ERROR ANALYSIS OF TIGRINYA – ENGLISH MACHINE TRANSLATION SYSTEMS

Negasi Haile *
Lesan.ai
negasihaile.abadi@gmail.com

Nuredin Ali *
Department of Computer Science
University of Minnesota
ali00530@umn.edu

Asmelash Teka Hadgu
Lesan.ai
asme@lesan.ai

ABSTRACT

Machine translation (MT) is an important language technology that can democratize access to information. In recent years, we have seen some progress in the development and deployment in production of MT systems for a handful of African languages. Evaluating the quality of such systems is fundamental to accelerating progress in this area. Tigrinya is a language that is spoken by more than 10 million native speakers mainly in Tigray, Ethiopia and Eritrea. In this work, we evaluated the current status of state-of-the-art MT systems that support the translation of Tigrinya to and from English: Google translate, Microsoft translator, and Lesan. We systematically collected a dataset for evaluating Tigrinya MT systems across four domains: Arts and Culture, Business and Economics, Politics as well as Science and Technology. The dataset contains snippets from 806 articles gathered from diverse sources. We performed an in-depth analysis of the errors current systems make using MQM-DQF standard error typology. We found that Mistranslation and Omission are the most frequent translation issues. We believe this work gives a methodology for evaluating other machine translation systems for low resource languages and we provide practical suggestions to improve current Tigrinya - English MT systems.

1 INTRODUCTION

Error analysis is an important part of the process of developing and evaluating machine translation systems because it allows researchers and developers to identify and understand the sources of errors in machine translation output Vilar et al. (2006) Popović & Ney (2011). By analyzing errors in machine translation, researchers can better understand the factors that contribute to translation errors and the patterns of errors that occur in different languages and contexts. This knowledge can then be used to improve machine translation systems by designing algorithms and systems that are better able to handle the types of errors that are commonly encountered. It's also important because it can help to identify areas where machine translation systems are performing well and areas where they are not. This can be useful in setting priorities for further research and development, as it can help researchers to focus on the areas where there is the greatest need for improvement. Additionally, error analysis can be used to evaluate the performance of different machine translation systems and to compare their relative strengths and weaknesses.

The performance of current state-of-the-art Machine Translation on low-resource language pairs still remains sub-optimal Maučec & Donaj (2019) compared to the high-resource counterparts, due to the unavailability of large parallel corpora Ranathunga et al. (2021). Recently, MT systems such as Google translate have added tens of new African languages Bapna et al. (2022). Microsoft translates

*Equal Contribution

also added nine new languages [1]. One of those newly added languages is Tigrinya. There are other systems such as Lesan Hadgu et al. (2022) that are dedicated to African languages. Tigrinya is a language spoken by more than 10 million people in the Tigray region of Ethiopia and most parts of Eritrea. It's a low-resourced language where there are limited resources over the internet to train machine translation systems.

Our main research objective is to quantify the most common translation issues present in current machine translation systems for Tigrinya to and from English. Through a comprehensive analysis of their weaknesses, we aim to provide practical suggestions for improvement.

2 Related Work

MT systems are evaluated using automatic and human methods. There are several automatic metrics such as BLUE (Bilingual Evaluation Understudy) Papineni et al. (2002), TER (Translation Error Rate) Snover et al. (2006), and METEOR (Metric for Evaluation of Translation with Explicit Ordering) Banerjee & Lavie (2005). These metrics measure the degree to which the machine-translated text matches the correct translation Castilho et al. (2018) Han (2016). BLEU simply calculates n-gram precision without explicitly taking into account intelligibility or grammatical correctness. METEOR is based on a generalized concept of unigram matching between machine-produced translations and human reference translations. These metrics play a huge role to evaluate the models at development time since human evaluation takes time. However, the capacity of such metrics to evaluate the syntactic and semantic equivalence is extremely limited Castilho et al. (2018) Han (2016). Previous work in Hadgu et al. (2020) evaluated MT systems: Google translate, Microsoft translator and Yandex translate for Amharic to and from English. They found an asymmetry in the quality of translation between English to Amharic and vice-versa. Several studies have shown that human evaluation generally provides more reliable and quality results than automatic evaluation of MT systemsFreitag et al. (2021) Chatzikoumi (2020). Mathur et al. (2020) compared different MT evaluation methods. They conclude that manual (human) evaluation should be the gold standard to establish significant improvements. Foradi et al. (2022) evaluated Google translate in translating English to Persian using the MQM-DQF error topology.

