Open Access | Original Research
Journal of Intercultural Communication, 24(1), 2024 | PP: 203–212
https://doi.org/10.36923/jicc.v24i1.171

A Comparative Study of the Lexical Ambiguity of Arabic, English, and French in Natural Language Processing

Bahia Zemni[1], Mimouna Zitouni[2], Farouk Bouhadiba[3], Mashael Almutairi[4]

Article History:
Received: 02-02-2023
Accepted: 19-10-2023
Publication: 01-03-2024

Cite this article as:
Zemni, B., Zitouni, M., Bouhadiba, F., & Almutairi, M. (2024). A Comparative Study of the Lexical Ambiguity of Arabic, English, and French in Natural Language Processing. *Journal of Intercultural Communication*, 24(1), 203–212. https://doi.org/10.36923/jicc.v24i1.171

Corresponding Author:
Mimouna Zitouni
Translation Department, Princess Nourah bint Abdulrahman University, Saudi Arabia. Email: MBZitouni@pnu.edu.sa

Abstract: Ambiguity in certain syntactic structures within the same language has consistently presented challenges for both human translators and machine translation systems. These challenges become even more complex in the machine translation of genetically unrelated languages, such as Arabic, English, and French. The lexical ambiguity of Arabic in Natural Language Processing (NLP) poses additional difficulties when the semantic nuances of Arabic words differ significantly from their English equivalents. This is often the case when two or more Arabic words correspond to a single word in English. Additionally, semantic gaps between the two languages contribute to ambiguity in NLP. This paper explores specific instances of ambiguity in machine translation between Arabic and English, as well as between Arabic and French. The issues examined include segmentation, determination/non-determination, coordination, and the function of words as meaningful units. The paper also discusses how constituents are segmented into grammatical categories and compares these structures in Arabic, English, and French.

Keywords: Non-Concatenation, Structure, Machine Translation, Vocalisation, Non-Vocalisation, Segmentation, Ligature.

1. Introduction

The Arabic language, a member of the Semitic language family, has its origins traced back by historians—and more recently by linguists—to ancient tongues such as Akkadian (circa 3000 BC) from Mesopotamia (Andrews, 2023; Trad, 2021). Languages like Phoenician and Abyssinian (Amharic) exhibit linguistic features akin to Arabic to varying degrees (Leslau, 2021; Royster, 2020). However, Hebrew is notably closer to Arabic, especially in morpho-syntactic and lexical aspects (Zaretsky & Russak, 2023).

This study investigates the nature of language, focusing on Arabic in the context of machine translation and natural language processing (NLP). We aim to analyze text-processing methods and their effects on translation into English and French and to review research findings on NLP and machine translation. The study highlights the challenges and uncertainties encountered in this domain, particularly in Arabic morphology and syntax (Shirko, Omar, Arshad, & Albared, 2010), which frequently impedes machine translation. Issues such as segmenting, parsing, and coordinating particles like "waw" have been significant hurdles in translating Arabic into English and French. This paper discusses the repercussions of these difficulties in translating Arabic semantic doublets into English or French, as well as managing the coordination copula 'waw' (و), which carries multiple meanings. Such ambiguities in machine translation arise because the copula's meaning depends on the entire sentence context, while its syntactic role as a coordination unit remains constant.

The impact of linguistic disparities on the output of machine translation is profound. Arabic's rich morphology, distinctive word formation, and unique grammatical features present considerable challenges to automatic translation systems (Alqudsi, Omar, & Shaker, 2019; Zakraoui, Saleh, Al-Maadeed, & Alja'am, 2020-2021). These challenges are not limited to vocabulary and syntax but extend to cultural, contextual, and expressive aspects (Khalatia & Al-Romanyb, 2020; Gamal, 2020). English and French, as Indo-European languages, bring their own complexities and idiosyncrasies that exacerbate translation difficulties.

From a morphological and syntactic standpoint, Arabic displays features that do not align seamlessly with Indo-European structures, leading to semantic and syntactic ambiguities in machine-translated content (Levshina, 2022). The primary goal of this paper is to provide various examples showing that without a tailored system and expert filters to mitigate these ambiguities, machine translation may not faithfully

[1] Translation Department, Princess Nourah bint Abdulrahman University, Saudi Arabia. Email: baalzemni@pnu.edu.sa
[2] Translation Department, Princess Nourah bint Abdulrahman University, Saudi Arabia. Email: MBZitouni@pnu.edu.sa
[3] Translation Department, Princess Nourah bint Abdulrahman University, Saudi Arabia. Email: bouhadibafaroukoran2@gmail.com
[4] Translation Department, Princess Nourah bint Abdulrahman University, Saudi Arabia: moalmutairi@pnu.edu.sa

reproduce the intended meaning of Arabic texts. Hence, linguists' contributions are crucial for computer analysts to translate these linguistic nuances into effective algorithms. This research delves into specific linguistic phenomena that illustrate the complexities of translation among these languages. The intricate Arabic root-based morphology and the divergent word orders in French and English each add a unique layer of complexity to the translation process.

1.1. Some Linguistic Landmarks

The Arabic language, just like the languages known as Semitic, is based primarily on a morph-syntax of non-concatenation. An example can be the Arabic root "ŸÉÿ™ÿ®" (k-t-b), which represents the core meaning of "writing." From this root, various words can be derived:

كَتَبَ (kataba) - The basic form, meaning "he wrote."

كِتَاب (kitab) - The noun form, meaning "a book."

كَتَبَتْ (katibat) - The feminine singular form, meaning "she wrote."

مَكْتَبَة (maktaba) - A derived form, meaning "a library" or "a bookshop."

This indicates that morphemes in Arabic are not linearly sequenced in the derivational and affixational processes, unlike Indo-European languages, where morpho-syntax involves a linear sequence of morphemes to construct sentences. For instance, the English word "write" undergoes linear affixation to form related words:

write (verb)

writer (noun)

writing (gerund)

wrote (past tense)

written (past participle)

In Indo-European languages, morphemes are sequentially added to the root, creating a linear structure in word formation, in contrast to the non-linear approach of Semitic languages like Arabic, where root consonants are central, and morphemes are interwoven non-linearly to express meaning.

