Quantifying the Contribution of MWEs and Polysemy in Translation Errors for English–Igbo MT

Adaeze Ngozi Ohuoba, Serge Sharoff, Callum Walker
Centre for Translation Studies,
School of Languages, Cultures and Societies
University of Leeds, LS2 9JT, UK

Abstract

In spite of recent successes in improving Machine Translation (MT) quality overall, MT engines require a large amount of resources, which leads to markedly lower quality for lesser-resourced languages. This study explores the case of translation from English into Igbo, a very low resource language spoken by about 45 million speakers. With the aim of improving MT quality in this scenario, we investigate methods for guided detection of critical/harmful MT errors, more specifically those caused by non-compositional multiword expressions and polysemy. We have designed diagnostic tests for these cases and applied them to collections of medical texts from CDC, Cochrane, NCDC, NHS and WHO.

1 Introduction

In recent years, there has been increased research into improving the quality of machine translation (MT) outputs (Wu et al., 2016; Hassan et al., 2018). Evidenced by the switch from rule-based and statistical MT systems to neural MT systems, this has led to visible improvement of MT outputs. However, these improvements are more common with 'high-resourced languages' that have sufficient data resources for training MT models. Thus, 'low/under-resourced languages' like Igbo, lag behind in this progress. The Igbo language (Ásusu Ìgbò) is one of the three major languages spoken in Nigeria, it is the native language of the Igbo people, an ethnic group in South-Eastern Nigeria. It is also, a recognised minority language in Equatorial Guinea and Cameroon[1].

Regarding language resources, Igbo language, for instance, has only 18,369 Wikipedia articles as of 02 October 2023 unlike the English and French languages that are in the top 5 languages used in Wikipedia with 6,722,185 and 2,557,357 articles respectively. Additionally, there are no single parallel corpora with Igbo language as a language pair in Sketch Engine[2] and only a few available in OPUS[3]. The amount of parallel data that includes Igbo language as one of the language pairs remains limited. Thus, "Igbo – any language" is a low-resource language pair.

Another critical aspect of the research into MT output is the evaluation of MT for health domains. This is especially important given the recent experience with the COVID-19 pandemic where the majority of the world's population had to be confined in isolation. Prior to the COVID-19 pandemic, there had been calls for increased research into translation in time of crisis championed by the International Network on Crisis Translation.[4] One of the goals of this network was "to make meaningful and effective contributions ... that enable accurate and timely translation-enabled crisis communication, with a particular focus on health-related content".

In crisis, access to information in one's L1 cannot be over-emphasized, and given the speed, cost-effectiveness, and easy availability of MT during crisis or in situations like the self-isolation necessitated by the coronavirus pandemic, it is safe to

[1] https://www.africanexponent.com/8-most-spoken-local-languages-in-africa/

[2] https://app.sketchengine.eu/

[3] https://opus.nlpl.eu/

[4] https://cordis.europa.eu/project/id/734211

assume as pointed out by O'Brien (2022), that MT would be the most logical 'go-to' tool for users. This has been evidenced many times, the most recent being the massive deployment of MT during the Ukraine crisis (Cirule, 2022)

Regarding the reliability of these MT tools, there have been recent claims comparing MT quality to be relatively close to human translations (Wu et al., 2016; Hassan et al., 2018). However, there are still some reservations on the perceived high quality of machine translation based on some qualitative evaluations (Läubli et al., 2020; Wang et al., 2021). These evaluations, which were carried out on news texts, found that translating between languages with varying word orders posed a challenge for machine translation systems. They also reported MT systems' weak error tolerance, which makes them susceptible to inaccurate translations due to minor punctuation or spelling errors that might be overlooked by a human. The inability to discern what should be emphasised or omitted; and data sparsity in low-resource language pairs and domains were also identified as hindrances to human-machine parity in machine translation. Dew et al. (2018) noted that for statistical-based machine translation systems, language resources matter, as they are known to perform better with language pairs that are well represented online. Other studies on neural machine translation systems (Lakew et al., 2018; Murthy et al., 2018; Singh and Singh, 2022) also agree that MT quality into low-resourced languages is relatively low.

In his empirical evaluation of the quality of machine translation of 20 phrases from English into 107 languages, using Google Translate (GT),[5] Benjamin (2019b)'s study recorded good, almost perfect outputs for high-resourced languages and lesser quality outputs for low-resourced languages. For Igbo language, he reported that 47.5 per cent of the texts were accurately translated while noting that GT was able to provide fairly meaningful translations 60 per cent of the time. Even so, his study was on short non-ambiguous phrases of common usage.