One of the main advantages of human evaluation is that it can lead to error analysis. Error analysis with a view to identifying different error types in machine translations can serve as a starting point to making MT systems better Koponen (2010). It provides a closer look into the errors made by the systems. There are no prior works that look into the type of errors that current MT systems make for Tigrinya, i.e., error types and severity levels for Tigrinya have not been previously explored. Building on these previous works, we will use human evaluation to evaluate current state-of-the-art systems: Google translates, Microsoft translate, and Lesan to quantify the type of errors these systems make for Tigrinya. We perform an error analysis using MQM-DQF error typology.

3 Dataset

We systematically gathered data to reflect diverse content across domains such as Art and Culture, Business and Economy, Politics as well as Science and Technology. Snippets from 806 articles where each domain contains 100 articles in each direction (Tigrinya and English) are collected. The domains are selected as they have diverse sources and different levels of availability on the internet. Politics, and Business and Economy have a better availability than Science and Technology, and Art and Culture. This would enable us to analyze the errors made over the different domains. We also gathered content that is diverse in scope, i.e., covering local and global issues. For each domain, 70% of the English source articles cover global content. For Tigrinya, 70% of the sources describe local content. This is to represent the availability of data in the internet. Where majority of English articles discuss global content. This systematic collection would provide representative data to analyze the systems from both directions.

[1] https://www.microsoft.com/en-us/translator/blog/2021/02/22/microsoft-translator-releases-nine-new-languages-for-international-mother-language-day-2021/

3.1 DOMAIN AND SOURCE

The four domains used in this dataset are selected based on various reasons: Arts and Culture, one of the main goals of MT systems is to reduce the barrier of communication between different languages. This domain includes broad cultural aspects such as food, dances, dress, games etc. Science and Technology, as there's enormous amount of scientific resource over the internet MT systems play a huge role on making these resources accessible. Business and Economics, is one of the main categories of any news papers. It's also one of the main areas where policy makers, investors from all over the world use to access the economic direction of a country. Politics, is one of the main topics of interest especially in the Tigrinya speaking community because of the political stability issues around the region. This diverse set of domains would provide an insightful inspection to evaluate the errors in the current Tigrinya MT systems.

Art and Culture: Text snippets are collected from Tigrinya student textbooks of grades 9 - 12, and web search by using keywords related to food, cultural clothes, traditional games, and dances as search keywords. For Tigrinya sources that contain local content news sources such as BBC Tigrinya, VOA (Voice of America) Tigrinya, Fana Tigrinya, student textbooks, and Social media sites such as Facebook have been used. For English sources that contain global content, cultures of various countries of the world (representing countries from each continent) from Wikipedia Arts and Culture contents [2] are selected.

Business and Economy: For Tigrinya sources, newspapers such as BBC Tigrinya, VOA Tigrinya and FBC Tigrinya are used. For English, Wikipedia's Business and Economy current events portal [3] is used as a main source.

Politics: For Tigrinya, news sources such as Fana Broadcasting Corporate S.C. (FBC) Tigrinya, BBC Tigrinya, VOA Tigrinya, and Haddas Eritrea are included. The news sources are diverse to evaluate whether the systems are biased toward a particular political group. This can be if they are trained from a source that has a biased political view. We included a filtering step to remove content with political view bias. For English sources, Wikipedia's current events portal is used as a main source. Ten countries from each continent (i.e. Africa, Europe, Asia, and North and South America) are included. For Politics and Business and Economy, the news covered between the years of 2018 - 2022 are collected to include diverse content across the years.

Science and Technology: For Tigrinya, 70% included concepts and examples from student textbooks between grades 7 and 8 are used as the main source, as there is a scarcity of online content in this area. The sub-categories include different fields such as Biology, Chemistry, Social Science, and Economics. Whereas 30% from BBC Tigrinya, FBC Tigrinya, and VOA Tigrinya category of Science and Technology. To make the English sources comparable, 70% were collected from student textbooks 9 - 12, in the same subject areas. 30% from Wikipedia Technology portal [4]. The content on Wikipedia is highly skewed containing information and news about developed countries. There was limited content about Tigray and Ethiopia in general. Collected text snippets were cleaned for formatting issues, proper spacing, misspellings, and duplicate topics.

