Journal of Scientific Research in Arts
ISSN 2356-8321 (Print)
ISSN 2356-833X (Online)
https://jssa.journals.ekb.eg/?lang=en

Un-paralleling the Parallel: A Contrastive Stylometric Analysis of H. G. Wells' *The War of the Worlds* Parallel Corpus

Rania A. El-Wakil
Faculty of Languages, October University for Modern Sciences and Arts (MSA), Egypt
relfattah@msa.edu.eg
Olfat N. Kerney
Faculty of Languages, October University for Modern Sciences and Arts (MSA), Egypt
olfat.noureldin@gmail.com

Received:15/3/2024 Revised:16/4/2024 Accepted: 5/10/2024 Published:30/10/2024

DOI: 10.21608/jssa.2024.277120.1620
Volume 25 Issue 7 (2024) Pp. 19-49

Abstract

Contrastive linguistic studies compare and contrast how texts are formed and interpreted in different languages and cultures. Recently, computational tools have been utilized to empirically conduct linguistic analysis. Stylometry is the quantitative study of literary style through computational text analysis. This study attempts a parallel-corpus contrastive stylometric analysis of H.G. Wells' The War of the Worlds (1898) and its Arabic translation (2012). The paper aims to demonstrate the various challenges of English/ Arabic parallel corpus alignment and to explore the effect of the intricate nature of the Arabic language on natural language processing (NLP) attempts by examining English adverbs and automatically recognized named entities of locations, people, and organizations in comparison to their Arabic renditions. For alignment, the heuristic-based NLTK sentence segmenter successfully produces valid alignments though some discrepancy occurs. The part-of-speech (POS) tagger is more trained on English texts. Most English tokens are accurately tagged; however, the tagger underperforms with Arabic tokens, either misidentifying parts of speech or by labelling them X, standing for unidentified. It is evident that Arabic renditions of adverbs fail to parallel those employed in the English source text featuring a variety of morpho-syntactic alternatives. NER tags manifest better results in both texts with the translator's tendency to transliterate named entities. The study concludes by shedding light on some of the factors that might have led to inaccurate alignment and annotation. The study also reflects on the translator's inconsistent choices in translating adverbs and entities of locations and organizations.

Keywords *corpus-based contrastive studies; stylometry; natural language processing (NLP) applications; part of speech (POS) tagging – named entity recognition (NER)*

Abstract

Contrastive linguistic studies compare and contrast how texts are formed and interpreted in different languages and cultures. Recently, computational tools have been utilized to empirically conduct linguistic analysis. Stylometry is the quantitative study of literary style through computational text analysis. This study attempts a parallel-corpus contrastive stylometric analysis of H.G. Wells' *The War of the Worlds* (1898) and its Arabic translation (2012). The paper aims to demonstrate the various challenges of English/ Arabic parallel corpus alignment and to explore the effect of the intricate nature of the Arabic language on natural language processing (NLP) attempts by examining English adverbs and automatically recognized named entities of locations, people, and organizations in comparison to their Arabic renditions. For alignment, the heuristic-based NLTK sentence segmenter successfully produces valid alignments though some discrepancy occurs. The part-of-speech (POS) tagger is more trained on English texts. Most English tokens are accurately tagged; however, the tagger underperforms with Arabic tokens, either misidentifying parts of speech or by labelling them X, standing for unidentified. It is evident that Arabic renditions of adverbs fail to parallel those employed in the English source text featuring a variety of morpho-syntactic alternatives. NER tags manifest better results in both texts with the translator's tendency to transliterate named entities. The study concludes by shedding light on some of the factors that might have led to inaccurate alignment and annotation. The study also reflects on the translator's inconsistent choices in translating adverbs and entities of locations and organizations.

Key words: corpus-based contrastive studies; stylometry; natural language processing (NLP) applications; part of speech (POS) tagging – named entity recognition (NER)

1. Introduction

Contrastive linguistics, an applied discipline of linguistics, aims primarily at examining similarities and differences across different languages (cross-linguistic contrastive studies) or within individual languages (intra-linguistic contrastive studies) in order to establish language-specific, typological and/or universal

linguistic patterns. Contrastive studies explore both abstract language systems as well as contextualized communicative instances. They either take theoretical standpoints providing descriptive accounts of the language features contrasted, or go beyond abstraction of linguistic features to manifest practical implications applied in related disciplines. Incorporating monolingual and multilingual corpora, corpus analysis measures and techniques have contributed to further developments in both micro-linguistic contrastive studies and interdisciplinary macro-linguistic research fields. The current study attempts a parallel corpus contrastive stylometric analysis of H. G. Wells' *The War of the Worlds* (1898) and its Arabic translation حرب العوالم (2012). The study is triggered by an inquiry on the potentials of computational methodologies employed in contrastive studies to unravel the linguistic and contextual complexities of the languages examined as well as the challenges faced. The study attempts to assess the "parallel" relationship between source text (ST) and target text (TT). By "un-paralleing" the two texts, patterns and instances of structural, lexical and morpho-syntactic disparity are detected. Employing Natural Language Processing (NLP) tools, the paper aims to: (1) demonstrate the various challenges of English/Arabic parallel corpus alignment, (2) demonstrate the effect of the intricate nature of the Arabic language on NLP attempts, (3) examine English adverbs with reference to their Arabic renditions, and (4) examine recognized named entities of locations, people and organizations across both texts.

2. Theoretical Background

2.1 Corpus-based Contrastive Studies

Contrastive linguistic studies, as a systematic examination of similarities and differences among selected languages or within the varieties of one language, assume different theoretical standpoints, adopt various methodological frameworks and bear diverse implications. Two core concepts are in question in contrastive studies: correspondence and methodology. Correspondence, *tertium comparationis*, pertains to the selection of particular features or instances of the languages in contrast that manifest either formal, functional or semantic degree of equivalence (Krzeszowski, 1990). Correspondence must be maintained to ensure initial validity of the comparison. Contrastive studies are conducted at various linguistic levels, witnessing major developments in scope and techniques. Studies vary from granular

structural perspectives to finer functional scopes. Under the influence of structuralism, early contrastive studies have attempted correspondence by abstracting selected features in a pair of languages. Micro-linguistic contrastive studies, such as those of Lado (1957) and Agard and Di Pietro (1965), adopt a structuralist approach probing decontextualized abstract structures and selected systems mainly of phonology, morphology, syntax and lexicon. The descriptive accounts of these studies manifest evident 'granularity' (Gast, 2012). However, they represent "a drastically limited view of language" as they lack "insights into how language actually functions in extralinguistic settings . . . [and] insights into how this structure is used to perform its numerous functions" (Krzeszowski, 1990, p.48). Scholarly attempts, such as James (1980), Krzeszowski (1990) and Gast (2012), aim to broaden the scope of contrastive linguistics, enhance methodology and set reliable procedural techniques. The contrastive scope has broadened 'vertically' as larger linguistic units and communicative instances are examined, and 'horizontally' by incorporating socio-cultural setting and further extralinguistic features (James, 1980, p.102). Emergence of macro-linguistic disciplines has given rise to such studies as contrastive pragmatics (Fillmore, 1984), contrastive sociolinguistics (Janicki, 1980; Bugarski, 1991), and contrastive discourse analysis (Sajavaara & Lehtonen, 1980).