More studies to evaluate and improve MT quality for health domain and especially into/from low-resourced languages is therefore considered imperative as the accuracy of health information received in times of crisis is vital. Our aim here is to provide evaluation of specific phenomena for automatic identification of translation problems. As such, an integral part of this research includes highlighting problem areas that negatively affect MT quality of health-related texts from English – Igbo, as this area has not yet been explored.

In our preliminary studies, we have identified MWEs and polysemy as the most common causes of critical errors.

Research questions:

- What proportion of critical errors can be identified via automatic detection of MWEs and polysemy?
- What is the most appropriate granularity for error detection: sentence, segment or full text?

2 Related Work

Multi-Word Expressions (MWEs): Multi-word expressions (MWEs) have been defined by Sag et al (2002) as a combination of words for which syntactic or semantic properties of the whole expression cannot be obtained from its parts. They are lexicalized combinations of two or more words that are exceptional enough to be considered as single units in the lexicon (Schneider et al., 2014). Non-compositionality in phraseology (difficulty in deriving meaning from individual components) and non-substitutability (components cannot be replaced with synonyms) have been reported by Premasiri and Ranasinghe (2022) as features of MWEs that are challenging for NLP (Natural Language Processing).

Zaninello and Birch (2020) in evaluating the effect of annotation and data augmentation in the English – Italian translation of MWEs in a neural machine translation system, note that for non-compositional MWEs, the translation quality was especially low. They report that following their study, there is clear indication that NMT systems find it difficult to translate non-compositional MWEs even for high-resourced languages, and that focusing on improving MWEs in a text can not only improve the quality of translation of MWEs in the text but also the overall machine translation quality. Arvi (2018) in his comparison of a rule-based machine translation system, (SALAMA), and Google Translate's translations of multi-word expressions in news texts from English – Swahili, discovered that Salama performed better than GT in translating MWEs. He therefore advocated the

[5] https://translate.google.com/

investment into rule-based system for translating highly inflectional low-resourced languages, as the rules can be adapted to similar languages and the accuracy of the translation would not be dependent on large parallel data.

Polysemy: Abdelaal and Alazzawie (2020) posit that Arnold et al. (1994)'s stance that ambiguity is a big challenge for MT no longer holds true for Google Translate since its switch to a neural system. Nevertheless, Xie et al (2021) report the occurrence of inconspicuous yet clinically significant medical and health (English–Chinese) machine translation mistakes suspected to be due to the limited ability of current neural MT systems to correctly interpret the meaning of polysemous words. Thus, leading to an increase in risks for end-users of machine translation systems. Meenal and Govindarajan (2023) whose research was on the challenge of machine translating polysemous words across various domains from French–English on Google Translate, concluded that the translations confirmed MT's current incapability to correctly translate polysemy even for a high-resourced language pair like French and English.

The above cases show that polysemy is a challenge for MT in high-resourced languages. This is also the case for low-resourced languages as seen in other studies. For instance, in their evaluation of machine translated texts into Lithuanian across two MT systems, Google Translate and VDU, Petkevičiūtė and Tamulynas (2011) report that the two systems recorded similar significant challenges in their translation of polysemous words. Likewise, Tudor (2017)'s research on machine translating polysemous Croatian words into English language, revealed a low level of translation accuracy. Abdulaal (2022) also reports that polysemous words should be considered while machine translating literary texts from English to Arabic as the texts could contain errors due to the machine translation system's inability to properly translate such linguistic phenomenon.

3 Methodology

3.1 Detection of MWEs

There have been a few studies on the detection and classification of MWEs for use in NLP. Zaninello and Birch (2020) report that they used manually compiled entries from a bilingual and a monolingual dictionary, instead of an automatic tool to extract MWEs in order to maximise accuracy during the extraction process. Simkó et al (2017) used POS tagging and dependency parsing to detect verbal MWEs.

For MWE identification, we treated them on the basis of syntax. We noted that our data (see 3.3) contained terminological units, named entities and light verb constructions. Thus to detect and extract MWEs in our study, Spacy's POS tagger and dependency parser were used to identify syntactic patterns on the texts. Due to the multi-word named entities in the texts, we did not apply n-gram restrictions during the extraction process. After tagging the texts, we subsequently extracted the error bearing segments and manually tagged them as a test set to compare Spacy's accuracy. We determined a precision and recall score of 0.91 each, which corresponds to Spacy's accuracy evaluation claim.

3.2 Detection of Polysemy

Analysis of the Corpus of Contemporary American English shows that most common English words have at least two senses, which produces 50/50 odds in the possible case that the target language uses different words for those different senses (Benjamin, 2019a). The word "back" for instance is reported to have 36 different senses, multiplying its translation possibility by 36 if each sense correlates to a different word in a target language. There are at least 6 different translations for "back" in Igbo ("azu", "nkezu" "nke gara aga", "n'azu", "ikwado", "ebe azu"). Amongst these six translations, "azu" can also mean at least 8 different things; (back, fish, train, behind, bum, shark, retreat, rear). Scenarios like this could lead to what Benjamin (2019a) describes as the multiplication effect of polysemy in translation.