3.2 ANNOTATION

'Sentence level', contains the evaluation of a snippet at a sentence level. Snippet level, 'Snippet level', is an evaluation of the article snippet. A single snippet could contain more than one sentence. 'Error Type', is the identified error type of the snippet. Figure 1 shows the overall structure of the dataset. The 2 annotators who are also the authors are native speakers of Tigrinya and could also speak English fluently. In total each annotator evaluated 50% of the dataset. The snippet level evaluation is the average result of the sentence level evaluation. To understand the context of sentences. The severity level for each snippet is included on the shared dataset.

The severity level for each snippet is included on the shared dataset.

The dataset is publicly available [5] for other researchers to build up on.

[2] https://en.wikipedia.org/wiki/Wikipedia:Contents/Culture_and_the_arts
[3] https://en.wikipedia.org/wiki/Portal:Current_events
[4] https://en.wikipedia.org/wiki/Portal:Technology
[5] https://anonymous.4open.science/r/Tigrinya-English-MT-Evaluation-4941

Domain	Source	Translation output	Sentence level	Snippet level	Error Type
Arts and Culture	Fiteer Baladi is Egyptian pizza which is super buttery and full of the calories, but oh so worth it! Fiteer is made of plenty of filo pastry layers that are cooked in a brick oven. The original is served plain however, it can be ordered savoury with meats, cheese and vegetables or sweet with syrup, honey or sugar.	ፊተር በላዲ ግብጻዊ ፒዛ ኮይኑ ሱፐር በትሪ ዘለዎን ካሎሪ ዝመልእን እዩ፣ ግን oh so worth it! ፊተር ካብ ብዙሓት ፊሎ ፓስተር ንጣር ዝተሰርሐ ኮይኑ ኣብ ፎርኖ ሕጡብ ዝብሰል እዩ። እቲ ኦርጅናል ስሩዕ ኮይኑ ይቕርብ ይኹን እምበር፣ ምስ ስጋ፣ በርበረን ኣሕምልትን ጣዕሚ ዘለዎ ወይ ድማ ምስ ሽሮፕ፣ መዓር ወይ ሽኮር ምቁር ክእዘዝ ይኽእል።	1,4,2	1	Untranslated
Science and Technology	The term chemotherapy has come to connote non-specific usage of intracellular poisons to inhibit mitosis (cell division) or induce DNA damage, which is why inhibition of DNA repair can augment chemotherapy.	ኤ. ንምጽጋን ዚግበር ዕንቅፋት ኬሞተራፒ ኼስፍሖ ዚኽእል በዚ ምኽንያት እዚ እዩ።	0	0	Omission
Business	ፋብረካ ሽኮር ወልቃይት ብሰብሃፍቲ ተጋሩ ክውነን ይግባእ ክብል ቤትምኽሪ ንግድን ዘፈር ማሕበራትን ትግራይ ገሊፁ። ሓላፊ እቲ ትካል ኣይተ ኣሰፋ ገብረስላሴ ሎሚ መዓልቲ ኣብ ዝሃብዎ መግለፂ፣ ፌደራል መንግስቲ ዝተፈላለያ ትካላት ልምዓት መንግስቲ ናብ ውልቀ ሰብሃፍቲ ንክዛወራ ምውሳኑ ኣዘኻኺሮም።	The Welkait sugar factory should be owned by the local authorities, said the Tigray Chamber of Commerce and sectoral Associations. Head of the agency, Asefa Gebreselassie, told Today that the federal government reminded the different government development agencies of its decision to go to individual development agencies.	1,2	1	Omission, Mistranslation
Politics	ሓገዝ ረድኤት ብዘይ ምንም ዓንቀፍቀፍ ንህዝብና ንምብፃሕ መንግስቲ ትግራይ ሎምውን ከም ወትሩ ድልው እዩ። ቅድሚ ኹሉ ዓለምለኻ ማሕበረሰብ፣ ተሓለቅቲ ሰብአዊ መሰለት፣ ትካላት ገበርቲ ሰናይን ኣብ ልዕሊ ህዝቢ ትግራይ ብወረርቲ ሓይልታት ዝበፀሖን ይበፀሖን ዘሎ ስነ ኣእሙራዊ ስብራት ሰብአዊ ህልቀትን ማሕበረ ኢኮነምያውን ቅልውላው ተረዲእኹም ንህዝቢ ትግራይ ንምሕጋዝ ዝገበርክምዎን ትገብርዎ ዘለኹም ሓገዝ መንግስቲ ትግራይ ብሽም ህዝቢ ትግራይ ምስጋንኡ ብፍሉይ ይገልፅ።	The government of Tigray, today, is always ready to reach our people with no aid. Separately, the support of the Tigray government on behalf of the people of Tigray is being appreciated and given to help the people of Tigray because of the humanitarian and community crisis that has been reported and visited by invading forces before every international community, human rights advocates, humanitarian agencies, and humanitarian organizations.	1,2	1	Mistranslation