Contrastive studies have witnessed categorical developments with the integration of computational tools. Parallel and comparable corpora with their computational affordances set an advantageous methodological footing for the examination of various linguistic features across languages (Aijmer & Altenberg, 2013). Seminal works by Johansson (2012) and others have given rise to 'a new era' of corpus-based studies, providing finer perceptions and more empirically-based comparisons of the language pair(s), rather that intuitive insights into form and function. The outset of corpus-based contrastive studies is to develop tools that align texts and add computational annotation of a preselected linguistic formal/functional category. The methodology yields quantitative results allowing a more systematic qualitative detection of correspondence/ divergence patterns in terms of structure, function and/ or semantics. Utilizing 'principled' sets of digitally processed source and translated texts (parallel corpora) or equivalent authentic texts (comparable corpora) takes contrastive studies away from the decontextualized abstraction of

linguistic features towards a more authentic framework where linguistic features are contrasted in context of use (Hasselgård, 2020).

Corpus-based contrastive studies manifest a variety of micro as well as macro linguistic comparisons of specific lexical categories, word combinations and collocations (Gundersen, 2004; Egan, 2012; Hasselgård, 2017); syntactic constructions (Egan 2018); discourse phenomena such as cohesion and thematic structure (Rørvik, 2003; Lewis, 2017) and pragmatics (Thormodsæter, 2020; Wu, 2022).

2.2 Corpus-based Translation Studies

Both corpus-based contrastive studies and translation studies share the empirical methodology of corpus analysis, though differing in purpose and research context. Contrastive analysis lies at the heart of translation studies. Corpus-based translation studies, introduced by Baker (1993), employ bilingual and multilingual parallel corpora in addition to comparable corpora to conduct statistical analysis of the features of translated texts (TTs) in comparison to source texts (STs) and/or non-translated equivalents. The plethora of corpus-based translation studies provide a rich empirical paradigm of quantitative and qualitative analysis pertaining to translation process in terms of linguistic features, conceptual techniques as well as socio-cultural factors contributing to the translated product (Laviosa, 1996; 2004; Olohan, 2004; Oakes & Ji, 2012). Contexts of corpus-based translation studies include: translation universals, linguistic features of translated texts, and translator's style.

Baker (1993) postulates that translated texts share a set of universal features – translation universals (TUs) – resulting from the translation process, regardless of the interlingual differences. TUs include explicitation, implicitation; simplification, and normalization. Examples of studies on translation universals are Olohan (2002), Steiner (2008), Álvarez de la Fuente and Fuertes (2015), Moghaddam et al. (2017), Molés-Cases (2019), Bartkute (2020), Liu and Afzaal (2021), and Kwok et al. (2023).

Studies on linguistic features of translated texts compared to non-translated texts essentially include lexical features, such as lexical density, type/token ratio and

high frequency words, n-grams, collocations and semantic prosody (Baker, 2004; Ebeling, 2013; Kotait, 2016), and syntactic features, such as sentence length and complexity, constructions load, punctuation, and contractions (Olohan, 2003).

Translator's style or "thumbprint" has a broad sense and a narrow sense. Broadly speaking, the study of translator's style pertains to specific selections of source texts – preferred translation strategies and structure of the translation text. In a narrow sense, a translator's style manifests in recurrent linguistic patterns (Baker, 2000; Hu, 2016). Such an approach employs comparable corpora of translated texts of several renditions or a corpus of translated texts by the same translator. Pertinent to the scope of the current research, a translator's style could be detected by means of examining specific items in the target text in relation to those in the source text, focusing on structural, lexical and semantic alterations intensified use or omission instances in context (Hu, 2016), in addition to characteristic corpus analysis measures. Study of translator's style and characteristic features of translated texts show direct relevance to stylometry.

2.3 Stylometry

For Lowe and Matthews (1995), stylometry is stylistic statistics; that is adopting computational methods to investigate characteristic features of a text. Stylometric investigation is closely related to natural language processing (NLP) and machine learning (ML) algorithms (Daelemans, 2013; Lagutina et al., 2019). The extracted features, stylometric entities, function as "interpretable statistical indicators" of the particular text(s) style (Langlois, 2021, p.53). Stylometric entities include lexical features, syntactic structures, structural features, and content specific features. In monolingual contexts of research, stylometry is an acknowledged methodology of authorship attribution, validation and profiling in addition to text – gender/sentiment classification. In a survey of 50 stylometric studies, Lagutina et al. (2019) categorize "the most popular stylistic features" under three main categories: character-level, word/token-level, and syntactic level (p.193). Specific detected features are character and word n-grams, type/token frequency, vocabulary (content-based), stop words, parts of speech (POS), sentence length and structure. Other less common features include embedded characters, errors, punctuation, semantic relations, rhythmic features of rhyme and stress. A variety of algorithm are employed

such as adjacency networks (Amancio, 2015; Stanisz, et al., 2019), sequential rules (Boukhaled & Ganascia 2015), Convolutional Neural Networks (Ruder et al., 2016; Sari et al., 2017), integrated syntactic graphs (Gómez-Adorno et al., 2016), Support Vector Machine (SVM) and Logistic Regression (Gómez-Adorno et al., 2018) among others.

Contrastive stylometric studies pertain to detection of style change or stylistic idiosyncrasies of authors/texts in contrast. Mostafa and Nebot (2018) conduct a contrastive stylometric analysis of the use of the word "Árabe" by three Spanish writers employing multiple algorithms: word frequency lists, a lexical variety index, concordancing in addition to multidimensional scaling, principal component analyses and cluster analysis. The results manifest distinct linguistic and stylistic features of each author although they belong to the same generation. Modoc and Gârdan (2020) offer a "pilot [stylometric] experiment" to differentiate genuine Romanian novels and other minor, tertiary novels published between 1920 and 1940 (p.49).

In a bilingual context, stylometric translation studies arise. El-Fiqi et al. (2011) develop a set of 'signatures', or translator's identifying features of two Quran translations. The model is based on the concept of network motifs. The text is represented as a network of nodes with adjacency links. The distribution of extracted 3-gram motifs function as "a signature for the corresponding translator" (p.2039). Rybicki and Heydel (2013) examine a collaborative Polish translation of a Virginia Woolf's *Night and Day* adopting authorship attribution techniques to identify translators, based on "a multivariate analysis of most-frequent-word frequencies" (p.708). The take-over point is successfully identified.

In a parallel vein, Lynch and Vogel (2015) investigate English and German translations of three plays by Henrik Ibsen, detecting "distinctiveness of textual contributions of characters" comparing "character homogeneity" between source and translated texts and among translated texts (p.1). The texts are parsed into character contributions followed by analysis of n-gram tokens to detect whether character idiosyncrasies are preserved. Lynch and Vogel (2018) examine translator's 'fingerprint'. Three translations of Ibsen's *Ghosts* are stylometrically analyzed against a reference corpus of English translations of Anton Chekhov's texts by other

translators. Distinctive textual features are retrieved through a variety of algorithms: Support Vector Machines, Simple Logistic Regression, Naïve Bayes and Decision Tree classifiers. Frequencies of these features are compared to corresponding frequencies in the reference corpus to establish a claim of translator's stylistic choices, rather than the effect of the source language and the limitations of the topic and genre.

In stylometric contrastive analysis, most frequent words are represented as feature vectors in contrasted texts. Text similarity is calculated in terms of how distant the texts are to one another, in technical terms, "the Manhattan distance of z-scores of the frequencies of n most frequent words in the collection" (Cinková & Rybicki, 2020, p.977). This does not allow a direct comparison between source text and target text due to language barriers. To overcome this issue, Cinková and Rybicki (2020) employ Universal Dependencies (UDs) for morphosyntactic markup; texts in the parallel corpus are parsed with the corresponding language model in UDPipe. A further step is generating cross-lingual 'pseudolemmas' of an aligned bilingual glossary.