To identify polysemous words in our corpus, we used NLTK's WordNet Interface (Miller, 1990), to identify the number of their word senses. We also incorporated the use of domain statistics (Hamilton et al., 2016) and used Word2vec-google-news-300 to identify the number of contexts/domains a polysemous word can occur in. This we did to determine if the number of word senses and number of contexts a word has, affects English–Igbo MT accuracy.

3.3 Data

We collated a total of 123 English texts, approximately 200,000 words, from the United States Centers for Disease Control and Prevention (CDC)[6], Cochrane[7], the Nigeria Centre for Disease Control (NCDC)[8], the United Kingdom's National Health Service (NHS)[9] and the World Health Organization (WHO)[10] websites. These texts, published between 2019 and 2022, are primarily about COVID-19 and are either instructional or informative texts with the exception of the Cochrane text which is majorly professional and academic. The first phase of this research involved a preliminary study. For this preliminary study, we selected one text from each source, comprising a total of 168 sentences and 2000 words . We thereafter grouped the selected data into two, considering variety in terminology. Flesch reading ease score was used to assess the linguistic difficulty of the English texts:

a) Reference Information texts for Health Professionals (henceforth 'PROF'): This text contained a lot of medical terminology. The Flesch-Kincaid score for this text is 27.2 and the Flesch-Kincaid Grade level is 13.2, and thus classed as very difficult to read for non-professionals. 28 of our selected sentences had this classification.

b) Instructional and Informative texts for Public/Patients (henceforth 'Info'): The Flesch-Kincaid reading ease score for this text is 57.2 and the Flesch-Kincaid Grade level is 8.5. The text is classed as a simple text with simple syntax. It is intended to be informative for the public and there is minimal use of highly specialised terminology and acronyms. Some part of the text also contains instructions. There were 140 sentences in the "Info" text.

Our study is based on output from Google Translate as in our preliminary evaluation of three MT systems, it emerged as the best tool for the English–Igbo language pair.

Given our aim to find which detection parameters from Sections 3.1 and 3.2 provides the most appropriate granularity for error-detection, we also vary the window for detection from tokens to segments and to full texts. For the word-level evaluation, our data contained 1490 tokens for the Info text and 388 for the PROF text. We also divided the texts into segments of meaning; 171 segments for the Info text and 45 segments for the PROF text.

3.4 Annotation Guidelines and Classification of Error Categories

Annotation guidelines were prepared to improve uniformity of MT quality evaluation across the texts. Given the absence of parallel data for our selected texts, we human translated the selected English texts into Igbo, as a gold standard for evaluation. For the error classification, we combined both linguistic and medical errors; linguistic errors here are errors that border on language fluency whereas medical errors refer to errors that though being linguistically fluent, contain errors that are medically significant and can cause harm. We thus grouped the MT errors into three error categories vis general errors, syntactic errors and terminology errors. So, if an error is neither a terminology nor syntactic error, it is tagged as a general error. Thereafter, if there are a lot of major and critical errors that have been classed as general errors, the error- causing words will then be further analysed. This phase aims to confirm Xie et al (2021)'s claim that terminology is not the major challenge in machine translation of health texts and also confirm if major and critical syntactic errors are made during English–Igbo MT. For the three error classes, if a segment had more than one error category, the category with the higher error severity was applied.

An error penalty was also associated to each error category according to the severity of the error. We carried out a three-level assessment scale for this study by modelling the error severity guidelines described by O'Brien (2012) and Comparin and Mendes (2017) which were adopted from the Multidimensional Quality Metrics (MQM) framework and Localization Quality Evaluation (LQE). Following the description of each level (written below), we found the three-level assessment to be a good fit for the preliminary error analysis. Error categories/potential for harm from inaccurate machine translations were thus grouped into three levels (Minor/no potential for harm, Major/potential for harm, Critical errors/life-threatening/ catastrophic/harmful). We also favoured an arithmetic progression of error penalty score (1,2,and 3) in place of the geometric scores used by Comparin and Mendes (2017)because we wanted to make linear distinctions among the error categories, thus making the difference between categories easier to interpret and apply consistently.

i) Minor: linguistically,the output is wrong, but the reader can decode the meaning of the sentence;

[6] https://www.cdc.gov/

[7] https://www.cochrane.org/

[8] https://ncdc.gov.ng/

[9] https://www.nhs.uk/

[10] https://www.who.int/

medically, the output is wrong but does not affect understanding nor cause any harm. Minor errors have a score of 1.

ii) Major: linguistically, this is a wrong output that hinders the understanding of the text; medically, the output can cause a degree of harm that is not life-threatening. Major errors carry a weighting of two points.

iii) Critical errors: errors that make the text incomprehensible, can cause harm or that connote meaning that is opposite of the source text. 3 points are assigned to each critical error. For instance:

- Source Text (ST): even if you've had a **positive** test result for COVID-19 before.