Arabic showcases derivational processes that are more pyramidal than linear or concatenative. Consequently, models that are effective for describing Indo-European languages—such as structuralism, functionalism, or generative models—may not adequately depict the non-linear, pyramidal nature of Arabic. Theories like the Auto-Segmental Model, as described by Goldsmith, J.A. (1976) or McCarthy (1979, 1981), and his Prosodic Theory of Nonconcatenative Morphology, provide a foundation for exploring the current Arabic language and its translation within NLP. These theories elucidate how phonological and morphological components within linguistic structures interact within a grammatical framework (McCarthy & Prince, 1996). Bais and Machkour (2023) endorses the prosodic morphology framework as optimal for analyzing Arabic diminutive formation, as it addresses the non-concatenative templatic morphology of the language.

Cavali-Sforza, Soudi, and Terulo (2000) presented research using a concatenative strategy from an auto-segmental perspective, demonstrating that infixes pose significant challenges in Arabic Morphology analysis. They propose treating infixation as a distinct process, separate from prefixation and suffixation, which can be analyzed more effectively from a concatenative viewpoint.

More recently, Kastner et al. (2019) suggested that Distributional Morphology (DM), combined with non-concatenative models, could accurately represent Arabic's morphological structure. They posited that DM allows for a straightforward modeling of morphology, where syntax informs the interfaces and non-concatenative morphology emerges from the mapping of syntax onto its interfaces.

Machine Translation (MT) outcomes illustrate that while prefixation and suffixation based on DM are generally unproblematic, infixation in Arabic presents significant challenges for MT. For instance, Reverso online's translations sometimes overlook the syntactic significance of infixes, leading to incorrect renderings. Accurate MT requires addressing both syntactic and semantic nuances in Arabic's complex morphological structure.

1.1.1. Cases of affixation that do not pose particular problems:

"إنما هو ابن أخي قدمت به من يثرب" (innama huwa ibnu akhi qadimtu bihi min Yathrib) is translated by Reverso as: "He is my nephew whom I brought from Yathrib."

"إنها على الأرجح كلمة شكر" (innaha 'ala al-arja'i kalimatu shukrin) is rendered as: "It's probably a word of thanks."

1.1.2. Cases of infixation that pose problems:

"إنقطع وصاح" (inqaṭaʿa wa ṣāḥa) is translated merely as "cut off and shouted," where the syntactic significance of the infix {ن} (n) is overlooked by the machine, and the root "قطع" (qaṭaʿa - "to cut") is used in the translation. A more accurate translation would be: "He interrupted and shouted," where "إنقطع" (inqaṭaʿa) should be rendered as "he interrupted."

"استغفر من ذنوبه" (istaghfara min dhunubihi) is incorrectly translated as "forgive his sins," where the machine neglects the reflexive infix {ت} (t) and uses the root "غفر" (ghafara - "forgive") instead. The most fitting translation would be: "He asked for forgiveness for his sins."

We observe that in instances of Arabic affixation that follow a linear sequence (prefixation and suffixation), Distributional Morphology (DM) seems to manage the Machine Translation (MT) quite effectively. However, difficulties emerge at a syntactic and semantic level when infixation is involved.

2. Literature Review

One of the distinguishing features of the Arabic language, compared to most Indo-European languages, is its non-concatenative nature, unlike the concatenative structure of languages like English or French. Arabic morphemes do not follow a linear, sequential pattern in word formation and syntax. Instead, Arabic uses non-linear morphological processes, such as root-and-pattern morphology, where roots and patterns interact to form words, and affixational processes that do not involve simple concatenation (Zitouni et al., 2022).

Linguistic models based on concatenation, i.e., linear sequencing of units, such as "he complained about his job," do not correspond with Arabic structure. For instance, the Arabic sentence "اشتكى عن عمله" (ištakaa ʕan ʕamalihi), literally 'He complained about his job,' shows that proceeding with parsing based on linear sequencing can compromise the integrity of the root in verbs like "شكى" (šakaa) 'to complain.' Additionally, Arabic's inflectional processes involve the copula "عن" (ʿan), providing the genitive vowel marking in "عمله" (ʕamalihi), and the accusative verb "يخص" (yakhuṣṣu) 'it concerns' producing inflectional suffixation in "يخصه" (yakhuṣṣuhu).

Machine Translation faces challenges with the multiple meanings of the coordination particle 'waw' (و), which changes its meaning and function depending on context. In an asyndetic structure like "استيقظت والشمس طالعة" (istayqaẓtu wa al-šššamsu ṭāliʕa), literally 'I woke up and the sun was rising,' which is better translated as 'I woke up at sunrise,' the 'waw' signifies 'while' rather than 'and.'

Semantic doublets, common in Arabic rhetoric, pose difficulties for MT. Dickins, Hervey, and Higgins (2016) propose truncating one of the items in the doublet, preferably the unmarked one. An example is "يمشي و يسير ببطء" (yamši wa yasīru bi buṭ'), literally 'he walks and walks slowly,' better translated as 'He walks slowly,' where the marked word "يمشي" (yamši) is retained and the unmarked "يسير" (yasīru) is omitted.

Dickins et al. (2016) suggest a strategy where a nominal semantic doublet is translated by its closest adjectival variant, as in "تحليل القيم والأخلاقيات" (taḥlīl al-qiyyām wa al-akhlāqiyyāt), translated as 'the analysis of values and ethics,' where 'moral values' in English conveys both Arabic concepts. They suggest 'partial grammatical transposition' as an effective strategy for handling Arabic's lexical semantic doublets.

"The challenges and intricacies when it comes to translating non-vocalized text from Arabic into English or French are indeed multifaceted. One significant aspect that deserves further consideration is the impact of context and surrounding words on disambiguating the meaning of a word. In Arabic, the lemma takes on multiple meanings based on context, and without the vocalic diacritics, disambiguation becomes a complex task. Furthermore, the direction of Arabic script from right to left adds an additional layer of complexity to translation. It requires a different approach to typesetting and formatting in languages that are written from left to right, like English and French (Ahmad et al., 2021).