Figure 1: Sample of the dataset after labeling

4 METHOD

After collecting and preprocessing the data, we used Google Translate API[6], Microsoft translate [7], and Lesan translation API[8] to translate text. These services were accessed between July and December 2022. To qualitatively describe these errors, we mapped each translation issue according to Multidimensional Quality Metrics (MQM) and Dynamic Quality Framework (DQF) error typology Lommel (2018). DQF-MQM error typology is a standard framework for defining translation quality metrics. It provides a comprehensive catalog of quality error types, with standardized names and definitions and a mechanism for applying them to generate quality scores. This approach to quality evaluation provides a common vocabulary to describe and categorise translation errors and create quality metrics that tie translation quality to specifications. This standardized terminology includes: accuracy (mistranslation, addition, omission, and untranslated), fluency (grammar, inconsistency, spelling, typography, and punctuation), style, and terminology. When evaluating a translation, it is typically not enough to know how many errors are present. Evaluators also need to know the severity level of the error types. By default, MQM supports five severity levels: critical, major, minor, neutral and kudos.

Two steps were followed for the evaluation. Identifying translation error severity and error type.

[6] https://cloud.google.com/translate/
[7] https://www.bing.com/translator
[8] https://api.lesan.ai/translate/v1

4.0.1 TRANSLATION QUALITY

First, to quantify the quality of translation output from the different MT systems we use a method on a scale of 5: 0 critical, 1 major, 2 minor, 3 neutral, and 4 kudos.

- **Critical**: Critical errors are those that by themselves render a project unfit for purpose. Even a single critical error would prevent a translation from fulfilling its purpose (e.g. by preventing the intended user from completing a task) and may have safety or legal implications. For example, if a translation of a text describing weight limits for an industrial centrifuge converts "2 pounds" into "2 kilograms" (instead of "0.9 kilograms"), it could result in destruction of the equipment or injury of its user, and is a critical error.
- **Major**: There is a serious problem in the translation. For example, there is addition of text not in source, some parts of the source are missing or mistranslated. It would be hard to match translation output with source text without major modifications.
- **Minor**: The translation has minor problems given the source text but requires some minor changes, e.g, changing a word or two to make it fully describe the source text.
- **Neutral**: The translation describes the source text; however, there may be some problems with style such as punctuation, and word order.
- **Kudos**: The output is a correct translation of the source text. It's both accurate and fluent.

This is done both at the sentence and snippet levels.

4.0.2 ERROR TYPES

We map the error of the translation according to the MQM-DQF typology to the particular error terminology. A detailed explanation of the error types can be found at Lommel (2018).

In our dataset, a single snippet contains more than one sentence. Hence, we rated at both sentence and snippet levels. As seen in Figure 1 multiple error types could exist in a single snippet.