  MT: ọbụlagodi na ị nwetala nsonaazụ nyocha **dị mma** maka COVID-19 na mbụ.

  Back Translation (BT): Even if you have had a **negative** test result for COVID-19 before.

  This segment was annotated as a terminological error and had an error score of "3" because the error is a critical error and could cause life threatening harm.

- ST: If you **book for** someone:...

  MT: Ọ bụrụ na ị **na-ede akwụkwọ maka** mmadụ:

  BT: If you are **writing about** someone

  This segment containing a linguistic error was annotated as a general error with an error score of "2".

- ST: Have a high **temperature**

  MT: Nwee **okpomọkụ** dị elu

  BT: Have a high **hotness**

  This segment was annotated as a terminological error with an error score of "1" as the meaning can be deduced.

Quantifying Error Severity

A cross-genre similarity is also identified, as error severity follows the same pattern for both the PROF and Info texts. The results in Table 1 indicate that there are more major and critical general errors, causing about 51 per cent of total errors annotated. Terminology based major and critical errors constitute about 24 per cent of total errors. We find that the MT system did not record any major or critical syntactic errors, as the only syntactic errors were minor and did not distort the meaning of the text. This conforms with Xie et al. (2021)'s opinion, that terminology is not the major challenge in machine translation of health texts which we sought to test its applicability to English–Igbo machine translation and thus decide if a fine-grained analysis on the exact typology of these errors would be beneficial. Subsequent to this finding, the next section is dedicated to an in-depth analysis of these errors.

3.5 Analysis of Major and Critical MT Errors

In this phase, we merge major and critical errors into one label (harmful) thus narrowing the criteria to either a negligible error or a harmful error. We also run further in-depth harmful-error analysis on the level of tokens. We therefore used Spacy's POS tagger to identify the parts of speech and dependencies of the tokens that caused harmful MT errors.

As evident in Table 2, adverbs and adjectives cause the highest amount of harmful errors in the Info text, whereas nouns and verbs account for the most errors in the PROF text (Table 3). Given the fact that the PROF text contains a lot of medical terminology which are mainly nouns and verbs and the Info text has a lot of descriptions as an informative text, we find that these distinctive features contribute to the difference in the ranking of the top 5 error-causing parts of speech for the two texts. However, a notable similarity in this scenario is that the error causing parts of speech for the Info also features in the top 5 harmful error causing parts of speech for the PROF text [ADV,ADJ,NOUN, VERB,PROPN]. We therefore analyse the features of the error-causing words to determine if there are more definite similar features between the cause of harmful errors for both text types.

3.6 Multi-word Expressions and their Impact as Cause of Critical Errors in English–Igbo MT

One other distinctive feature of the error-causing POSs was that they formed part of multi-word expressions. This part of the experiment thus served to reveal if the top error-causing parts of speech for both text types have similar linguistic classifications as part of MWEs.

Thus, to analyse the effect of multi-word expressions on English–Igbo MT, we would focus on their frequency, syntactic constructions and semantic properties in the source text.

Text Type	General Errors			Syntactic Errors			Terminology Errors		
	Negligible**	Harmful***		Negligible	Harmful		Negligible	Harmful	
	Min.	Maj.	Crt. *	Min.	Maj.	Crt.	Min.	Maj.	Crt.
PROF	10	13	8	1	0	0	0	10	5
Info	24	38	26	5	0	0	2	13	12

Table 1: Error severity by segments. *Critical errors **Negligible: scores < 2 *** Harmful errors: scores ≥ 2

POS	Freq.	Harmful Error	% Error
ADV	71	15	21.13%
ADJ	118	24	20.34%
NOUN	330	45	13.64%
VERB	246	29	11.79%
PROPN	72	8	11.11%
AUX	72	5	6.94%
SCONJ	50	3	6.00%
ADP	161	9	5.59%
DET	78	4	5.13%
PRON	149	7	4.70%
PART	52	2	3.85%
CCONJ	73	2	2.74%
NUM	14	0	0.00%
X	2	0	0.00%

Table 2: Frequency of Harmful-Error causing POS (Info)