In the same vein of research, the current study attempts a stylometric analysis of adverb phrases and named entities of locations, people and organizations in H. G. Wells' *The War of the Worlds* (1898) and their Arabic renditions in the Arabic translation (2012).

2. Methodology

3.1 Building the Corpus

To compile the parallel corpus, the entire English text and its Arabic translation are segmented into sentences. The Natural Language Tool Kit (NLTK) sentence tokenizer is a trained model that identifies sentences (Bird et al., 2009). The tokenizer is a heuristic model that relies on rules to identify sentence boundaries. Table 1 shows how the model succeeds in identifying sentence boundaries.

Table 1: Example of Successful Sentence Segmentation Across both Texts

English Text	Arabic Text
I crept forward, saying "Good dog!" very softly; but he suddenly withdrew his head and disappeared.	تسللت للأمام قائلًا بصوت خافت: «أيها الكلب المطيع!» لكنه سحب رأسه فجأة، واختفى.
The risk is that we who keep wild will go savage—degenerate into a sort of big, savage rat . . . You see, how I mean to live is underground.	الخطر يكمن في أننا — نحن الذين سيرفضون الخضوع لهذا الترويض — سنعود إلى بربريتنا ... إنني أنوي الحياة تحت الأرض.

In the first pair, the model is able to recognize that the exclamation point does not break the sentence, for it is enclosed in quotation marks. In the second pair, it recognizes the dots as a case of ellipses, instead of being mistaken for full stops. However, while the segmentation of the English text results in 3062 tokens, the Arabic translation produces 2877 tokens. Table 2 explains with examples some of the reasons that led to the 185-token discrepancy.

Table 2: Examples of Alignment Discrepancy

#	English Text	Arabic Text
1.	Eastward, over the blackened ruins of the Albert Terrace and the splintered spire of the church, the sun blazed dazzling in a clear sky, and here and there some facet in the great wilderness of roofs caught the light and glared with a white intensity.	-
2.	He proposed, he said, to make his way Londonward, and thence rejoin his battery—No. 12, of the Horse Artillery.	قال إنه يعتزم الذهاب باتجاه لندن، ومِن هناك يعاود الانضمام إلى سريته؛ التي تحمل الرقم (١٢) والتابعة لمدفعية الخيالة.

3.	"He is dying fast, and very thirsty. It is Lord Garrick." "Lord Garrick!" said my brother; "the Chief Justice?" "The water?" he said. "There may be a tap," said my brother, "in some of the houses.	«إنه يحتضر، وقد بلغ به الظمأ مبلغه. إنه اللورد جاريك.» قال شقيقي: «لورد جاريك؟ قاضي القضاة؟» قال الرجل: «الماء؟» قال شقيقي: «ربما يكون هناك صنبور في أحد المنازل.
4.	About eleven, as nothing seemed happening, I walked back, full of such thought, to my home in Maybury. But I found it difficult to get to work upon my abstract investigations.	نحو الحادية عشرة — عندما بدا لي أنه ما من جديد — عدت أدراجي إلى منزلي في «مايبري» تستبد بي تلك الفكرة، لكني واجهت صعوبة في بدء العمل على أبحاثي النظرية.

As the table illustrates, the first sentence has no equivalence in the Arabic translation. This is the case with 12 instances, in which the translator provides no equivalence. This accounts for complete sentences that are captured by the tokenizer, regardless of phrases that are also ignored within the sentences. For example, the English sentence "'Nay,' shouted the curate, at the top of his voice, standing likewise and extending his arms." drops the adverbial phrase describing the curate's posture in the Arabic translation.

As for the second pair, it can be seen that the tokenizer has mistaken the dot after the 'No.' abbreviation for a full stop, causing the sentence to break into two segments. Since Arabic uses the complete word form 'رقم' (number), the outcome is one segment. The third example illustrates the kind of errors that arise from the structure and punctuation of quotes. The translation attempts to maintain the same structure, by placing the reporting verbs within the quotes. As a result, the tokenizer results in 3 segments for both. It can be seen, though, that the second segment is different. In the English text, the tokenizer is trained to recognize quotation marks as part of sentences. Hence the break only occurs after the full stop following the

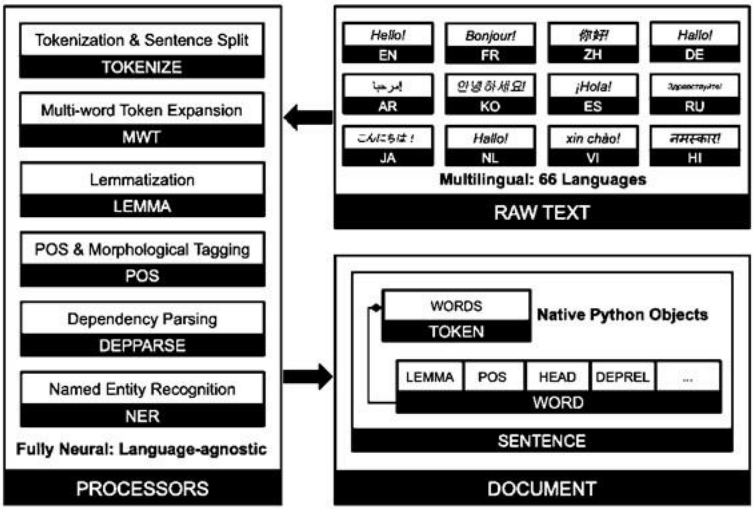
reported verb. In the Arabic text, the reporting verb does not occur between "Lord Garrick!" and "the Chief Justice?" as in the English text, it occurs at the beginning of the quote to result in "قال شقيقي: «لورد جاريك؟" (Lord Garrick!" said my brother) and "قاضي القضاة؟» قال الرجل" ("the Chief Justice?" he said.).

The fourth pair epitomizes discrepancies that may result due to different punctuation marks. Whereas the English text presents two sentences, joined by 'but', the Arabic texts offers one sentence joined by 'لكني'. The full stop breaks the two sentences, whereas the comma causes the tokenizer to recognize them as one. To remedy the discrepancies, a manual review was conducted. The review results in 2758 pairs.

3.2 Annotating the Corpus

Stanza is a Python NLP package of tools for linguistic analysis that supports more than 70 languages. It is used for part-of-speech (POS) tagging and named entity recognition (NER) tasks (Peng, et al., 2020). Figure 1 below summarizes the linguistic analysis tools provided by Stanza.

Figure 1: Overview of Stanza's Tools (Peng, et al. 2020)



The stylometric analysis begins by drawing a generic comparison between the two texts making up the parallel corpus. Table 3 exhibits some numerical findings, from which it can be seen that whereas both the English text and its Arabic translation manifest a similar number of tokens, the Arabic text is richer in lexical

variation. Not only do unique words (where each word is counted only once, regardless of how many times it appears) make up 18.5% (11266) of the overall count of words (60793), unique lemmatized words constitute 12.6% (7691) of the total number of words. This is to be contrasted with 12.4% (7287) and 9.5% (5569) in the English text, respectively.

