apple), Alexa (Amazon), and Google Assistant have become a major interface between people and computers. They make use of Automatic Speech Recognition (ASR) in interpreting spoken speech and Text–to–Speech (TTS) in producing speech.

Coming up with ASR/TTS systems is even more of a data intensive task than the text based NLP. It needs thousands of hours of audio recording transcriptions of various speakers. Naturally, such systems can only be offered with a minor portion of the languages across the globe. Even the most common voice assistants do not support over 100 languages as of 2025 and their presence can be even considerably poorer in a situation when talking to a child with non–standard ac–

cent, even with the high-resource language (Koenecke et al., 2020, as cited in a technical report).

The implication of this voice divide is grave. With the increasing voice-activated nature of technology (in vehicles, homes, and other community services), the speakers of unrecognizable languages will not have an opportunity to do so. This filtering is especially acute in those languages that are mostly spoken and have a small written tradition because it may be a more natural voice mode of interaction than text-based interfaces.

4.4. The Challenge of Data Scarcity

Most of the languages in the world are not developed due to insufficient data to develop AI. Brute force State-of-the-art deep learning is a type of learning animal; they require millions of examples to learn a task. In situations where there are few speakers of a language and there is scarce presence of digital information, such data simply do not exist. It is a big challenge to come up with. The cost of creating data is costly and consumes more labor. It requires:

• Digitizing current texts: Scanning and OCR-ing books, newspapers, and documents:

• Transcribing audio: Manually typing out spoken language recordings.

• Annotating data: Attaching linguistic descriptions (such as parts of speech) to text necessary to numerous NLP models. Such tasks demand financial expenditure as well expertise of native speakers and linguists.

To most communities especially those with an economic disadvantage, it is a mammoth task mobilizing such resources (Bender, 2019). Unless there is concerted effort to solve this information choke point, the linguistic gap in AI will continue to expand.

5. Discussion: Interpreting the Implications

5.1. Algorithmic Bias and its Linguistic Consequences

The results obtained clearly show that the use of AI systems is not a neutral tool. They are the biases that the data they are trained on has. When a few hegemonic languages are overrepresented in the training data, it results in some sort of algorithmic bias with significant linguistic implications. It goes beyond poor performance in the face of low resources, and it is a structural problem that upholds the existing dis-

parities (O›Neil, 2016).

An example is that a sentiment analysis model that has been trained mainly on American English writings would possibly not understand sarcasm or politeness rules in other types of English, not even other languages. A content moderator may consider harmless phrases in a minority language offensive because it cannot put this in its cultural context. It may contribute to censorship and silencing of the online marginalized voices. It has the impact of creating an online world that implicitly goes after speakers of majority languages, where the rest must change or remain marginalized.

5.2. The Economic Drivers of Linguistic Inequality

The creation of AI is keyed mainly by the business interests. NLP is researched by major technology companies in billions of dollars; however, this research is inherently geared towards markets with the highest amounts of affluent users. It does not make much sense to invest in the development of AI on a language with only a few thousand speakers, some of whom might lack easy access or purchasing power to internet services (Muhnhagesh, 1996). Such market logic produces a strong and enduring feedback mechanism otherwise known as a vicious cycle as shown in Figure 1.

Figure 1: The Vicious Cycle of the Digital Language Divide. This diagram illustrates how a poor AI support due to the lack of digital data of low-resource languages. This lowers the usefulness of the language on the internet and hence the designing of new internet material is discouraged and the cycle of data shortage is created again. This cycle starts by the original scarcity of digital information. There is no massive market to be exploited to benefit the company, which implies that there is no substantial business reason to invest in the costly data production. No efficient AI-based systems can be constructed without data.

These tools are absent and, therefore, this language becomes less helpful and prestigious in the digital world. This, on its turn, promotes speakers of other languages to shift to the dominant language to get to the economic and social mobility, further killing the user base and the production of the content in their native languages. This is not so novel, but AI can also increase this process by turning the digital world into an even more essential component of the modern world.