POS	Freq.	Harmful Error	% Error
NOUN	127	40	31.50%
VERB	47	13	27.66%
PROPN	20	4	20.00%
ADJ	34	6	17.65%
ADV	10	1	10.00%
ADP	43	3	6.98%
CCONJ	26	0	0.00%
DET	22	0	0.00%
AUX	16	0	0.00%
PRON	15	0	0.00%
PART	10	0	0.00%
SCONJ	8	0	0.00%
NUM	6	0	0.00%

Table 3: Frequency of Harmful-Error causing POS (PROF)

Frequency Distribution of MWEs in Source Text: Further analysis of the syntactic information of the error causing POSs and their collocates highlights that they form part of MWEs. Table 6 shows the frequency of these MWEs and the percentage errors caused by the MWEs for both text types.

Syntactic Properties of Harmful Error-causing MWEs: Classes of error-causing MWEs were distinguished based on their categorical properties and their syntactic features. It is note-worthy that most compounds in the data used are open (i.e. the compound words are written with spaces), non-compositional and non-hyphenated. This could contribute to a machine translation system's inability to properly distinguish its linguistic features and give an accurate translation. This was the case for 38 per cent of open compounds in the PROF text.

Using Spacy's syntactic dependency parser, we also analysed the dependency tags (Schuster and Manning, 2016) of the words that caused harmful errors in the MT output.

Dep	Count	Error	% Error
quantmod	5	2	40%
oprd	3	1	33%
acomp	20	5	25%
appos	4	1	25%
amod	99	20	20%
ccomp	28	5	18%
advmod	88	15	17%
nsubjpass	6	1	17%
dobj	141	23	16%
xcomp	31	5	16%
relcl	19	3	16%
npadvmod	13	2	15%
pcomp	14	2	14%
pobj	131	17	13%
advcl	53	6	11%
auxpass	10	1	10%
conj	114	10	9%
compound	81	7	9%
neg	12	1	8%
acl	13	1	8%
mark	29	2	7%
prep	152	9	6%
det	77	4	5%
aux	80	4	5%
nsubj	92	4	4%
cc	73	2	3%

Table 4: Error Frequency>1 by Dep. (Info)

We see in Table 4 and Table 5 that modifiers e.g. "quantmod", "amod", and complements e.g. "acomp", "xcomp" which form part of noun phrases, compound nouns and verb phrases are frequent causes of harmful errors.

Noun Phrases (NP) and Compound Nouns: Syntactically, compound nouns in English are typically left-branching i.e., the modifiers come before the noun whereas the reverse is the Igbo language case (right-branching: the modifiers come after the

Dep	Count	Error	% Error
prt	1	1	100%
acl	4	3	75%
npadvmod	3	2	67%
xcomp	4	2	50%
conj	23	10	43%
compound	34	14	41%
amod	31	11	35%
pobj	41	12	29%
dobj	25	5	20%
relcl	6	1	17%
advmod	13	2	15%
advcl	8	1	13%
prep	39	2	5%
ROOT	21	1	5%

Table 5: Error Frequency>1 by Dep. (PROF)

noun) e.g., hand-cream: ude aka (ude=cream, aka= hand). For noun phrases, English language accepts both forms of modification whereas Igbo accepts only post-modification. (Orji et al., 2022).

We find that 33 and 44 per cent of compound nouns and noun phrases in the Info and PROF text are causes of harmful errors, these errors nevertheless were not caused by the syntactic difference between compound nouns and noun phrases in English and Igbo (see Table 6).

Compound Verbs, Verb Phrases (VP) and Light Verb Constructions (LVC): Syntactically, compound verbs in English are typically right-branching i.e. the modifiers come after the verb, however English language accepts both forms of modification e.g. double-click (pre-modified) or throw up (post-modified). Whereas in Igbo, compound verbs are strictly post-modified e.g."weta = we-ta [to bring]". Despite this syntactic difference between English and Igbo, we did not record any harmful compound verb/VP/LVC syntactic translation error in both texts. Babych et al. (2009)'s observation that rule-based MT often mistranslates LVCs, still holds true in this English–Igbo neural MT experiment as seen in Table 7 and this mistranslation was caused by the semantic implication of LVCs.

Multi-Word Named Entities: Our experiments proved single word named entities did not cause any harmful errors. Multi-word named entities, on the other hand, were responsible for some harmful errors (20 per cent in PROF text). Thus, the need to have (multi-word) named-entities as part of the multi-word expressions. Results from our syntactic analysis of multi-word named entities further proved that the major challenge of English–Igbo MT is not primarily syntactic as Google Translate did not output any significant syntactic errors in its translation of MWEs.