Table 3: Generic Stylometric Analysis of the Parallel Corpus

Item of Comparison	English Text	Arabic Text
Number of Tokens	58415	60793
Number of Unique Tokens	7287	11266
Number of Unique Lemmas	5569	7691
Number of Unique Tokens with Unidentified POS	3	1273

As for the Stanza NER Tagger, it supports 4 tags across many languages, including Arabic and English. These tags are PERSON (people and characters), LOC (locations), ORG (organizations), and MISC (miscellaneous). The Stanza English model also supports another 18-tagset introducing more detailed entities such as nationalities, products, events, works of art, laws, and other numerical information. Nevertheless, for the sake of consistency, the PERSON, LOC, and ORG of 4-tagset are to be explored in this study across the parallel corpus.

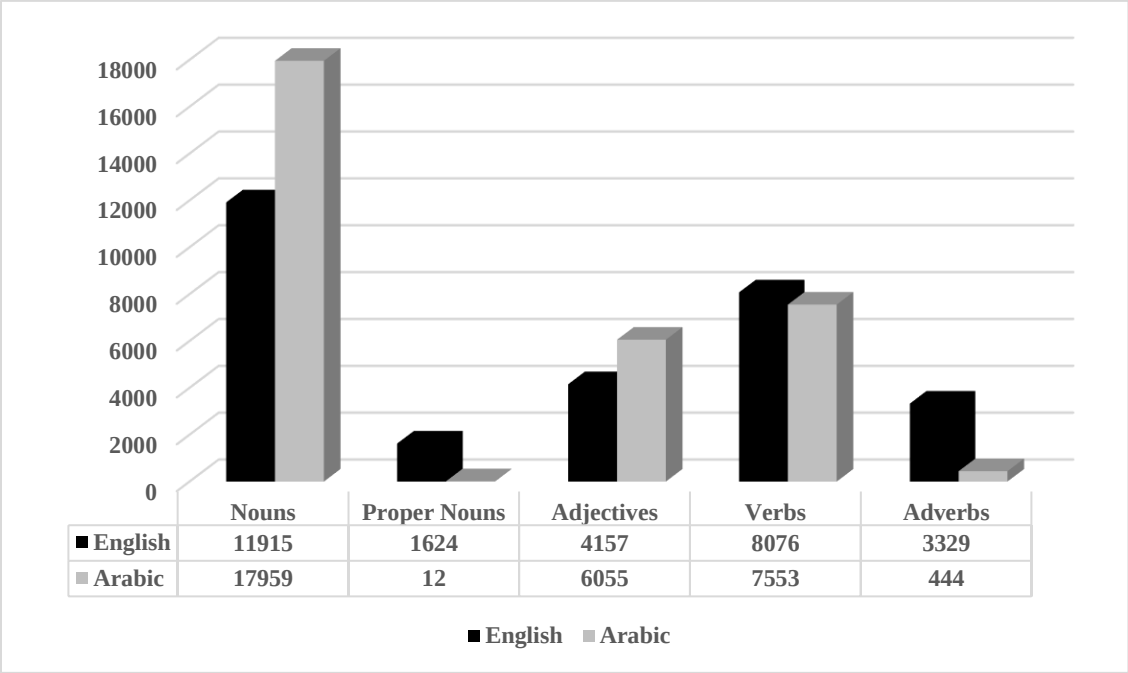
3. Results and Analysis

4.1 POS Tagging: Adverbs

One of the challenges met during the stylometric analysis is the POS tagging. It can be seen from Table 3 that the Stanza tagger is better trained on English data, for it only fails to identify the parts of speech of 3 words; in fact, these three words are the Martian siren 'Ulla' and its lower-cased version 'ulla' and the Latin number 'V'. On the other hand, the model fails to identify 1273 words in the Arabic text. Some of the unidentified words are transliterations of English proper nouns, which might explain why the system fails to contextualize them. Nevertheless, many

Arabic words fail to be identified, such as 'سددتها' (I had given him), 'صدغي' (my temple), 'أترى' (see), 'فيُزاحون' (go out), 'أمتعتي' (luggage), and 'لانهائي' (infinite). Figure 2 below gives a detailed account of the distribution of tags among the two texts.

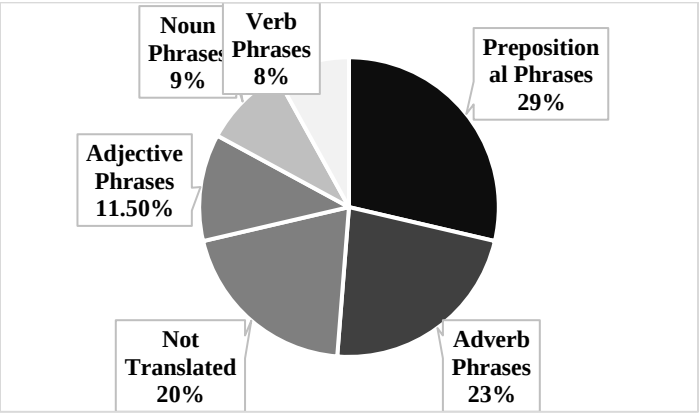
Figure 2: POS Tags in the English and the Arabic Texts



	Nouns	Proper Nouns	Adjectives	Verbs	Adverbs
English	11915	1624	4157	8076	3329
Arabic	17959	12	6055	7553	444

As the Figure 2 illustrates, proper nouns and adverbs display the widest discrepancies; while the tagger identifies 1624 proper nouns in the English text, it only identifies 12 in the Arabic translation. In fact, 11 unique proper nouns are identified. It can be argued that such proper nouns are common English places and hence were identified by the tagger. More strikingly is the discrepancy between the number of identified English adverbs and their Arabic counterparts. Whereas 2154 unique adverbs out of a total of 3329 are identified in the English text, only 23 unique adverbs out of a total of 475 are identified. To explore this discrepancy, all adverbs are explored across the parallel corpus. Figure 3 illustrates how the 2154 unique English adverbs are translated into Arabic.

Figure 3: POS Tags of Arabic Translations of English Adverbs



Per the chart, almost 58% of Arabic renditions of English Adverbs take various forms: 28.5% prepositional phrases (PP), 8% verb phrases (VP), 9% noun phrases (NP), 11.5% adjective phrases (AdjP). Only 23% are translated into their expected equivalent Adverb phrases. 20% adverb instances are not translated. A detailed account of the results is illustrated in Table 4.

Table 4: Detailed Account of Adverb Translations into Arabic

PPs		AdvPs		AdjPs		NPs		VPs	
P + N	65.3%	Circumstantial Accusative	38%	ADJ	84.5%	N	55%	V	77.5%
P + N + ADJ	14.3%	Adv of Time	34%	ADJ + N	8.5%	N + N	27%	V + N	14%
P + N + N	12%	Adv. Of Place	27%	ADJ + ADJ	7%	N + ADJ	17%	V + N + (ADJ/ N)	5.6%
P + ADJ	5.2%							V + (V/ ADJ/ N)	2.8%
P + (ADJ/ N/ V)	3%								

Per Table 4, structural divergence is evident. Inconsistency of translation could result from a number of factors; first, adverbs in English constitute a distinct lexical word class with different functional categories: manner, time, place,

frequency, degree, and conjunctive adverbs. Morphologically, they vary in form; adverbs either have a fixed or a derivative form. Functionally, adverbs modify adjectives, verbs, clauses as well as other adverbs; hence, they vary in syntactic positioning. The semanto-syntactic equivalent to adverbs in Arabic are: circumstantial accusative (الحال), adverbs of time and place (ظرفا الزمان والمكان). The three equivalences fall under the umbrella word class NOUN. Morpho-syntactically, circumstantial accusative (الحال) is similar to adjective (النعت); both are derived modifiers that follow the modified noun in number and gender. Adjectives follow the noun in definiteness/indefiniteness and case whereas circumstantial accusative occurs in an indefinite form and in the accusative case. There is no consistent direct equivalence between English and Arabic adverbs; some adverbs manifest direct semanto-syntactic equivalence, such as now (الآن), above (أعلى/ فوق), below (تحت) and eastward (شرقا). On the other hand, the semantic equivalent sometimes necessitates a syntactic divergence. Table 5 displays some examples.