For instance, the word "علم" may mean "flag," "knowledge," or "learned" in English, and ‘drapeau,’ ‘savoir,’ or ‘appris’ in French, depending on the context. This contextual dependency underscores the importance of not only understanding the individual word but also the surrounding words and the broader context in which the word is used. Translators often face the challenge of accurately capturing these nuances in the target language.

Another intriguing aspect is the presence of semantic gaps in Arabic compared to English or French. Certain words in Arabic, such as "خالٌ" and "عمٌ," transcribed as /khālun/ and /‘amun/, translate to ‘uncle’ in English and ‘oncle’ in French, respectively, but they carry specific cultural or contextual connotations that may not have direct equivalents in English or French. Translators must navigate these gaps and find the most suitable terms or phrases to effectively convey the intended meaning.

Additionally, the unique phonetic features of Arabic, including gutturals and emphatics, present challenges for pronunciation and transcription in languages with different phonetic systems. An example can be the letter "ع" (ʕayn) in Arabic. The Arabic letter "ع" (ʕayn) has no direct phonetic equivalent in French or English. Therefore, it is important to provide a transliteration and an explanation to help readers understand the unique phonetic feature being referred to. This showcases the challenge of accurately conveying Arabic phonetics in languages with different phonetic systems.

Lastly, Arabic's rich morphological processes, such as affixation and reduplication, as well as its complex morphosyntactic features like tenses and modes, demand a deep understanding of Arabic grammar and syntax to ensure accurate and meaningful translation."

The Arabic language presents a complex challenge in the field of Natural Language Processing (NLP). Arabic NLP has developed numerous tools using MT techniques to analyze the language in both written and spoken forms (Marie-Sainte, Alalyani, Alotaibi, Ghouzali, & Abunadi, 2019). There have been various proposals for processing Arabic, such as Leavitt's (1994) MORPHE, a morphological rule compiler; Soudi and Cavalli-

Sforza's (2003) work on interfacing Arabic morphology with sentence generation systems; Habash's (2010) introduction to Arabic NLP; and Mohamed and Sadat's (2015) hybrid translation approach. This approach couples an Arabic–French statistical MT system using the Moses decoder with additional morphological rules to simplify the source language morphology. Another advancement includes creating Arabic language interfaces for databases, applying supervised learning to transform natural language queries into unambiguous expressions (Bais & Machkour, 2023).

A notable difficulty in Arabic NLP is the absence of vocalization in texts (Debili & Achour, 1998; Bekraoui & Daoudim, 2015). For instance, a non-vocalized root sequence like <ktb> could be interpreted in multiple ways by a machine. The system first seeks the "kernel meaning" of the root to derive related lemmas, such as "كَتَبَ" (kataba - 'He wrote'), "كُتِبَ" (kutiba - 'It was written'), and "كُتُب" (kutub - 'Books').

Today, it is possible to automate the processing of Arabic text, using electronic dictionaries, spell-checkers (Al-Jefri & Mahmoud, 2013), and word-order error detection tools (Azmi, Almutery, & Aboalsamh, 2019). This marks significant progress in Arabic NLP since the 1970s.

The Arabic language has twenty-eight phonemes, with some easily Latinized, such as the labials, dentals, and velars. However, the representation of guttural consonants like "ح" (ḥāʾ), "خ" (khāʾ), "غ" (ghayn), "ع" (ʿayn), and emphatics like "ظ" (ẓāʾ) and "ط" (ṭāʾ) remains problematic. These issues arise particularly with gemination, signaled by the diacritic 'shadda' (ّ), which is not always easily recognized by machines, especially in geminated or causative verb forms. Additionally, vowel lengthening is indicated by the consonants 'alif' (ا), 'waw' (و), and 'ya' (ي), which can also serve as consonants, posing the question of whether machines can accurately determine their role based on context.

3. Research Methodology

The primary objective of this study is to explore and analyze the intricacies of translating the Arabic language, with a particular emphasis on non-concatenative morphology's impact on machine translation. Employing a descriptive and analytical approach, the study examines the unique aspects of Arabic language structure and their implications for machine translation into English and French. We aim to address the issue of ambiguity in machine translation at the levels of morphology, lexis, and syntax by posing the following questions:

- How do non-concatenative morphological processes in Arabic affect machine translation into languages such as English and French?
- What semantic and syntactic ambiguities emerge when translating Arabic text?
- What are the constituent segments and syntagms that compose Arabic words and sentences, and how do these elements contribute to their formation and meaning?

These initial questions will guide our examination of the items under study (words/sentences) to illustrate their construction and meaning determination. We will also analyze the elements within the Arabic words/sentences and their rendition in English and French through the lens of machine translation. This analysis will be qualitative, aiming to assess the extent to which machine translation conveys the meaning of the source items, comparing it to human translation.

For our corpus, we have selected a range of vocalized and non-vocalized examples that form the basis of our analysis.

4. Results and Discussion

4.1. Some Aspects of the Morpho-syntax of Arabic

In morpho-syntax, an Arabic word does not necessarily correspond to the same schema or construction as a word in Indo-European languages. Arabic, primarily constructed on a root system (mainly triliteral and quadriliteral roots), uses these roots to convey core meanings. These roots consist of consonant sequences, whereas Indo-European language roots usually combine consonants and vowels to form a lemma.