5.3. Counter–Narratives: AI for Language Revitalization

Nevertheless, with the unpromising outlook the growth of commercial AI development presents, an increasing number of trends seem to be in favor of utilizing AI as engine of good in language revitalization. These are usually the initiatives of academics, non–profits and of the linguistic communities themselves. They are also interested in creating technology to meet the requirements of the individual speakers instead of focusing on commercial gain.

The application of AI to create language learning tools is one of the strong examples. An example of such a project is the First Voices project in Canada, which collaborates with Indigenous communities to create online dictionaries, language applications, and keyboards layouts to their language.

AI can enhance these efforts. An example is that a TTS system can be trained to use the recordings of elders and to provide a voice of a language, which can be used by the learners even without having access to fluent speakers. The other important area is documentation. AI based forced alignment software can receive a recording of audio and a crudely transcribed text and automatically identify the words that correspond with the audio. This radically accelerates the work of generating transcribed corpora which is vital in the linguistic analysis as well as the construction of additional NLP devices (Gales, 2011).

5.4. The Role of Community–Led Initiatives

More importantly, efficient language revitalization efforts based on AI are efforts that are self–led by the language communities. Top–down methods in which technologists come up with tools without consulting the speakers do not always work. The people living in the community are the best people who understand their language and culture. They are in the best position to determine their needs, direct the process of development and make sure that the technology that is born is culturally relevant and actually of use (Czaykowska–Higgins, 2009). This is a neighborhood–based approach to the conventional paradigm of tech development. It is more so– sovereign, and the linguistic information obtained is not under the administrative control of the community. Such projects as Common Voice by Mozilla are among the first attempts to adopt the approach by crowdsourcing speech data in many languages, becoming the voice of any speaker himself or herself

in order to create open–source datasets (Mozilla, 2002). These projects prove that they can develop high quality data even on the low resource languages as long as there is a collective effort among them.

6. Conclusion

6.1. Summary of Findings

This dissertation has explored the multifaceted and two–sided nature of the interaction between the Artificial Intelligence and the linguistic diversity. The study discovered that the trend in AI development which has been majorly influenced by business interests is enhancing a digital language gap. Modern NLP technologies do not support even the vast majority of the languages of the world, which is also based on the lack of digital representation in such languages. This technological inequality relegates the speakers of the low–resource languages and is a great threat to the existence of the endangered languages because it depreciates them in the digital sphere that is gaining so much importance. Based on case studies of machine translators and voice assistants, we see a sharp difference between the performance of high and low resource languages, which is the bias of the algorithm at work. Nonetheless, another interesting finding of the research was the existence of a strong counter–narrative. AI provides language documentations and preservations as well as revitalization opportunities never seen before. The AI can be applied to language acquisition and documentation through developing these tools that will be invaluable when developed in tandem with the linguistic communities. Such efforts are a sign of a way ahead where a focus is made on linguistic equity rather than profit.

6.2. Limitations of the Research

The paper is based on a qualitative meta–analysis of sources available, and hence is restricted by the quantity and extent of the same information available to people. The sphere of AI is developing at a very fast rate, and any picture of the contemporary state of technology is prone to be outdated very soon. Moreover, this study is a general picture of a global problem; more localized work on its effect on different communities of language should be done. The absence of a unified reporting on the provided language assistance by major technology corporations also complicates the quantitative analysis undertaken on the subject matter. .