Table 5: Examples of Syntactic Divergence in Arabic Translations of English Adverbs

Adverb	Translation	Notes
very	شديد / كثيرا / جدا	The Arabic equivalents are usually tagged as adjectives.
now and again	من وقت لآخر	The Arabic translation is 2 prepositional phrases each made up of a preposition and a noun.
more/ less	الأكثر / الأقل	Whereas the English adverbs are used in comparative structures, the Arabic equivalents are usually adjectives used for both comparative and superlative structures.
Indoors	في منازلهم	The Arabic translation is a prepositional phrase made up of a preposition, a noun and a pronoun.

Almost	كاد / تكاد / أكاد	The English adverb is translated into a verb and is used in various forms (past, third-person present, first-person present).
speechlessly	قده القدرة على الكلام	The Arabic translation presents an equivalent clause that captures the same meaning; it would be back translated as 'It made him lose his ability to speak.'

It is noteworthy that the translator's stylistic choices vary. For example, one adverb may be translated in different ways, as evident in Table 6.

Table 6: Examples of Translator's Stylistic Choices when Translating Adverbs

Adverb	Translation	POS Tags
northward	شمالا	Adverb of Place
	نحو الشمال	ADV + N
	متجهين نحو الشمال	Circumstantial Accusative + AdvP
sluggishly	على مهل	PP
	يتحركون بخطى ثقيلة	VP + PP
alone	بمفردي	PP
	وحيدا	Circumstantial Accusative
nearer	نحوي	Adverb of Place
	تقترب	V
	مع اقترابي	PP

In addition to the stylistic choices of changing the English adverb to a different part of speech, the translator tends to extend the Arabic equivalent structure adding more semantically specific parts of speech for further clarification of meaning, or embedding a noun phrase – cognate accusative for emphasis of meaning; Table 7 presents some examples.

Table 7: Examples of Explicitation when Translating English Adverbs

Adverb	Translation	POS Tagging	Back Translation
above	في السماء	PP (P + N)	in the sky
unnaturally early	عن موعدها المعتاد	PP (P + N + PRON + ADJ)	in its usual time
	على غير العادة	PP (P + ADJ + N)	in an unaccustomed way
hurriedly	بخطوات (واسعة) سريعة	PP (P + N + ADJ)	in fast steps
now	في هذه اللحظة	PP (P + PRON + NP)	in this moment
	في تلك اللحظة	PP (P + PRON + NP)	in that moment

It is important to note that per Figure 3, 23% of the total adverb instances in ST are translated in their direct Arabic equivalent: circumstantial accusative, adverb of place or time. However, they are incorrectly POS tagged as adjectives or nouns. For example, in translating the adverb 'just', the translator opts for 'مباشرة', which is a circumstantial accusative, inaccurately tagged as a noun. Similarly, many adverbs of time and place such as حديثاً (newly/ recently), باكرا (early), حول (around), دائماً (always) are also incorrectly tagged as nouns.

4.2 NER Tagging

After running the NER models on the parallel corpus, it is concluded that the model is not very accurate, for it does not capture all tags. In fact, some tags are identified in some instances and ignored in another. For example, one of the most common LOC entities detected by both the Arabic and the English models is 'Horsell'/ 'هورسيل', which is repeated 24 times in the English texts and equally 24 times in its Arabic translation. However, the English NER model only identifies 6 instances, while the Arabic identifies 20. In addition to the missing tags, some tags are misidentified. One striking example is the name 'Oglivy', which is identified by the English model 13 times as ORG and 8 times as PERSON. In total, the English model identifies 329 entities (162 unique entities), only 195 of which are accurately

tagged (48.7%). In contrast, the Arabic model identifies 480 accurate tags from a total of 513 entities (227 unique entities). Table 8 gives a detailed account of the findings.

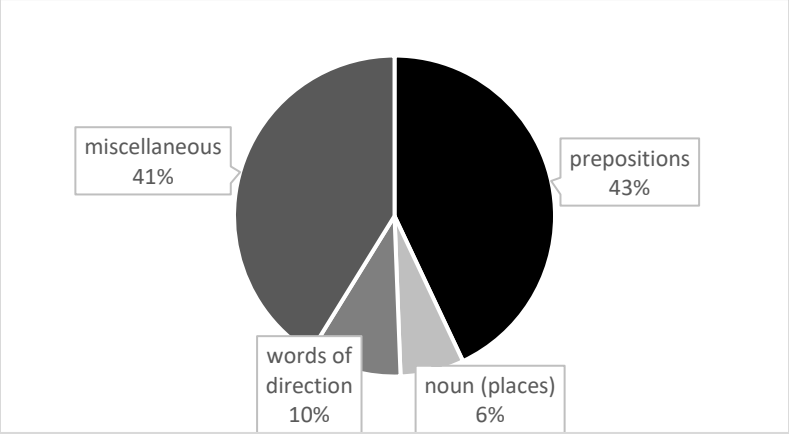
Table 8: Overview of NER Tags across the English and the Arabic Texts

Item	English	Arabic
LOC	111 (unique = 42)	502 (unique = 218)
Inaccurate LOC	8 (accurate = 103/111 = 92.7%)	29 (accurate = 473/502 = 94.2%)
PERSON	133 (unique = 61)	0
Inaccurate PERSON	74 (accurate = 59/ 133 = 44.3%)	-
ORG	85 (unique = 59)	11 (unique = 9)
Inaccurate ORG	52 (accurate = 33/85 = 38.8%)	4 (accurate = 7/11 = 63.6%)
Total number of entities	329 (unique = 162)	513 (unique = 227)
Inaccurate tags	134 Accurate = 195/ 329 (48.7%)	33 Accurate = 480/513 = 93.5%

It can be noted accordingly that Arabic NER model produces more accurate results. However, it fails to identify any PERSON entities, which can be attributed to the fact that in Arabic the absence of capitalization may make it difficult for the machine to identify proper nouns, especially that these proper nouns are in fact foreign to the language. Surprisingly, the model successfully identifies the transliterated LOC entities (recognizing 502 entities with 94.2% accuracy), despite failing to do so in the English model (recognizing only 111 entities with 92.7% accuracy). One suggestion is the use of context clues, which can act as a pattern that can be easily detected by the model. For example, for the 502 entities identified by

the Arabic model, around 80% are marked by context clues. Figure 4 maps the types of context clues that can be traced.

Figure 4: Context Clues of Location in the Arabic Texts



Per Figure 4, the dominant context clue is prepositions. Places are also associated with words of directions, which while cannot be categorized as prepositions also constitute around 10% of the clues. For instance, the words اتجاه and ناحية (toward) make up around 68% and 34% of direction clues. The high percentages prove that these words are very common in Arabic and their association with places help the model detect LOC entities accurately. Strikingly, these context clues are already available in English; prepositions and adverbs such as 'toward' are recurrent; however, the model still fails to recognize the LOC entities.