Semantic Properties of Harmful-Error Causing MWEs: Given the results above,we thereafter investigated the semantic properties of the MWEs in our data. Dickins (2020) defined multi-word expressions in relation to their semantic compositions. He also classified them into three viz: "Type 1: fully non-compositional, i.e. none of the words has an independent sense; Type 2: at least one of the words has a sense which is independent but is only found in the context of this expression; and Type 3: at least one of the words has a sense which is independent but is only found in definable limited contexts of which this context is one."

A greater percentage of the MWEs in our data are open compounds and endocentric or copulative, this corresponds with Dickins (2020)'s type 2 and 3 MWEs. Less than 5 per cent of our dataset contained closed compounds (i.e. the compound words are written with no spaces or punctuation) and these closed compounds did not cause any harmful errors. One other semantic feature of note is that some harmful-error causing multi-word expressions which are type 2 and type 3 compounds (independent contextual senses) contained individual polysemous words e.g. 'positive test result'. This will be discussed in the next section.

3.7 Polysemy and its Impact as a Cause of Critical Errors in English–Igbo MT

Collocational relations and context are meant to be helpful in neural machine translation systems; nonetheless, polysemy is one of the linguistic phenomena that has been noted as a challenge to MT especially when the probability of the accurate translation of the word in context is statistically low i.e. not the most frequent sense, or its sense is insignificant in the MT system's training data for the languages in contact.

Error severity by word senses: In Table 8 and Table 9, we record the frequency and severity of the polysemous words in the data by their word-senses. We investigated if the error- causing rate of a polysemous word is directly proportional to the number of word-senses it has. One constant is that there is a similar frequency in the percentage of the errors/harmful errors caused by polysemous words in both text types. Furthermore, in at least 76 per cent of the time across all word-senses and text

MWE	Freq Info	Errors	Error %	Freq Prof	Errors	Error %
Compound Nouns/NP	82	27	33%	41	18	44%
Compound Verbs	23	9	39%	1	1	100%
Multi-Word Named Entities	24	1	4%	5	1	20%

Table 6: Manually annotated MWE errors in both texts

Source Text	Machine Translation	Back Translation
Take a break	Were ezumike	Collect a break
Get Vaccinated	Were ogwu mbochi	Collect a vaccine

Table 7: Example cases of inaccurate machine translation of VPs and LVCs

Info	**≥2**	**≥5**	**≥7**	**≥10**	**≥15**	**≥20**
Frequency	**807**	**418**	**307**	**227**	**133**	**67**
Error (**201**)	133	98	70	50	30	13
Harmful Error (**153**)	106	82	57	38	24	10
% of error is polysemous	66%	49%	35%	25%	15%	6%
% of harmful error is polysemous	69%	54%	37%	25%	16%	7%
% of Polysemous word is a harmful error	13%	20%	19%	17%	18%	15%
% of polysemous error is harmful	80%	84%	81%	76%	80%	77%

Table 8: Error severity by word senses (Info)

Prof	**≥2**	**≥5**	**≥7**	**≥10**	**≥15**	**≥20**
Frequency	**220**	**105**	**80**	**43**	**24**	**3**
Error (**82**)	61	40	32	14	5	1
Harmful Error (**67**)	55	39	31	14	5	1
% of error is polysemous	74%	49%	39%	17%	6%	1%
% of harmful error is polysemous	82%	58%	46%	21%	7%	1%
% of Polysemous word is a harmful error	25%	37%	39%	33%	21%	33%
% of polysemous error is harmful	90%	98%	97%	100%	100%	100%

Table 9: Error severity by word senses (PROF)

types, the polysemous error is a harmful error. This reveals that polysemous words do not just cause MT errors; they cause harmful errors in English – Igbo machine translation of medical texts. Another important point is that the percentage of errors for polysemous words of word-senses greater than 10 is comparatively lower than words of word- senses 7 and below.

Error severity by Polysemy domain/context: This part of the study sought to analyse if the error- causing rate of a polysemous word is directly proportional to the number of domains or contexts (con) it occurs in. We thus varied our experiments to account for different context lengths; words occurring in greater than "1,2,5,and 10" domains/contexts. However, the results show that at least 30 per cent of the polysemous words in both the Info and PROF texts were causes of harmful errors irrespective of the number of contexts the polysemous word has. For the PROF text, all the polysemous words that had up to ten contexts caused not just errors but harmful errors (Table 10 and Table 11).

Below are examples of errors caused by polysemous words in this study.

i) Source Text (ST): Always call before **visiting** your doctor or health facility.

Machine Translation (MT): Na-akpọ oku mgbe niile tupu **ịga leta** dọkịta gị ma ọ bụ ụlọ ọrụ ahụike.

Back Translation (BT): Always call before you **pay a social visit** to your doctor or health facility.