6.3. Recommendations for an Equitable Future

Judging by the results, the following are the recommendations in this dissertation towards a more linguistically inclusive digital future:

- To the Policymakers and Funding Bodies: Governments and other international bodies such as UNESCO must make an investment in open source digital corpora of low-resource languages. This would give the supporting data required to develop AI. There must be the establishment of policies, where digital services should promote the official and the indigenous languages in the areas they are conducted.
- In the case of the Tech Industry: Companies struggle to avoid solely market outpouring in respect to language support. They ought to use part of their earned profits to invest in developing technologies of the low-resource languages as a corporate social responsibility. Another priority should also be to develop the methods of low-resource NLP that can be applied with fewer amounts of data.
- To the Research Community: Computer scientists and linguists need to work more in liaison with language communities. The points of interest of the research should be creating community-centered design methodologies as well as designing tools that are transparent and flexible and owned by the communities that it serves.
- In the case of Language Communities: Communities must be enabled to have their own digital language projects. Basic data collection and technology use training programs may empower speakers to become the actors in the making of a digital future of their language.

6.4. Concluding Remarks

The point between language and AI is a key-point of address that will create the future of the way humanity communicates. Unchecked by now, AI is expected to contribute to the tendencies toward language loss and homogenization. This is not also a necessary consequence, though. Technology is an instrument and its influence relies on the intentions and values of people who created and used it. It can be purposefully oriented to ensure that AI leads to the future where technology is helpful, not a replacement of the diverse web of human language, which is possible through a conscious decision to follow the path of linguistic diversity, as well as through collaboration between technologists, researchers, and speakers. This is a difficult task, but the

rewards, our continued human heritage, are far greater.

7. References

1. Abney, S., & Bird, S. (2010). The Human Language Project. Proceedings of the Workshop on NLP and Linguistics: Finding the Common Ground.

2. Bender, E. M. (2019). The #BenderRule: On Naming the Languages We Study and Why It Matters. The Gradient.

3. Bird, S. (2020). Decolonising Speech and Language Technology. Proceedings of the 28th International Conference on Computational Linguistics.

4. Chowdhary, K. R. (2020). Natural Language Processing. In Fundamentals of Artificial Intelligence. Springer.

5. Crystal, D. (2000). Language Death. Cambridge University Press.

6. Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork. In J. Newman (Ed.), The linguistics of giving. Mouton de Gruyter.

7. FirstVoices. (n.d.). About Us. First Peoples› Cultural Council.

8. Fuchs, C. (2014). Social Media: A Critical Introduction. SAGE Publications.

9. Gales, M. J. F. (2011). Challenges in speech recognition for under-resourced languages. Spoken Language Technology Workshop.

10. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

11. Harrison, K. D. (2007). When Languages Die: The Extinction of the World›s Languages and the Erosion of Human Knowledge. Oxford University Press.

12. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of NLP in the Wild: A Case Study on Low Resource Indian Languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

13. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft).

14. Kornai, A. (2013). Digital Language Death. PLoS ONE, 8(10).

15. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

16. Mozilla. (2002.). Common Voice. Mozilla Foundation.

17. Mühlhäusler, P. (1996). Linguistic Ecology: Language Change and Linguistic Imperialism in the Pacific Rim. Routledge.

18. Nettle, D. (1999). Linguas de Povoação: The Worldwide Extinction of Language. Oxford University Press.

19. O›Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

20. Pimienta, D. (2007). The language divide in the digital world. UNESCO.

فصلية مُحكمة تُعنى بالبحوث والدراسات العلمية والإنسانية والفكرية
العدد (١٧) السنة الثالثة جمادى الآخرة ١٤٤٦ هـ كانون الأول ٢٠٢٥م

Al-Thakawat Al-Biedh Maga-

Website address
White Males Magazine
Republic of Iraq
Baghdad / Bab Al-Muadham
Opposite the Ministry of Health
Department of Research and Studies
Communications
managing editor
07739183761
P.O. Box: 33001
International standard number
ISSN 2786-1763
Deposit number
In the House of Books and Documents
(1125)
For the year 2021
e-mail
Email
off reserch@sed.gov.iq
hus65in@gmail.com

فصلية مُحكمة تُعنى بالبحوث والدراسات العلمية والإنسانية والفكرية
العدد (١٧) السنة الثالثة جمادى الآخرة ١٤٤٦ هـ كانون الأول ٢٠٢٥م