Consider the Arabic root <ج ع ل> (ʒ ʕ l), where the radicals' order implies the kernel meaning 'to do something, produce, undertake,' as in "جَعَلَ" (jaʿala). Conversely, rearranging these radicals to <ع ل ج> (ʕ l ʒ) changes the meaning to 'to look after, to relieve, treat...' as in "عَلَجَ" (ʿalaja). Similarly, the radicals /ʕ, l, m/ (ع ل م) in this order form the root <ʕ l m>, producing the non-vocalized "علم" (ʿilm) and vocalized forms such as "عَلَمٌ" (ʿalamun - flag), "عَلِمَ" (ʿalima - to know)[5], "عِلْم" (ʿilmun - knowledge), and so on. Alternatively, the sequence /l, m, ʕ/ (ل م ع) creates the root <l m ʕ>, meaning 'to shine, to illuminate' as in "لَمِعَ" (lamiʿa) or "لَمْعَةٌ" (lamʿatun - brightness), to which the feminine marker ة is suffixed as ـة. The same process applies to quadrilateral roots such as <Twrt> ((ة ،ر ،و ،ث 'revolution ثَوْرَةٌ ' while <Trwt > (ة،و،ر،ث) gives "richness, resources, ثَرْوَةٌ ...'.

It is essential to recognize that arbitrary combinations of radicals do not inherently create new meanings in Arabic. The language enforces constraints imposed by the sonority hierarchy of Arabic sounds. For instance,

[5] . Vowel length is represented here as V_1V_1 in order to draw ~~a~~ parallelism with geminated consonants C1C1 and to abide by Syllable Structure Constraints of Arabic which disallow the sequence $V1V1C1C1$ in a syllable.

Arabic does not permit initial consonant clusters like C_1C_2 or C_1C_1 or C_1 C_2 (C_3)[6], unlike French and English, which allow clusters in words like "produire," "traduire," "produce," and "translate." In Arabic, the root <ʕhd> (عهد - 'the idea of promising') is permissible because pharyngeal sounds are more sonorous than glottal sounds.

Word morphology becomes more complex with the inclusion of inflection and derivation processes like prefixation, infixation, and reduplication. Native speakers can readily identify a word such as "مُستَسمَحٌ" (mustasmaḥun), composed of the instrumental prefix "مُ" (mu-) and the root "سمح" (smḥ - 'to forgive'), translating to 'allowed' in English or 'permis' in French. However, whether machines can identify roots and construct words in affixation processes as adeptly as native or proficient speakers remains uncertain.

This example presents a challenge for machines to recognize words based on their components to find equivalents in English or French. It also necessitates proper segmentation in Arabic, which significantly differs from that in Indo-European languages. Proposals to model segmentation in English or French for adaptation to Arabic, such as the MORPHE[7] morphological system, are still debated.

In machine translation from Arabic to English or French, substantial difficulties arise with elementary parts of speech such as articles, prepositions, conjunctions, and pronouns, as well as with adjectives, adverbs, nouns, and verbs. Questions arise about what segmentation can be used in language modeling, especially when identifying equivalents in Arabic compared to French and English. For example, syntactic ambiguities with the determinate/indeterminate copula, the coordination particle "وَاوٌ" (waw), and semantic doublets in Arabic, like "بِصِفَةٍ مُستَمِرّةٍ مُتَوَاصِلَة" (bișifatin mustamirratin mutawāșilatin), cannot be translated directly as 'continuously continuous' in English or French. Instead, human translators might opt for truncation, resulting in 'in a continuous manner' or 'de façon continue.'

4.2. Parts of Speech and Syntactic Ambiguity

Many substantial difficulties are encountered in the machine translation of Arabic into English or French, particularly with basic parts of speech such as articles, prepositions, conjunctions, and pronouns, as well as with categories like adjectives, adverbs, nouns, and verbs. One fundamental question concerns the type of segmentation that should be used in language modeling, especially when comparing genetically related languages like French and Spanish, and to a lesser degree, English and French. More specifically, we will discuss cases that are problematic in identifying equivalents in Arabic as compared to French and English. These include determinate/indeterminate copulas, the coordination particle with the conjunction "وَاوٌ" (waw), and semantic doublets in Arabic, such as "بِصِفَةٍ مُستَمِرّةٍ مُتَوَاصِلَة" (bișifatin mustamirratin mutawāșilatin). These expressions cannot be directly translated into English as *' in a continuously continuous manner' or into French as *'de façon continuellement continue.' Instead, human translators must truncate the second word to produce phrases such as 'in a continuous manner' or 'de façon continue.'

In broader terms, Arabic differs from English and French primarily in its syntactic structure, verb tenses, and modes. Arabic frequently employs the VSO (Verb-Subject-Object) order as a marked structure, whereas the SVO (Subject-Verb-Object) order is considered the unmarked structure. Unlike English or French, Arabic does not typically use the passive voice. Instead, it uses a passive construction where the agent is not specified. The language operates with three tense forms—the perfective, the imperfective, and the participle—and five modes: the indicative, the imperative, the subjunctive, the jussive, and the energetic.

Here are some examples: Structure VSO: **كتب الكاتب كتاب**, vocalized as **كَتَبَ الكاتب كتاب,** /kataba al-kaatib kitaab/. Literal translation: 'Wrote the writer a book.' This is rendered in English as 'The writer wrote/has written a book.' The tense in English depends on the context in Arabic, where the verbal form is used: 'wrote' or 'has written'.

- Structure VSO: The sentence "كتب الكاتب كتاب" is vocalized as "كَتَبَ الكاتب كتاب")/kataba al-kaatib kitaab/). Literal translation: 'Wrote the writer a book.' This is rendered in English as 'The writer wrote a book' or 'The writer has written a book,' depending on the context, in Arabic, where the verbal form is used.
- The Passive form: The phrase "كتب الكتاب" is vocalized as "كُتِبَ الكتاب")/kutiba al-kitaab/). Literal translation: 'Written the book.' This is rendered in English as 'The book was written,' where the agent is unspecified, in contrast to English constructions that may specify the agent with 'by X' or in French as 'par Y'.
- The Perfective Tense: "كَتَبَ")kataba) serves as the base form of the verb, since Arabic does not have infinitive forms like 'to write' in English or 'écrire' in French. The infinitive in Arabic is equivalent to the 3rd person masculine singular of the perfective, as in "كَتَبَ")/kataba/) 'He wrote'.
- The Imperfective Tense: "يَكْتُبُ")/yaktubu/) 'He writes,' where prefixation and suffixation act as inflectional markers.
- The Past Participle: "كُتِبَ")/kutiba/) 'was written' (masculine singular).