5 RESULTS

5.1 QUANTIFYING MOST COMMON ERROR TYPES

We use snippet-level scores to determine whether translation quality is good or not. Each annotator was provided with 50 snippets to evaluate the inter-annotator agreement. The snippets were chosen from all systems translation outputs randomly, all domains and 25 English to Tigrinya and 25 Tigrinya to English. The annotators labeled the translation outputs for both error type and severity level identification. They had 72% agreement in labeling the error topology. In our dataset, 38.6% of the translations had no translation quality issues. As seen in Figure 2, the most common error types with 66.2% are Mistranslation and Omission. When we break down by translation direction, as shown in Figure 3, this is generally true especially when translating Tigrinya to English. Omissions are common when translating Tigrinya into English compared to the other direction. Figure 4 shows the distribution of error types by domain. Mistranslation and Omission are the two most common error types across all domains. Arts and Culture is the most challenging followed by Science and Technology in current systems.

5.2 ANALYZING TRANSLATION ISSUES

Mistranslation - This is the most common type of translation issue in current systems. These are commonly terminologies which could be technical, e.g., oxide. We observed many such errors where a system translates a given terminology by taking a part of the source token. Another common type of Mistranslations are words having different meanings depending on context(see Figure 5 for examples). Finally, we observed many occurrences of words translated with their antonyms (kick-off to start is translated as to finish).

Omission - The second most prevalent type of error is omission. The main type of omission is cases where current systems leave out an expression at the start, middle or end of a sentence. Usually

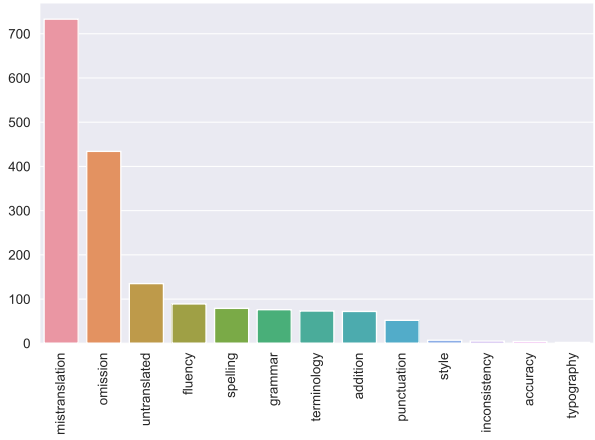

Figure 2: Distribution of Error types

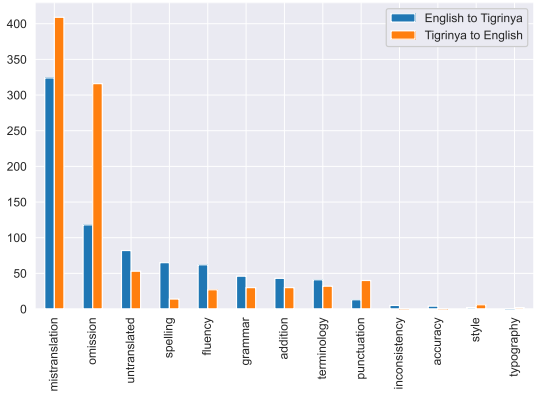

Figure 3: Error types by translation direction

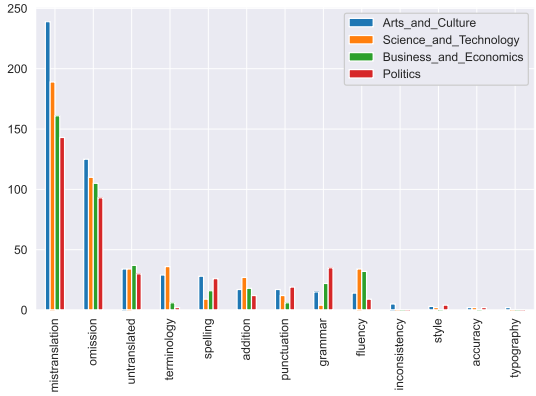

Figure 4: Error type distribution by domain

the systems translate the main idea of a sentence and leaves out supportive clauses. We observed in some systems, complete sentences are omitted from a snippet.

Untranslated - The other common type of error is untranslated tokens. These are usually abbreviations that refer to measurements (GWH), named entities e.g., political party names, currency etc. Tigrinya uses the Ge'ez script. Another common problem we observed is where some systems copy terms in the source language to the target language resulting in code-mixed output.

6 Discussion

It's important to highlight that there are several dialects of Tigrinya. We observed Microsoft uses a slightly different dialect, one that's found mostly in religious texts, than the other systems.