The idea here is one of going to a health centre to be seen by the health professional, not a 'social visit' as translated.

No of Domain	**$\geq$1**	**$\geq$2**	**$\geq$5**	**$\geq$10**
Frequency	**213**	**167**	**46**	**10**
Error (**201**)	97	78	20	4
Harmful Error (**153**)	75	63	15	3
% of error is con	48%	39%	10%	2%
% of harmful error is con	49%	41%	10%	2%
% of con word is harmful error	35%	38%	33%	30%
% of con error is harmful	77%	81%	75%	75%

Table 10: Error severity by no. of domains (Info)

No of Domain	**$\geq$1 (77)**	**$\geq$2 (56)**	**$\geq$5 (17)**	**$\geq$10 (5)**
Frequency	**77**	**56**	**17**	**5**
Error (**82**)	48	34	10	5
Harmful Error (**67**)	42	29	9	5
% of error is con	59%	41%	12%	6%
% of harmful error is con	63%	43%	13%	7%
% of con word is a harmful error	55%	52%	53%	100%
% of con error is harmful	88%	85%	90%	100%

Table 11: Error severity by no. of domains (PROF)

ii) ST: ...such as the emergency department or **dedicated** COVID□19 clinics.

MT: dị ka ngalaba mberede ma ọ bụ ụlọ ọgwụ COVID-19 **raara onwe ya nye**.

BT: Such as the emergency unit or COVID-19 hospital that has **committed itself**.

The translation of 'dedicated' here is that of a person instead of an allocated item.

iii) ST: ... at both title and abstract, and full□text **stage**.

MT: ... ma aha ma nke nkịtị na **ọkwa** ederede zuru oke.

BT: ... at both title and normal, and full-text **podium** . *'Stage' is translated as a theatre stage instead of its accurate connotation of a process.*

iv) ST: **Cough** or **sneeze** into a **tissue**.

MT: **Ukwara** ma ọ bụ **uzere** n'ime **anụ ahụ**.

BT: **A cough** or **a sneeze** into the **body**.

'Tissue' translated as 'anụ ahụ': body tissue, instead of its implied context of 'tissue paper'

v)ST: Talk about your concerns – anxiety at this time is **normal**.

MT:Kwuo banyere nchegbu gị - nchegbu n'oge a bụ **ihe nkịtị**.

BT: Talk about your concerns- anxiety at this time is **insignificant**.

Here the polarity of 'normal' is misinterpreted. The translation does not adequately represent the sentiments expressed. It unfortunately stifles the emotions of anxiety.

vi) ST: even if you've had a **positive** test result for COVID-19 before.

MT: ọbụlagodi na ị nwetala nsonaazụ nyocha **dị mma** maka COVID-19 na mbụ.

BT: Even if you have had a **negative** test result for COVID-19 before.

4 Discussion

Ambiguity in MWEs: Even though most of the MWEs in the source texts were endocentric, 50 per cent of the MWEs in the PROF text contained at least one polysemous word (type 3 MWE) which posed a challenge for MT and made the semantics not easily predictable from the expression. This caused an error in at least 64 per cent of MWEs with polysemous constituent words. Examples ii, iii and vi above highlight some of the cases. Google Translate was unable to recognise cases in which expressions with seemingly positive connotations are used for expressing a negative idea e.g., positive covid-19 test result in example (vi) was translated as 'nyocha dị mma maka COVID-19' implying a negative test result, as the MT system uses the connotation that 'positive' implies something good. This is different to its medical meaning showing the presence of an organism/disease. This reveals that polysemous words and multi-word expressions are to be analysed independently as polysemous words can be part of MWEs but not vice versa. The results from our study can also aid in evaluation-guided pre-editing (Babych et al., 2009) for English – Igbo machine translation and the resulting MT output could be re-evaluated to quantify pre-editing impact.

5 Conclusion and Future Work

In our paper, we have identified and quantified what linguistic features of English as a source language, create challenges for a machine translation system to accurately translate a medical text into Igbo. Our findings confirm that a medical text filled with multi-word expressions and polysemous words is not suitable for English to Igbo machine translation as Google Translate is still unable to correctly translate such linguistic properties from English to a low-resource language like Igbo. We also find that syntactic differences between the two languages do not contribute to harmful MT errors. For polysemous words, focusing on their word senses reveals an error- peak point of seven word senses, whereas all levels of domain/context numbers had a high percentage of harmful errors. As such, future work to determine if these challenges will still persist on a larger data set will be primarily on word-senses less than and equal to seven and there will be no focus on the number of contexts. Token- level analysis of our data resulted in more detailed findings and would also form the guideline of further work. As part of our wider objectives, we intend to use the results from this preliminary study to develop machine learning classifiers in order to predict medical texts that could be catastrophic for MT users to machine translate from English to Igbo. Finally, we hope our findings can also serve as a guide to evaluate/detect causes of critical MT errors for low-resourced languages especially other Niger-Congo language families.