[6]. The Arabic dialects do not have these syllable structure constraints. Initial and final consonant clusters are attested in most varieties of the Arabic language.

[7]. The Morphe system was conceived to treat a type of concatenative morphology and it was adapted to Arabic (Leavitt, 1994. MORPHE: Morphological Rule Compiler has. Technical Carryforward, Cmu-cmt-94-memo).

- The Imperative: "أَكْتُبْ")/uktub/) for 'Write!' directed at a masculine singular you.
- The Subjunctive: Used with particles like "عَلَيْكَ أَنْ" or "لازم")/ʔalaika an/ or /lazim/), meaning 'must' or 'have to,' as in "لازم تكتب")/lazim taktub/) or "عَلَيْكَ أَنْ تَكْتُب")/ʔalaika an taktub/), conveying 'You must write' or 'It is necessary that you write.'
- The Jussive: Used with the negation particle "لَمْ")/lam/) as in "لَمْ أَكْتُب")/lam aktub/), translating to 'I have not written.' Note that in English, this is expressed in the negative form, while in Arabic, it is in the jussive mood.

The use of certain particles in Arabic, compared to English and French, can be problematic for machine translation. The definite article in Arabic, "ال" (al), is singular and invariable, akin to "the" in English, while in French, it varies as "le," "la," or "les." Placed at the beginning of a word, it indicates determination as opposed to indetermination, which is unmarked in Arabic. For example, "الكتاب" (al-kitaab) demonstrates determination, whereas "كتاب" (kitaab) indicates indetermination. In English, the counterparts are "the book" and "a book," while in French, they are "le livre" and "un livre," respectively.

Gender in Arabic is marked by the final "ة" (ta marbuta), serving as the singular feminine marker, as seen in "مكتبة" (maktaba(tun)) for 'library.' Although rare, it is possible to find Arabic male names with "ة," such as "حمزة" (Ḥamzatun) for Hamza. The feminine plural is created through suffixation using the extended vowel "آت" (aat), as in "مكتبات" (maktabaat).

Given the definite article "ال" (al) in Arabic, a machine will recognize its equivalents in English ("the") or in French ("le," "la," "les"). Additional considerations include the vowel alternation processes in Arabic that apply to gender and number. For instance:

- "الكتاب" (al-kitaab) - 'the book'
- "المكتبة" (al-maktaba(tu)) - 'the library'
- "الكتب" (al-kutubu) - 'the books'
- "المكتبات" (al-maktabaat) - 'the libraries'

With the indefinite article, we have:

- "كتاب" (kitaab) - 'a book'
- "مكتبة" (maktabatun) - 'a library'
- "كتب" (kutubun) - 'books'
- "مكتبات" (maktabaatun) - 'libraries'

If we exclude the less common cases of irregular plurals in Arabic, which activate different linguistic mechanisms, the issue of equivalence for definiteness/indefiniteness between Arabic and English or French does not pose significant challenges for isolated word formations. However, the difficulty arises in syntactic structures where the definite article "ال" (al) is attached to the word it specifies, such as in "الجامعة" (al-jaamiʿa(tu)) - 'the university.' Unlike English or French, where articles are separate from the word they modify, the Arabic definite article is attached to the specified word.

There are words in Arabic that begin with "ال" (al) where it is an integral part of the word and not a definite article[8]. Examples include "لعبة" (luʿba(tun)) - 'a game,' "لاعب" (laaʿib(un)) - 'a player,' "لهب" / "التهاب" (lahab, iltihaab) - 'a flame' or 'inflammation' from the quadriliteral root (lthb), "لطيف" (laṭiif) - 'kind,' and "لاجئ" (laajiʿ) - 'a refugee' from the triliteral root (laʒaʿ) 'refugee.' In the determined form, the word for 'flame' takes a doubled "ل" as in "اللهب" (allahab) - 'the flame,' "اللاجئ" (allaajiiʿu) - 'the refugee,' and the noun "اللجوء" (al-lujuuʿ) - 'refuge.' The challenge for a machine is to discern when "ال" (al) functions as the definite article and when it is part of the lemma itself.

The ambiguity lies here in the fact that unlike English or French where the definite or indefinite articles are separated from the following word they specify or unspecify, the definite article {al} of Arabic is attached to the word it specifies[9].

In some instances, the definite article in the Arabic syntagma does not translate directly into English or French. This is common in fixed expressions or so-called frozen sentences, such as "ممنوع التدخين")/mamnuuʿ at-tadkhiin/), which translates to 'No smoking' in English and 'Interdit de fumer' in French, where the definite article does not appear. It is noted that for the translation of the Arabic definite article, the English equivalent is "the,"

[8]. We shall not discuss here cases of lunar vs. solar consonants where le definite article assimilates regressively to the following solar consonant. We shall not also discuss cases of idhafa where the noun is unspecified followed by a specified marker as the case may be for prossessive nouns, rendered in Arabic syntax as mudhâf (the attached marker) and mudhâf ilayh (the unspecified noun which becomes determinate in the absence of the definite article {al}.

[9]. We shall not discuss here cases of lunar vs. solar consonants where le definite article assimilates regressively to the following solar consonant. We shall not also discuss cases of idhafa where the noun is unspecified followed by a specified marker as the case may be for prossessive nouns, rendered in Arabic syntax as mudhâf (the attached marker) and mudhâf ilayh (the unspecified noun which becomes determinate in the absence of the definite article {al}.

while in French, additional information on gender and number is required to provide the correct equivalent for the definite article "le," "la," or "les."

In an Arabic sentence like "كيف الحال؟")/kaifa al-ḥaal?/), the translation is "How are you?" rather than the literal "How is the situation?" which is often incorrectly produced by automatic translation software on the Web. Without determination, Arabic and English exhibit some similarities, particularly in the plural form of non-specified words, both using the empty symbol Ø, as in "books" for /kutub/. English uses "a/an" for unspecified nouns, whereas Arabic employs vowel alternation to move from the singular "كتاب")/kitaab/) 'a book' to the plural "كتب")/kutub/) 'books.'