6.1 Recommendations

Based on the main translation issues identified in current MT systems, we suggest the following recommendations.

- Incorporation of terms in more contextual examples during training may increase the robustness of systems in accurately determining the meaning of words when used in different contexts. (Based on mistranslation issues).
- Increasing domain diversity, specifically by incorporating more content related to Art and Culture, and the inclusion of terminologies specific to Science and Technology is recommended to further improve the accuracy of these systems. (Based mistranslation issues)
- Incorporation of abbreviations and named entities in the right script as well as cleaning of the training corpus, specifically with regard to script, is recommended to avoid code-mixing in translation output. (Based on untranslated issues).
- Utilization of diverse data sources may aid in addressing issues with handling multiple dialects and styles. (This is not considered an error since any dialect of a language is a correct translation. However, outputting the standard dialect used widely would result in an easier understanding of the outputs).

7 Conclusion

Performing error analysis in evaluating the quality of MT systems is a fundamental step to accelerating the progress of Machine translation. In this case, leveraging human evaluation techniques plays an important role. In this work, we evaluated state-of-the-art MT systems that support the translation of Tigrinya to and from English: Google translate, Microsoft translators, and Lesan. We systematically collected a dataset for evaluating Tigrinya MT systems across four domains: Arts and Culture, Business and Economics, Politics as well as Science and Technology. We performed an in-depth analysis of the errors current systems make using MQM-DQF standard error typology. We found that Mistranslation and Omission are the most frequent translation issues. We believe this work gives a methodology for evaluating other low-resource languages and provides directions on the areas to focus on for Tigrinya MT research to fill current gaps.

Limitations

The work used human evaluation to analyze the errors of the MT systems. Evaluating translations produced by machine translation systems typically require human evaluators to read and assess large volumes of text, which can be time-consuming and tedious. This kind of error analysis is not suitable to quickly iterate while developing systems but should be used to determine the next development direction. The dataset size used to evaluate the system is relatively small. Snippets from a total of 806 articles across different domains are used in this study. A larger dataset could provide a much broader insight and new error categories as well.

Four domains (Politics, Business and Economics, Arts and Culture, and Science and Technology) are used to perform the study. Other domains such as Health, Entertainment (e.g., Sports), and other settings such as social media could be added to make the findings more holistic.

References

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*, 2022.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. Approaches to human and machine translation quality assessment. In *Translation quality assessment*, pp. 9–38. Springer, 2018.

Eirini Chatzikoumi. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161, 2020. doi: 10.1017/S1351324919000469.