Furthermore, Arabic names of places (اسم مكان) can be created using a certain word form (مفعل); accordingly, words such as 'مرصد' (observatory), 'محطة' (station), and 'منزل' (house) can form a pattern that can be easily recognized by the model. As for the miscellaneous category, which comprises 41% of the chart, it mainly refers to words that are usually found in contexts where places and locations are discussed. For example, words like نهر (river), طريق (road), شارع (street) help the machine accurately tag LOC entities. It is important to note that the word نهر (river) is usually added by the translator to help with the identification of the place. Only in one instance can it be considered as a translation of the word 'water'.

Despite the inaccuracies, by mapping the tagged entities across the parallel corpus, some stylistic findings are traced. For example, in the translated texts most

PERSON and LOC entities are transliterated. Clearly, all names need to be transliterated, but while an Arabic equivalent of other entities might be available, the translator almost always resorts to transliteration. It has to be mentioned, though, that in some instances, additions and translations are provided. To illustrate, the English NER model identifies 38 compound LOC entities (25 unique ones), the most repeated of which is 'Primrose Hill' – identified 6 times. The translation is a case of inconsistency, for the translator opts to provide a transliteration of 'Primrose Hill' once, a translation of 'hill' alongside a transliteration of 'Primrose' twice, and a transliteration of 'Primrose Hill' and an additional translation of 'Hill' thrice. The different renderings of the same location may confuse the target reader, for the place may be perceived differently. This perception is further intensified when the same place is translated differently. For example, 'Addlestone' is transliterated once as 'أديلستون' and twice as 'أدلستون'. Similarly, 'Cobham' is transliterated once as 'تشوبهام' and thrice as 'كوبهام'. As can be seen, the spelling variations of the same entity may confuse the reader. In the case of 'Addlestone', the different spellings may indicate different locations, which can be misleading. 'Cobham' is confused with 'Chobham', in one instance. It can be argued, thus, that transliteration of LOC entities is the most frequent rendition. Nevertheless, the translator in many instances opts for explicitation by adding translated entities to the transliterations. Table 9 illustrates some examples, which clarify how the translator attempts to create a sense of place that dissipates some of the vagueness surrounding the foreign places: while the invasion takes place in a foreign country, Martians attack familiar places such as roads, colleges and buildings.

Table 9: Examples of Transliteration and Explicitation of LOC Entities

English Text	Arabic Text	Notes
So some respectable dodo in the **Mauritius** might have lorded it in his nest . . .	ربما فعل أحد طيور الدودو المنقرضة في عشه في **جزيرة موريشيوس** . . .	The word 'جزيرة' (island) is added.

About eleven a company of soldiers came through **Horsell** . . .	نحو الحادية عشرة جاءت كتيبة جنود عن **طريق هورسيل**. . .	The word 'طريق' (road) is added.
. . . I saw the tops of the trees about the **Oriental College** burst into smoky red flame. . .	. . . رأيت قمم الأشجار حول **كلية «أورينتال كوليدج»** تشتعل بلهب أحمر دخاني. . .	The word 'college' is both transliterated (كوليدج) and translated (كلية).
From **Ripley** until I came through Pyrford I was in the valley of the Wey . . .	من **شارع «ريبلي»** إلى أن وصلت «بيرفورد» كنت أمر بوادي «واي» . . .	The word 'شارع' (street) is added.
I made my way by the police station and the **College Arms** towards my own house.	مررت بقسم الشرطة و**مبنى «كوليدج آرمز»** متجهًا إلى منزلي.	The word 'college' is transliterated (كوليدج), but the word 'مبنى' (building) is added.

On the other hand, most of the ORG tags identified by the English models are translated. As for those that are transliterated, added words are provided. Table 10 illustrates some examples, which prove how the translations create a sense of urgency as news about the invasion slowly dominate newspaper articles. In addition, they trigger the army's reaction, as different troops attempt to control the situation but to no avail.

Table 10: Example of Transliteration and Explicitation of ORG Entities

English Text	Arabic Text	Notes
Yet the next day there was nothing of this in the papers except a little note in **the Daily Telegraph** . . .	غير أن صُحف اليوم التالي خلت من الحديث عن هذا الأمر باستثناء تعليق موجز في **صحيفة «ديلي تليجراف**. . .	The word 'صحيفة' (paper) is added.
In addition, Ogilvy's wire to **the Astronomical Exchange** had roused	إضافة إلى ذلك، فإن برقية أوجيلفي إلى **جريدة «أسترونوميكال إكستشينج»** قد	The word 'جريدة' (journal) is added.

every observatory in the three kingdoms.	أثارت انتباه جميع المراصد داخل الممالك الثلاث.	
. . . their idea was that a dispute had arisen at **the Horse Guards**.	. . . كانوا يظنون أن نزاعًا قد نشب في **فرقة الخيالة**.	The word 'guards' is translated as 'فرقة' (troop) to sound more familiar to the Arabic reader.
The Cardigan men had tried a rush, in skirmishing order, at the pit, simply to be swept out of existence.	حاول **الفرسان** الهجوم على الحفرة، عن طريق بعض المناوشات، لكن قُضي عليهم تمامًا.	The ORG entity is translated into 'الفرسان' (knights).

4. Discussion and Conclusion

The study attempts a stylometric analysis of a parallel corpus of H. G. Wells' *The War of the Worlds* and its Arabic translation, aiming at exploring stylistic translation choices and how they might affect the cultural and contextual rendition of the source text. In this regard, the study utilizes NLP python tools that help with sentence segmentation for the sake of building the aligned corpus. While the heuristic-based NLTK sentence segmenter successfully produces valid alignments, some discrepancy occurs. It can be deduced that inaccuracy stems from the segmenter's unfamiliarity with dealing with quoted material in the Arabic text, as the translator chooses the symbols «» to enclose quoted material. In addition, the translator's choices to place reported verbs and to leave some sentences untranslated have contributed greatly to the alignment discrepancy.

As for the annotation process, it was found that the POS tagger is more trained on English texts. For it mostly tags all tokens accurately; however, when it comes to the Arabic text, the tagger underperforms, whether by misidentifying parts of speech or by labelling them X, standing for unidentified. Scanning through the unidentified tags suggests that the tagger is either trained on a small or domain-specific datasets; that is, it is not exposed to many literary expressions that are commonly used in narratives.

By zooming in on English adverbs and their Arabic renditions, it is evident that 'un-paralleling' is a distinctive feature of the target text. More than 50% of adverb instances are translated into non-adverb counterparts, leaving out 20% untranslated. Though renditions maintain semantic equivalence, structural divergence is an idiosyncrasy. This could be attributed to the following reasons: a) cross-lingual differences in word class categorization, b) translator's stylistic choices, c) explicitation, d) incorrect POS tagging of adverbs and circumstantial accusatives, and e) incorrect tagging of other parts of speech and inconsistent tagging. It is important to highlight here that Stanza POS tagger fails to capture the intricate nature of Arabic syntax, especially when it comes to adverb (circumstantial accusative) use mainly because it agrees in form with adjectives, nouns, and other adverb types.

As for exploring NER tags across the parallel corpus, it was revealed that, unlike the POS tagger, the NER tagger produces much more accurate results, especially of LOC entities when it comes to the Arabic text. The study attempted to account for such striking results by tracing context clues that might have drawn a pattern recognized during training. It is worth noting that while these context clues are commonly used in the English text, it seems that the model is more attracted to English punctuation patterns, such as capitalization, which is missing in the Arabic language.