The issue of translating unspecified nouns from Arabic to French is more complex, while the translation from French to Arabic is more straightforward. French uses a range of indefinite articles "un," "une," "des," "du," "de la" for unspecified words. In Arabic to French translation, these are all represented by Ø in Arabic, as Arabic uses the tanwîn (نون التنوين /tanwiin/) for indeterminate forms. However, when translating from Arabic to French, machines often struggle to choose the appropriate French indefinite article.

For example:

- "The writer's book" translates to "كتاب الكاتب")/kitaab al-kaatib/).
- "The poet's book" translates to "كتاب الشاعر")/kitaab ash-shaʿir/), where "ال" assimilates to the "ش")/sh/).
- "The book of poets" translates to "كتاب الشعراء")/kitaab ash-shuʿaraaʔ(u)/).

But when translating such sentences from Arabic to French, machines may fail to accurately render the indefinite nature in Arabic into French indefinite articles "un," "une," "des," "du," "de la." For instance, "كتاب شعراء")/kitaab shuʿaraaʔ(un)/) could translate to either "Le livre des poètes" or "Le livre de poètes."

Another source of ambiguity in translation from Arabic to English or French concerns coordination. Arabic uses nine particles of coordination, which are frequently employed in both written and spoken language.

In Arabic, coordination is expressed through nine particles, which are frequently used in both written and spoken forms. These are "مع, لكن, بل, أو, أم, حتى, ثم, ف, واو")/maʕa, laakin, bal, aw, am, ḥatta, ṯumma, fa, waw/)[10], which approximately translate to "with, although, however, or (aw), or (am), until, then, then, and" respectively. The particle 'waw' is particularly ambiguous in Arabic machine translation. It connects syntagmas without necessarily implying a logical relation between them, unlike the English "and" or the French "et." Its high frequency of occurrence and various meanings make 'waw' one of the most challenging particles to translate. In Arabic, 'waw' carries both syntactic and semantic loads, serving different purposes such as coordination, relationship, swearing, doubt, separation, etc. These particles, although clear in Arabic syntax and semantics, pose ambiguity and problems in Arabic Natural Language Processing. Consequently, assigning an equivalent in English or French for 'waw' can be challenging.

For instance, in the Arabic syndetic structure "جلست وجلست السيدة ورداء")/jalastu wa jalasati as-sayidatu warda/), 'waw' conveys both simultaneity and the independence of two actions performed by separate actors. In this case, 'waw' could be translated into English as "and," producing "Mrs. Warda and I sat down." The French equivalent could be "Je me suis assis ainsi que Madame Warda" or "Nous nous sommes assis, Warda et moi." The English translation conveys simultaneity, whereas the French allows for variability in expressing this simultaneity.

Another example is "إن تركوه هلك وهلكوا")/in tarakūhu halaka wa halakū/), where 'waw' does not directly translate to "and" as it would in "If they leave him, he will perish, and they will perish." A more nuanced translation might be "If they leave him, he will perish, and so will they," indicating the collective fate. In French, automatic translation might produce an awkward sentence. A more accurate translation could omit the conjunction "et," as in "S'ils l'abandonnent, ils périront tous," removing 'waw' from the sentence.

The construction of Arabic words from roots to form lemmas is unique due to the use of radicals and inserted vowels, known as "حركات")/ḥarakāt/). The root <ktb> signifies 'to write,' and with thematic vowels, we get different forms such as "كتب")/kataba/), "كتب")/kutiba/), "كتاب")/kitaab/). Derivation processes can produce words like "مكتبة")/maktaba/) for 'bookshop' or 'library,' "كتيب")/kutayyib/) for 'a small book,' and "اكتتب")/iktataba/) for 'to correspond.'

In Western linguistic tradition, a word is a sign used in a grammatical form, consisting of an acoustic or graphic form and a meaning, forming a semantic, phonological, and syntactic unit. Smaller units, morphemes, also

[10]. Caution being taken here not to translate these coordination copulas of Arabic into English or French, simply because most of them may not take the same meaning. The Syntactic-semantic meaning is inferred from the sentences or the syntagms in which they are used. We chose ‘waw’ because it is the most problematic coordination particle when it comes to the transfer from Arabic to other languages; in this case English and French.

carry semantic, phonological, and grammatical properties. The relationship between lexical and grammatical morphemes varies across languages, often leading to ambiguities in machine translation from Arabic. For instance:

Arabic: "زهرة صغيرة")/zahratun ṣaghīratun/) - 'small rose,' with two words and four morphemes.

English: "little rose" - two words and two morphemes.

French: "petite rose" - two words and three morphemes (considering the gender agreement).

Spanish: "poca flor" - two words and two morphemes. Italian: "fiorellino" - one word with two morphemes. German: "wenig Blume" - two words and two morphemes.

NB (These translations were taken from online automatic translators).

5. Conclusion

To summarize, it is evident that machines cannot systematically analyze languages of different genetic relationships using uniform models of segmentation, as each language has its own unique characteristics. The task seems more manageable in cases of semantic gaps than in cases of syntactic relations within a language, which may sometimes be similar in languages of the same family. This observation is valid in machine translation across all levels of analysis, including morphological, syntactic, and lexical.

For instance, consider the Arabic syntagma "من هو المترجم")/man huwa al-mutarjim/), which consists of four words and five morphemes ("man," "huwa," "al," "mu-tarjim," with "mu-" being a prefix marker for the instrumental derived from the verbal root "ت ر ج م")trjm)). It translates into English as "Who is the translator," which has four words but only three morphemes.

The challenge becomes more pronounced in French due to the translation of "من")/man/) from Arabic, which can be rendered as "who," "what," "which" in English, and as "qui," "lequel," "laquelle," etc., in French.