Error type	Source Sentence	Translation output	Translation Issue
Mistranslation	The Lion dance, for example, is used to **kick off** the Chinese New Year celebrations, and it's characterized by a parade of clowns lined up in the streets.	ንኣብነት ፡ ትልሂት ኣንበሳ ፡ ንበዓላት ሓድሽ ዓመት ቻይናውያን **ንምጥፋእ** የገልግል እዩ ፤ ኣብ ጐደናታት ድማ ፡ ኣብ ባህ ዘብል ጋብላታት ይቕመጡ እዩ ።	Kick off mistranslated with a term that has the opposite meaning "**ንምጥፋእ** (remove)". The correct translation in this context should be "ንምጅማር (kick off)".
	But, all metal **oxides** are not necessarily basic oxides; for example **Al2O3** and ZnO are amphoteric oxides,which will be discussed in part (C).	እንተኾነ ግን: ኩሎም ኣቶም ብሓጺን ኣተሰርሑ ኣብውር ግድን መሰረታዊ ዝኾኑ **ኣብውር** እዮም ማለት ኣይኮነን ፤ ንኣብነት: **ኣል2ኦ3** ከምኡውን ዝኖ ሓደ ብኸፊል (ሐ) ዝዝረበሎም ኣምፎርቲክ ኦክሲድ እዮም ።	The word **oxides** is mistranslated to" **ኣብውር** (oxen)", but in this context it should be translated as "ኦክሳይድስ (oxides)".
	National Electoral Board of Ethiopia has discussed with political parties on complaints raised by parties concerning some issues during the State of Emergency.	ሃገራዊ ቦርድ ምርጫ ኢትዮጵያ ብዛዕባ እቲ ሰልፊታት ኣብ እዋን ህጹጽ ኩነታት ዘበገስዎ ጥርዓን ምስ ፖለቲካዊ ሰልፊ ተመያይጡ ኢዩ።	Party is mistranslated into "ሰልፊታት (demonstrations) or ሰልፊ (demonstration)". The correct translation of party in this context should be "ፓርቲ (party)".
	ኣቶ **ተስፋ ድራር**፣ ብዛዕባ ጥዕናዊ ረብሓ ምብላዕ ጣፍ ክዛረብ ከሎ፡ ነቲ ብስርናይ ዝስራሕ መግቢ በቲ ካብ ግሉትን ናጻ ዝኾነ ጣፍ ብምትካእ ጥዕናና ንሓሉ ይብል።	Mr. **Hope Feeding**: When talking about health benefits of eating teff, he says that we protect our health by replacing the wheat meal with a gluten-free teff.	Person name should not be translated. The name "**ተስፋ ድራር** (Tesfa Drar)" is translated to " **Hope Feeding".**
	ብ ትካላት **ትእምት** ዘጋጠመ ዕንወት ብናይ ውሽጢ ዓቅምና ብዝገበርናዮ መፅናዕቲ 11-ቢልዮን ቅርሺ ዝግመት ክሳራ እንተታጋጥም ሕዚ ንተካእዮ እንተኢልና ግን ካብ 50-60- ቢልዮን ቅርሺ'ዩ ዘድልዮ ክብሉ ኣባል ኣመራርሓ ብሓደራ ዘመሓድሮ ዘሎ ቦርድ ትእምት ኣቶ ሙሉኣለም ብርሃነ ገሊፆም።	The damage caused by the **IT** companies is estimated at 11-billion birr according to our internal capacity study but if we want to replace it now, it will need 50-60- billion birr, said Mululem Berhane, a member of the board of trustees.	The acronym **ትእምት** is mistranslated to **"IT"** instead of **EFFORT** which stands for (Endowment Fund for the Rehabilitation of Tigray).
	ብመሰረት **ዕድመ** ፕረዚደንት ኢሳይያስ ኣፈወርቂ፡ ኣብ ኤርትራ ንኣርባዕተ መዓልታት ናይ ስራሕ ምብጻሕ ዝፈጸሙ ፕረዚደንት ሰማል ሓሰን ሸኽ ማሕሙድ ሎሚ ኣብ ሰዓታት ቅድሚ ቀትሪ ናብ ሃገሮም ተመሊሶም።	Somali President Hasan Sheikh Mahmoud, who has spent four days on a working visit to Eritrea according to the **age** of President Isaiah Afwerkhi, returned to his country hours earlier today.	The word **ዕድመ** is unctextuatly translated to **age**. In this context it should be translated as "**invitation**".
Omission	ኢትዮ ቴሌኮም ኣስታት **45 ብምኢቲ ትካላቱ ብምኽኒያት** ሳንሚ ኣገልግሎት እንተይሃበ 61 ነጥቢ 3 ቢሊዮን ብር ኣቶት ምእካብ ምኽእሉ እውን ሓቢረን።	Ethio Telecom has collected 61.3 billion birr in revenue without providing services due to conflict, she said.	The **45 ብምኢቲ ትካላቱ ብምኽኒያት (because 45% of the company)** phrase is omitted.
	ኣብ ኩናት ሕድሕድ ኢትዮጵያ ንዝተፈፀሙ ገበናት ኩናትን ግህሰት ሰብኣዊ መሰላትን ንምፅራይ ዝቕመ ኮምሽን ሰብኣዊ መሰላት ሕቡራት ሃገራት ንፈለማ እዋን ናብ ኢትዮጵያ ኣትዩ።	The UN Human Rights Commission (UNHRC) has arrived in Ethiopia for the first time.	ኣብ ኩናት ሕድሕድ ኢትዮጵያ ንዝተፈፀሙ ገበናት ኩናትን ግህሰት ሰብኣዊ መሰላትን ንምፅራይ ዝቕመ (Established to investigate war crimes committed in violation of human rights during the civil war of Ethiopia) is omitted from the translated text.
	ብዶ/ር ኣረጋዊ በርሀ ዝምራሕ ንጅለ ኣባላት ዲሞክራሲያዊ ምትሕብባር ትግራይ (ዲ.ም.ት) ቅድሚ ሳልስቲ ኣዲስ ኣበባ ኣትዩ'ሎ። ዕላማ'ታ ንጅለ 'እቲ ብዶ/ር ኣብዩ ኣሕመድ ተጀሚሩ ዘሎ መስርሕ ንከይቕልበስ መታን ድጋፍ ንምብርካትን ህዝቢ ተረባሓይ ንምግባርን' ምኗኑ'ውን ተገሊጹ'ሎ።	A coalition led by the Tigray (DMT) has entered a new flower before the tigray tigray.	A complete sentence which is ዕላማ'ታ ንጅለ 'እቲ ብዶ/ር ኣብዩ ኣሕመድ ተጀሚሩ ዘሎ መስርሕ ንከይቕልበስ መታን ድጋፍ ንምብርካትን ህዝቢ ተረባሓይ ንምግባርን' ምኗኑ'ውን ተገሊጹ'ሎ (It's described that the main aim of the group is 'To continue and support the people for the initiative started by Dr. Abiy Ahmed ') is omitted from the translated text.
	Ethiopia, Namibia **Sign MoU** To Increase Cooperation In Diplomacy, Establish Political Consultation	ኢትዮጵያ ፤ ናሚብያ ኣብ ዲፕሎማሲ ምትሕብባር ከተዕዝዝ ፤ ፖለቲካዊ ሕልናኣ ከተዕቢ ትእዝዝ	**Sign MoU** is omitted from the translated text.
Untranslated	The Ethiopian Electric Power has announced that it was able to generate more than 15,400 GWh of energy in the already concluded Ethiopian fiscal year.	የኢትዮጵያ ኤሌክትሪክ ኃይል በተጠናቀቀው በጀት ዓመት ከ15,400 GWh በላይ ኃይል ማመንጨት መቻሉን አስታወቀ።	GWh is untranslated. It should be ጌጋ ዋት ኣብ ሰዓት (GWh).
	ዲሞክራሲያዊ ምንቅስቓስ ህዝቢ ትግራይ/ዲምህት/ ድሕሪ ናይ 17 ዓመታት ዕጥቃዊ ቃልሲ ደሙ ምባሉ **ንመግለሂ** ማእኸላይ ኮሚቴ ናይቲ ውድብ ብምጥቓስ ፋና ብሮድካስቲንግ ኮርፖሬት **ፀብሂባ**።	After 17 years of armed struggle, the Tigray/Democratic People's Movement **ፀብሂባ** the Torch of Broadcasting Corp., **ንመግለሂ** the organization's interim committee.	The words **ንመግለሂ (to explain)** and **ፀብሂባ (reported)** words are untranslated.