The study also concludes that in translating NER tags, the translator opts for transliteration. In this regard, it is worth noting that H. G. Wells maps the events of his story around Surrey and London; the narrator gives a first-person account of the Martian invasion in Surrey and reports his brother's rendition of the events in London. The detailed narration aims to create a sense of familiarity that invites the English reader to experience the Martian attack. This may also intensify the impact as the reader feels the threat of having the familiar places under attack. Such evoked feelings may be lost to the Arabic reader, for not only are the places transliterated, but different variants are offered, adding to a sense of alienation. Nevertheless, to make up for such lost impact, the translator in many instances adds lexical items or provides some translations along with the transliterations to give a sense of place. While this does not help with the familiarity evoked in the source text, it definitely

shows how the invasion quickly overtakes many places around the cities. Combining translating and transliterating techniques in transferring ORG entities also helps maintain the sense of urgency as readers of the target language feel that many organizations are involved in dealing with the crisis in question.

Thus, the findings of the study fall under two categories: computational processing of the parallel corpus and stylistic analysis of the translator's choices. In the former, it is recommended that NLP packages are more exposed to Arabic data, covering multiple genres. It is also recommended that UDs are employed along with heuristics that pay attention to Arabic morphology and syntax. As for the latter, it is believed that corpus translation studies should work to the end of developing translators' stylistic choices in the hopes of achieving more 'faithful' translations. Accordingly, by conducting a computational scan of the corpus in question before translation, translators can gain insights about intentional or unintentional patterns that once maintained in the TT, it would produce more parallel structural and content alignment.

English References:

Agard, F. B., & Di Pietro, R. J. (1965). *The grammatical structures of English and Italian.* University of Chicago Press.

Aijmer, K., & Altenberg, B. (2013). *Advances in corpus-based contrastive linguistics: Studies in honour of Stig Johansson*. Benjamins.

Álvarez de la Fuente, E., & Fernández Fuertes, R. (2015). Translation universals in the oral production of bilingual children. *Translation and Translanguaging in Multilingual Contexts*, 1(1), 49–79. https://doi.org/10.1075/ttmc.1.1.03alv

Amancio, D. R. (2015). A complex network approach to stylometry. *PloS One*, 10(8), e0136076. https://doi.org/10.1371/journal.pone.0136076

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

Baker, M. (2000). Towards a Methodology for investigating the style of a literary translator. *International Journal of Translation Studies*, 12(2), 241–266. https://doi.org/10.1075/target.12.2.04bak

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9(2), 167–193. https://doi.org/10.1075/ijcl.9.2.02bak

Bartkute, D. (2020). Spatial deictics and translational implicitation: Evidence from a corpus-based analysis of English and Lithuanian fictional discourse. *International Journal of Language and Linguistics*, 8(5), 229–239.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python*. NLTK. https://www.nltk.org/

Boukhaled, M. A., & Ganascia, J.-G. (2015). *Using function words for authorship attribution: Bag-of-words vs. sequential rules*. HAL open science. https://hal.science/hal-01198407

Bugarski, R. (1991). Contrastive analysis of terminology and the terminology of contrastive analysis. In V. Ivir & D. Kalogjera (Eds.), *Languages in contact and contrast: Essays in contact linguistics* (pp. 73–82). De Gruyter Mouton. https://doi.org/10.1515/9783110869118.73

Cinková, S., & Rybicki, J. (2020). Stylometry in a bilingual setup. *International Conference on Language Resources and Evaluation*, 977–984.

Daelemans, W. (2013). Explanation in computational stylometry. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 451–462). Springer. https://doi.org/10.1007/978-3-642-37256-8_37

Ebeling, S. O. (2013). Semantic prosody in a cross-linguistic perspective. *Studies in variation, contacts and change in English*, 13, 1–14.

Egan, T. (2012). Using translation corpora to explore synonymy and polysemy. In T. Egan & H. Dirdal (Eds.), *Crossing paths: English corpus linguistics* (pp. 1–17). https://varieng.helsinki.fi/series/volumes/12/egan/

Egan, T. (2018). The FAIL TO construction: A contrastive perspective. *Bergen language and linguistics studies*, 9(1), 173–186. https://doi.org/10.15845/bells.v9i1.1525

El-Fiqi, H., Petraki, E., & Abbass, H. A. (2011). A Computational linguistic approach for the identification of translator stylometry using Arabic-English text. In *IEEE International Conference on Fuzzy Systems*, 2039–2045. https://doi.org/10.1109/FUZZY.2011.6007535

Fillmore, C. J. (1984). Remarks on contrastive pragmatics. In J. Fisiak (Ed.), *Contrastive linguistics: Prospects and problems* (pp. 119–142). De Gruyter Mouton. https://doi.org/10.1515/9783110824025.119

Gast, V. (2012). *Contrastive linguistics: Theories and methods*. Academia. https://www.academia.edu/11678460/Contrastive_linguistics_Theories_and_methods

Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., & Gelbukh, A. (2016). Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. *Sensors (Basel)*, 16(9), 13–74.

Gómez-Adorno, H., Posadas-Duran, J.-P., Ríos-Toledo, G., Sidorov, G., & Sierra, G. (2018). Stylometry-Based approach for detecting writing style changes in literary texts. *Computación Y Sistemas*, 22(1), 47–53.

Gundersen, K. (2004). Norwegian Preposition + at-clause and its correspondences in english. *Gothenburg Studies in English*, 89, 113–127.

Hasselgård, H. (2017). Lexical patterns of place in English and Norwegian. In T. Egan & H. Dirdal (Eds.), *Cross-linguistic correspondences: From lexis to genre* (pp. 97–119). John Benjamins.

Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast*, 20(2), 184–208.

Hu, K. (2016). *Introducing corpus-based translation studies*. Shanghai Jiao Tong University Press. https://doi.org/10.1007/978-3-662-48218-6

James, C. (1980). *Contrastive analysis*. Longman.

Janicki, K. (1980). Contrastive sociolinguistics. Some methodological considerations. *PSiCL*, 10, 33–40.

Johansson, S. (2012). Cross-linguistic perspectives. In M. Kytö (Ed.), English corpus linguistics: Crossing paths (pp. 43–68). Brill. https://doi.org/10.1163/9789401207935

Kotait, R. (2016). On translating semantic prosody: A corpus-based cognitive-semantic approach. *The Proceedings of the First International Conference of the Department of English in Literature, Linguistics and Translation, Travelling, Theories: Origins and Manifestations*, 252–282.

Krzeszowski, T. P. (1990). *Contrasting languages: The scope of contrastive linguistics*. De Gruyter Mouton. https://doi.org/10.1515/9783110860146

Kwok, H. L., Laviosa, S., & Liu, K. (2023). Lexical simplification in learner translation: A Corpus-based approach. *Research in Corpus Linguistics*, 11(2), 103–124. https://doi.org/10.32714/ricl.11.02.06

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.

Lagutina, K. V., Lagutina, N. S., Boychuk, E. I., Vorontsova, I. A., Shliakhtina, E. V., Belyaeva, O., Paramonov, I. V., & Demidov, P. G. (2019). A survey on stylometric text features. *25th Conference of Open Innovations Association (FRUCT)*, 622(25), 184–195. https://doaj.org/article/0fd51dbf17b5447fb00918e6f0c89d20

Langlois, J. (2021). When linguistics meets computer science: Stylometry and professional discourse. *Training, Language and Culture*, 5(2), 51–61.