In fact, the task of Arabic natural language processing and translation into English or French is not as straightforward as it may seem. Problems emerge not only on the semantic level, as illustrated by the English word "uncle," which has two specific meanings in Arabic: "عم")/ʿamm(un)/) for paternal uncle and "خال")/khaal(un)/) for maternal uncle. Similarly, a single word in English, like "lion," can correspond to several words in Arabic within the same semantic field, such as "أسد," "ضرغام," "قشم")/asad, ḍirghaam, qasham/).

The primary issue lies in the area of segmentation and parsing, as these languages do not exhibit the same word formation or syntactic/semantic parsing as Arabic. Ambiguity at the semantic level is generally less challenging than the ambiguity encountered at the syntactic level for machine translation. This is compounded by cultural features specific to each language and the problems of cultural transposition that may arise.

We will attempt to address some aspects of the segmentation of Arabic compared to other languages, particularly English and French, along with phraseology cases and the divergences they cause at various levels. These are challenges that researchers in automatic translation (Arabic/English/French) are likely to encounter.

Notes

[1] . Vowel length is represented here as V_1V_1 to draw a parallelism with geminated consonants C1C1 and to abide by Syllable Structure Constraints of Arabic which disallow the sequence $V1V1C1C1$ in a syllable.

[2]. It is important to note that dialects of Arabic generally lack these strict syllable structure constraints. Many varieties of Arabic allow for initial and final consonant clusters.

[3]. The Morphe system was conceived to treat a type of concatenative morphology and it was adapted to Arabic (Leavitt, 1994. MORPHE: Morphological Rule Compiler. Technical Carryforward, CMU-CMT-94-memo).

[4] The passive construction in the Arabic language is known as the *mabni lil majhul* or the construction with no agent. This state of affairs has a religious connotation based on the belief that Allah (God) is the only one who knows and who can predict the future.

[5]. We will not discuss here cases of lunar vs. solar consonants where le definite article assimilates regressively to the following solar consonant. We will not also discuss cases of idhafa where the noun is unspecified followed by a specified marker as the case may be for possessive nouns, rendered in Arabic syntax as mudhâf (the attached marker) and mudhâf ilayh (the unspecified noun which becomes determinate in the absence of the definite article {al}.

[6]. Caution being taken here not to translate these coordination copulas of Arabic into English or French, simply because most of them may not take the same meaning. The Syntactic-semantic meaning is inferred from the sentences or the syntagms in which they are used. We chose ‘waw’ because it is the most problematic coordination particle when it comes to the transfer from Arabic to other languages, in this case, English and French.

Acknowledgment statement: The study was funded by the Literature, Publishing and Translation Commission, Ministry of Culture, Kingdom of Saudi Arabia, under [10/2022] as part of the Arabic Observatory of Translation.

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contribution statements: Conceptualization, MZ, and BZ; Methodology, FB., and MZ; Resources, MA, FB; Writing-Original Draft, MZ., FB, BZ, and MA; Writing-Review & editing, BZ, FB, MZ; Supervision, BZ and MZ; Project administration, MA; Funding Acquisition, BZ, and MA.

Funding: The research team received funding from the Literature, Publishing and Translation Commission, Ministry of Culture, Kingdom of Saudi Arabia, under [10/2022] as part of the Arabic Observatory of Translation.

Data availability statement: Data is available at request. Please contact the corresponding author for any additional information on data access or usage.

References

Al-Jefri, M., & Mahmoud, S. A. (2013). Context-Sensitive Arabic Spell Checker Using Context Words and N-Gram Language Models. *In the Proceedings of Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, Al-Madinah al-Munawwarah, Saudi Arabia, 1, 258-263. *https://doi.org/10.1109/nooric.2013.59*

Alqudsi, A., Omar, N., & Shaker, K. (2019). A hybrid rules and statistical method for Arabic to English machine translation. *In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1-7). IEEE.

Andrews, E. D. (2023). Introduction to the Text of the Old Testament: From the Authors and Scribes to the Modern Critical Text. *Christian Publishing House.*

Azmi, A. M., Almutery, M., & Aboalsamh, H. (2019). Real-Word Errors in Arabic Texts: A Better Algorithm for Detection and Correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8), 1308–1320. *https://doi.org/10.1109/taslp.2019.2918404*

Ahmad, I. (2021). Emotional Regulation as a Remedy for Teacher Burnout in Special Schools: Evaluating School Climate, Teacher's Work-Life Balance and Children Behaviour. *Frontiers in Psychology*, 12, 1–10. https://doi.org/10.3389/fpsyg.2021.655850

Bais, H., & Machkour, M. (2023). A rule-induction approach for building an Arabic language interface to databases. *The International Arab Journal of Information Technology*, 20(1), 49 – 56. *https://doi.org/10.34028/iajit/20/1/6.*

Bekraoui, F., & Daoudim, M. (2015). Conception Et Réalisation D'un Vocaliseur Automatique De Texte Arabe "Tashkil*." El-Wahat Journal for Research and Studies*, 8(2). *https://doi.org/https://doi.org/10.54246/1548-008-002-036*

Cavali-Sforza, V, Soudi, A & Terulo, M (2000). Arabic Morphology Generation Using a Concatenative Strategy. *In Proceedings of 1st Meeting of NAACL*, 86 -93, Seattle, USA.

Computation approaches to Semitic languages, Montreal, Canada. *https://doi.org/10.3115/1621753.1621761*

Debili, F., & Achour, H. (1998). Voyellation automatique de l'arabe. *In the Proceedings of the Workshop on*

Dickins, J., Hervey, S., & Higgins, I. (2016). Thinking Arabic Translation: A course in translation method: Arabic to English. *London: Routledge.* https://doi.org/10.4324/9781315012650

Gamal, M. Y. (2020). Context, field, and landscape of audiovisual translation in the Arab world. *ESSACHESS-Journal for Communication Studies*, 13(1(25)), 73-105.

Goldsmith, J.A. (1976). Auto-segmental Phonology. *Doctoral Thesis*, Massachusetts MIT, USA.