Acknowledgements: We thank the anonymous reviewers for their time and their insightful suggestions and comments.

References

Abdelaal, Noureldin Mohamed and Abdulkhaliq Alazzawie. 2020. Machine translation: The case of Arabic-English translation of news text. *Theory and Practice in Language Studies*, Vol. 10(No. 4):408–418.

Abdulaal, Mohammad Awad Al-Dawoody. 2022. Tracing machine and human translation errors in some literary texts with some implications for EFL translators. *Journal of Language and Linguistic Studies*, 18(Special Issue 1):176–191.

Arnold, Doug, Lorna Balkan, Siety Meijer, R. Lee Humphreys, and Louisa Sadler. 1994. *Machine Translation: an Introductory Guide*.

Arvi, Hurskainen, 2018. *Sustainable language technology for African languages*, book section Sustainable language technology for African languages. Routledge.

Babych, Bogdan, Anthony Hartley, and Serge Sharoff. 2009. Evaluation-guided pre-editing of source text: improving MT-tractability of light verb constructions. In *Proc 13th European Association for Machine Translation*, pages 36–43, Barcelona, May.

Benjamin, Martin. 2019a. The astounding mathematics of machine translation, 01 April 2019.

Benjamin, Martin. 2019b. Empirical evaluation of Google Translate across 107 languages, 2019.

Cirule, Gunta. 2022. Tilde has developed Ukrainian machine translation systems to help refugees.

Comparin, Lucia and Sara Mendes. 2017. Using error annotation to evaluate machine translation and human post-editing in a business environment.

Dew, Kristin N., Anne M. Turner, Yong K. Choi, Alyssa Bosold, and Katrin Kirchhoff. 2018. Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*, 85:56–67.

Dickins, J. 2020. An ontology for collocations, formulaic sequences, multiword expressions, compounds, phrasal verbs, idioms and proverbs. *Linguistica Online*, 23:29–72.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation.

Lakew, Surafel M., Marcello Federico, Matteo Negri, and Marco Turchi. 2018. Multilingual neural machine translation for low-resource languages. *Italian Journal of Computational Linguistics*, 4(1):11–25.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *The Journal of Artificial Intelligence Research*, 67:653–672.

Meenal, T S and P Govindarajan. 2023. The challenges of using machine translation while translating polysemous words. *2023*, 11(2):5.

Miller, George. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4).

Murthy, Rudra, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2018. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages.

O'Brien, Sharon. 2012. Towards a dynamic quality evaluation model for translation. *The Journal of Specialised Translation*, (17):55–77.

O'Brien, Shanon. 2022. Crisis translation: A snapshot in time. *INContext: Studies in Translation and Interculturalism*, 2(1).

Orji, Ifeoma Maryann, Sylvanus Okwudili Anigbogu, Oluchukwu Uzoamaka Ekwealor, and Ukamaka Betrand Chidi. 2022. Enhanced machine learning algorithm for translation of English to Igbo language. *Machine Learning Research*, 7(1):8–14.

Petkevičiūtė, Inga and Bronius Tamulynas. 2011. Computer-based translation into lithuanian: Alternatives and their linguistic evaluation. *Studies about Languages*, (18):38–45.

Premasiri, Damith and Tharindu Ranasinghe. 2022. BERT(s) to detect multiword expressions. *ArXiv*, abs/2208.07832.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. *Multiword Expressions: A Pain in the Neck for NLP*.

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 455–461. European Language Resources Association (ELRA).

Schuster, Sebastian and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Simkó, Katalin Ilona, Viktória Kovács, and Veronika Vincze. 2017. USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 48–53, Valencia, Spain, April. Association for Computational Linguistics.

Singh, Salam Michael and Thoudam Doren Singh. 2022. Low resource machine translation of English–Manipuri: A semi-supervised approach. *Expert systems with applications*, 209.

Tudor, Atena. 2017. *Machine Translations of Polysemous Croatian Words in Various Text Genres*. Thesis.

Wang, Haifeng, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. *Engineering*.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Xie, Wenxiu, Meng Ji, Riliu Huang, Tianyong Hao, and Chi-Yin Chow. 2021. Predicting risks of machine translations of public health resources by developing interpretable machine learning classifiers. *International Journal of Environmental Research and Public Health*, 18(16):8789.

Zaninello, Andrea and Alexandra Birch. 2020. Multiword expression aware neural machine translation. Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3816–3825. European Language Resources Association.