Laviosa, S. (1996). *The English comparable corpus (ECC): A resource and a methodology for the empirical study of translation* [Unpublished doctoral dissertation]. University of Manchester.

Laviosa, S. (2004). Corpus-based translation studies: Where does it come from? Where is it going?. *Language Matters*, 35(1), 6–27.

Lewis, D. (2017). Coherence relations and information structure in English and French political speeches. In K. Aijmer & D. Lewis (Eds.), *Contrastive analysis of discourse-pragmatic aspects of linguistic genres* (pp. 141–161). Springer. https://doi.org/10.1007/978-3-319-54556-1_7

Liu, K., & Afzaal, M. (2021). Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *PloS One*, 16(6), e0253454. https://doi.org/10.1371/journal.pone.0253454

Lowe, D., & Matthews, R. (1995). Shakespeare vs. Fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities*, 29, 449–461. https://doi.org/10.1007/BF01829876

Lynch, G., & Vogel, C. (2015). Chasing the ghosts of Ibsen: A computational stylistic analysis of drama in translation. *ArXiv*. https://doi.org/10.48550/arXiv.1501.00841

Lynch, G., & Vogel, C. (2018). The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations. *Computer Speech and Language*, 52, 79–104. https://doi.org/10.1016/j.csl.2018.05.002

Modoc, E., & Daiana, G. (2020). Style at the scale of the canon. A stylometric analysis of 100 Romanian novels published between 1920 and 1940. *Metacritic Journal for Comparative Studies and Theory*. https://doi.org/10.24193/mjcst.2020.10.03

Moghaddam, M. Y., Shahraki Deh Sukhteh, S., & Delarami Far, M. (2017). Explicitation in translation: A case of screen translation. *Journal of Language Teaching and Research*, 8(1), 75–80. https://doi.org/10.17507/jltr.0801.09

Molés-Cases, T. (2019). Why typology matters: A corpus-based study of explicitation and implicitation of manner-of-motion in narrative texts. *Perspectives*, 27(6), 890–907. https://doi.org/10.1080/0907676X.2019.1580754

Mostafa, M. M., & Nebot, N. R. (2018). A corpus-based computational stylometric analysis of the word 'Árabe' in three Spanish generación del 98 writers. *Journal of Language Teaching and Research*, 9(5), 928–938. https://doi.org/10.17507/jltr.0905.05

Oakes, M. P., & Ji, M. (2012). *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*. John Benjamins.

Olohan, M. (2002). Leave it out! Using a comparable corpus to investigate aspects of explicitation in translation. *Cadernos de Tradução*, 1(9), 153–169.

Olohan, M. (2003). How frequent are the contractions?: A study of contracted forms in the translational English corpus. *Target. International Journal of Translation Studies*, 15(1), 59–89. https://doi.org/10.1075/target.15.1.04olo

Olohan, M. (2004). *Introducing corpora in translation studies*. Routledge.

Peng, Q., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A python natural language processing toolkit for many human languages*. Stanza. https://stanfordnlp.github.io/stanza/

Rørvik, S. (2003). Thematic progression in translation from English into Norwegian. *Nordic Journal of English Studies*, 2(2), 245–264.

Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *ArXiv*. https://arxiv.org/abs/1609.06686

Rybicki, J., & Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's night and day in Polish. *Literary and Linguistic Computing*, 28(4), 708–717. https://doi.org/10.1093/llc/fqt027

Sajavaara, K., & Lehtonen, J. (1980). Papers in discourse and contrastive discourse analysis. *Jyvaskyla Contrastive Studies, 5. Reports from the Department of English, University of Jyvaskyla*, 6.

Sari, Y., Vlachos, A., & Stevenson, M. (2017). Continuous N-gram representations for authorship attribution. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 267–273.

Stanisz, T., Kwapień, J., & Drożdż, S. (2019). Linguistic data mining with complex networks: A stylometric-oriented approach. *Information Sciences*, 482, 301–320. https://doi.org/10.1016/j.ins.2019.01.040

Steiner, E. (2008). Explicitation: Towards an empirical and corpus-based methodology. *Meaning in Context: Implementing Intelligent Applications of Language Studies*, 235–278.

Thormodsæter, Ø. (2020). *The idiomaticity of emotion in English and Norwegian a corpus-based contrastive investigation of the phraseology of the three English-Norwegian verb pairs enjoy-nyte love-elske and like-like* [Unpublished Dissertation]. University of Oslo.

Wells, H. G. (1898). *The War of the Worlds*. Project Gutenberg. https://www.gutenberg.org/ebooks/36

Wells, H. G. (2012). *The War of the Worlds* (S. A. Taha, Trans.). Hindawi Institution. https://www.hindawi.org/books/82686040/

Wu, R. (2022). A Corpus-based contrastive analysis of English-Chinese main negatives. *International Conference on Social Sciences and Humanities and Arts*, 528–531.

تحليل القياس الاسلوبي للذخائر المتوازية: دراسة اسلوبية مقارنة بين رواية "حرب العوالم" وترجمتها باللغة العربية

رانيا عبد الفتاح الوكيل
كلية اللغات، جامعة أكتوبر للعلوم الحديثة والآداب، جمهورية مصر العربية.
relfattah@msa.edu.eg

ألفت نور الدين قرنـي
كلية اللغات، جامعة أكتوبر للعلوم الحديثة والآداب،جمهورية مصر العربية.
olfat.noureldin@gmail.com

المستخلص:
تعمد الدراسات اللغوية المقارنة إلى تحليل تكوين النص وفهمه في اللغات والثقافات المختلفة. وحديثا يتم استخدام الادوات الحاسوبية للتحليل اللغوي. ويقوم علم القياس الاسلوبي على الدراسة الكمية لأسلوب الأدبي باستخدام الادوات والوسائل الحاسوبية. تهدف هذه الدراسة إلى القيام بقياس اسلوبي مقارن بين رواية حرب *العوالم* والنص المترجم للغة العربية لبيان الصعوبات والتحديات التي تفرضها الطبيعة الخاصة والمعقدة للغة العربية امام تطبيقات معالجة اللغات الطبيعية وبناء الذخائر المتوازية. وتتوصل الدراسة إلى أن مقسم الجمل في NLTK نجح في إنتاج ذخيرة لغوية موازية على الرغم من حدوث بعض الاختلافات. أما بالنسبة لتصنيف أجزاء الكلام، فاتضح أن الأداة مدربة أكثر على قاعدة بيانات إنجليزية حيث تم تعيين علامات معظم الكلمات الإنجليزية بدقة بينما دلت نتائج النص العربي على أخطاء كثيرة وخاصة في تعريف أجزاء الكلام المترجمة عن الظروف في اللغة الإنجليزية. ويرجع هذا إلى الاختلاف الصرفي والنحوي المتبع في الترجمة العربية. أما أداة التعرف على الكيانات المسماة فإن نتائجها أدق بكثير وتكشف عن تفضيل المترجمة للجوء إلى نقل الأسماء من الإنجليزية إلى العربية دون ترجمة. وتكشف الدراسة عن عدم تطابق الاختيارات الاسلوبية للمترجم نتيجة للطبيعة المختلفة للغتين.

الكلمات المفتاحية: الدراسات اللغوية المقارنة القائمة على الذخائر المتوازية – علم القياس الاسلوبي – تطبيقات معالجة اللغة الطبيعية - تصنيف أجزاء الكلام - التعرف على الكيانات المسماة