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis lectures on human language technologies*, 3(1), 1-187. https://doi.org/10.1007/978-3-031-02139-8

Kastner, I., Tucker, M. A., Alexiadou, A., Kramer, R., Marantz, A., & Massuet, I. O. (2019). Non-concatenative morphology. *Ms., Humboldt-Universität zu Berlin and Oakland University.*

Khalatia, M. M., & Al-Romanyb, T. A. H. (2020). Artificial intelligence development and challenges (Arabic language as a model). *Artificial Intelligence*, 13(5), 83-105. http://dx.doi.org/10.17977/um056v7i1p83-105

Leavitt, J. R. (1994). MORPHE: A morphological rule compiler. *Technical Report, CMU-CMT-94-MEMO.*

Leslau, W. (2021). An annotated bibliography of the Semitic languages of Ethiopia (Vol. 1). *Walter de Gruyter GmbH & Co KG.*

Levshina, N. (2022). Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*, 26(1), 129-160. https://doi.org/10.1515/lingty-2020-0118

Marie-Sainte, S.L., Alalyani, N., Alotaibi, S., Ghouzali, S., & Abunadi, I. (2019). Arabic Natural Language Processing and Machine Learning-Based Systems. *IEEE Access*, 7, 7011–20. *https://doi.org/10.1109/access.2018.2890076.*

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12(3), 373-418. https://scholarworks.umass.edu/linguist_faculty_pubs/26

McCarthy, J. J., & Prince, A. (1996). Prosodic Morphology. *In J. Goldsmith (Ed.), The Handbook of Phonological Theory*, (pp. 318-366). Cambridge, MA: Blackwell.
McCarthy, J.J. (1979). Formal Problems in Semitic Phonology and Morphology. *Doctoral Thesis*, Massachusetts MIT, USA.
Mohamed, E., & Sadat, F. (2015). Hybrid Arabic–French machine translation using syntactic re-ordering and morphological pre-processing. *Computer Speech & Language*, 32(1), 135–144. *https://doi.org/10.1016/j.csl.2014.10.007*
Royster, P. D. (2020). The Geography of Racial Bias. *In Decolonizing Arts-Based Methodologies*, (pp. 110-149). Brill.
Shirko, O., Omar, N., Arshad, H., & Albared, M. (2010). Machine Translation of Noun Phrases from Arabic to English Using Transfer-Based Approach. *Journal of Computer Science*, 6(3), 350–356. *https://doi.org/10.3844/jcssp.2010.350.356*
Soudi, A., & Cavalli-Sforza, V. (2003). Interfacing an Arabic Morphology and Sentence Generation with an English-to-Arabic knowledge-based Machine Translation System. *In the Proceedings of the workshop on Information Technology*, Rabat, Morocco, March (pp. 17-19).
Trad, A. (2021). The Societal Transformation Framework: The Nation of Semites–The Phoenicians. *In Management and Conservation of Mediterranean Environments*, (pp. 227-259). IGI Global. https://doi.org/10.4018/978-1-7998-7327-3.ch015
Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2020, April). Evaluation of Arabic to English machine translation systems. *In 2020 11th International Conference on Information and Communication Systems (ICICS)* (pp. 185-190). IEEE.
Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic Machine Translation: A Survey With Challenges and Future Directions. *IEEE Access*, 9, 161445-161468. https://doi.org/10.1109/ACCESS.2021.3132488
Zaretsky, E., & Russak, S. (2023). Using oral narratives to examine the acquisition of English verb morphology among multilingual Arabic and monolingual Hebrew speakers: finding similarities with monolingual English-speaking SLIs. *International Journal of Multilingualism*, 1-22. https://doi.org/10.1080/14790718.2023.2232376
Zitouni, M., Alzahrani, A., Al Kous, N., Almutlaq, S., Abdul-Ghafour, A.-Q., & Zemni, B. (2022). The Translation of Selected Cultural Items in Al-Nawawi's Forty Hadiths: A Descriptive and Analytical Study. *Journal of Intercultural Communication*, 22(3), 43-53. https://doi.org/10.36923/jicc.v22i3.74

About the Author(s)

Bahia Zemni received her PhD in linguistics from Sorbonne-Nouvelle III University. Since 2012, she has been a professor at Princess Nourah bint Abdulrahman University, where she has headed the Languages Faculty Research Center. Currently, she leads a research unit in the Translation Department and contributes to the research project titled "Translation from Arabic to French and Vice Versa in Contextual Dictionaries: mechanisms and strategies." Additionally, she heads the project "Artificial Intelligence and Audiovisual Translation." Bahia has published several translations in collaboration with the Louvre Museum and publishing houses such as Skira in France and Alsaqui in Lebanon. She has participated in several national and international conferences and has published extensively in well-established journals on the subjects of Linguistics and Translations. She recently co-edited a Special Issue of Kervan - International Journal of Afro-Asiatic Studies entitled "Pratiques langagières en arabe: variables culturelles et défis de la traduction."

Mimouna Zitouni is a professor of Translation Studies and Linguistics at the College of Languages, PNU. In 2017–2018, she served as a distinguished Fulbright SIR visiting scholar at Coastal Carolina University (USA). She has participated in several international conferences and has published extensively on the subjects of language use and translation studies.

Farouk Bouhadiba is a professor of linguistics at the College of Foreign Languages, University of Mohamed Ben Ahmed, Oran-Algeria. He specializes in Maghrebi Linguistics and Dialectology from a Sociolinguistics perspective. He also conducts research in the field of Didactics, with a special focus on the LMD implementation in Algeria. Currently, he is engaged in Natural Language Processing and is involved in building a WordNet Arabic. **Dr. Bouhadiba** has participated in numerous international conferences and has published extensively on various subjects related to language use and translation.

Mashael Almutairi is an assistant professor at Princess Nourah bint Abdulrahman University. She obtained her PhD in Translation from Leicester University in 2018 and subsequently assumed the role of Head of the Translation Department at PNU. Almutairi served as the Dean of the College of Languages from 2019 until May 2021 when she was appointed as the Dean of the Development and Consulting Services Institute. Throughout her professional career, she has been actively involved in various academic associations. She currently serves as an associate editor and reviewer at Altralang Journal and has authored several papers and translated several books.