Figure 5: Distribution of Error types

Zahra Foradi, Jalilollah Faroughi, and Mohammad Reza Rezaeian Delouei. Assessing the performance quality of google translate in translating english and persian newspaper texts based on the mqm-dqf model. *Journal of Language and Translation*, 12(4):107–118, 2022.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021.

Asmelash Teka Hadgu, Adam Beaudoin, and Abel Aregawi. Evaluating amharic machine translation. *arXiv preprint arXiv:2003.14386*, 2020.

Asmelash Teka Hadgu, Abel Aregawi, and Adam Beaudoin. Lesan–machine translation for low resource languages. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 297–301. PMLR, 2022.

Lifeng Han. Machine translation evaluation resources and methods: A survey, 2016. URL https://arxiv.org/abs/1605.04515.

Maarit Koponen. Assessing machine translation quality with error analysis. In *Electronic proceeding of the KaTu symposium on translation and interpreting studies*, 2010.

Arle Lommel. Metrics for translation quality assessment: a case for standardising error typologies. In *Translation Quality Assessment*, pp. 109–127. Springer, 2018.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*, 2020.

Mirjam Sepesy Maučec and Gregor Donaj. Machine translation and the evaluation of its quality. *Recent Trends in Computational Intelligence*, pp. 143, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Maja Popović and Hermann Ney. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688, 2011.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*, 2021.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, 2006.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf.